EXPLOITING MODE CONNECTIVITY FOR OUT-OF-DISTRIBUTION DETECTION VIA LOW-DIMENSIONAL SUBSPACE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting Out-of-Distribution (OoD) from In-Distribution (InD) samples is fundamental to trustworthy deep learning. To characterize InD-OoD disparities, existing detection methods largely focus on outputs from a single local optimum, or *mode*, which provides only a limited and potentially noisy view of the underlying InD-OoD data-model relationship. To move beyond this local perspective, we study OoD detection through the lens of *mode connectivity*, where two independently trained modes can be connected by a low-loss path on InD data, suggesting that this path potentially encodes richer InD and OoD information than any isolated mode alone. We first reveal that models sampled along a low-loss connecting path exhibit nearly consistent InD accuracy but markedly different OoD detection performance, implying that OoD discriminability is not uniformly preserved along the low-loss path. Based on this finding, we propose a low-dimensional subspace modeling approach over the connecting path. Our analysis shows that low-variance principal directions are closely associated with unstable OoD behavior, whereas dominant directions capture stable structure from the path. By removing the former and aggregating the latter, we construct a *principal mode*, which is compatible with existing detection methods, delivers stronger OoD detection than even mode ensembling, and maintains competitive InD recognition. Our study highlights low-dimensional parameter subspace as a new and effective perspective for understanding and improving OoD detection.

1 INTRODUCTION

Deep Neural Networks (DNNs) generalize well on data from In-Distribution (InD) (Zhang et al., 2021), i.e., the training data distribution. However, their predictions become unreliable on unseen Out-of-Distribution (OoD) data, i.e., inputs drawn from a different distribution (Nguyen et al., 2015). Therefore, identifying whether a new sample is from InD or OoD, known as OoD detection (Yang et al., 2024), serves as a crucial step toward building reliable deep learning systems and also understanding how DNNs operate under distributional shifts.

The key to OoD detection lies in correctly characterizing the InD-OoD disparities by leveraging the outputs from DNNs, e.g., logits (Hendrycks & Gimpel, 2016; Liu et al., 2020), features (Sun et al., 2022; Fang et al., 2026), and gradients (Huang et al., 2021; Seleznova et al., 2026). For instance, logit-based scores exploit the peaked predictions on InD samples (Hendrycks & Gimpel, 2016), feature-based methods seek feature subspaces with enhanced InD-OoD separability (Fang et al., 2026), and gradient-based approaches leverage the low-rank structure of InD gradients (Seleznova et al., 2026). While these approaches have achieved notable success, they focus solely on information from a single local optimum. This view is incomplete because OoD detection seeks to capture the underlying InD-OoD data-model relationship that should remain stable across different optima. In this sense, any individual optimum provides only a local, potentially noisy realization of that relationship. This is evidenced by the substantial variation in OoD detection performance across multiple independently-trained local optima, as observed in (Fang et al., 2024a).

To move beyond this limitation, we explore OoD detection from a novel parameter-space perspective, building upon the concept of *mode connectivity* in loss landscapes of DNNs. Recent studies

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

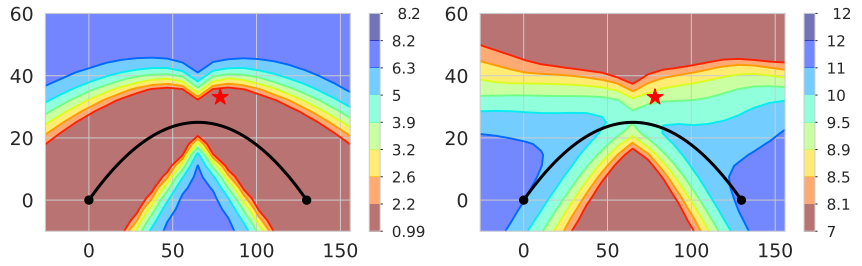


Figure 1: **Left**: Loss landscape on InD data (lower is better). **Right**: Energy score (Liu et al., 2020) landscape on OoD data (lower is more likely OoD). The black curve denotes the mode-connecting path, along which low-losses on InD data are achieved with markedly varied OoD detection results. The red star marks our principal mode: it attains low InD loss and low energy scores on OoD data, indicating effectiveness on both tasks. The visualization technique follows (Garipov et al., 2018).

(Garipov et al., 2018; Wortsman et al., 2021; Zhou et al., 2023; Zhao et al., 2025; Theus et al., 2025) have revealed that two independently trained local optima, i.e., *modes*, are often connected by a low-loss path in the parameter space, along which every point maintains high performance on InD data. This connecting path implies infinitely many models that all solve the InD task effectively, and thereby encodes far richer information about InD, and potentially about InD-OoD disparities, than any single mode can provide. Leveraging this connecting path can yield deeper insights into InD-OoD disparities, which opens new opportunities for more powerful OoD detection.

Motivated by this, we first investigate how OoD detection behavior varies along the low-loss connecting path. By sampling models from this path and evaluating their OoD detection performance, we uncover a significant phenomenon: although all these models perform nearly equally well on InD data, their OoD detection results vary substantially with a high variance, as discussed in Sec.3.1 with Figure 2. It indicates that while this path preserves structures essential for InD accuracy, OoD-sensitive variation does not arise uniformly across models along the path. Rather, as we will uncover next, such variation resides in certain parameter-space directions.

Building on this observation, we further investigate how the mode-connecting path can be exploited to improve OoD detection. Directly extending common strategies, such as output ensembling or weight averaging, to models along the path faces clear limitations, detailed in Sec.3.1. In this work, we study the path from a low-dimensional subspace perspective by Principal Component Analysis (PCA) (Pearson, 1901). Our analysis reveals that principal directions explaining only a small fraction of the parameter variation can nevertheless exhibit high OoD sensitivity under controlled perturbations. In contrast, directions with dominant variance display more stable and favorable behavior for both InD recognition and OoD detection. Therefore, we discard the low-variance principal directions and aggregate the dominant ones to construct a *principal mode*. This principal mode preserves useful knowledge contained in the connecting path while effectively eliminating unstable components that degrade OoD detection. As shown in Figure 1, our principal mode resides in a desirable region characterized by both low losses on InD data and low Energy scores on OoD data.

Notably, our principal mode is derived from a low-loss path connecting two independently trained modes. As a single model, it is compatible with existing OoD detection methods and incurs no additional detection cost at inference time. Moreover, it achieves even stronger detection performance than the ensemble of the two endpoint modes, while maintaining competitive InD accuracy simultaneously. Extensive experiments on the challenging large-scale ImageNet-1K dataset (Deng et al., 2009) with residual networks (He et al., 2016) and different types of detection methods validate the effectiveness of our approach. The main contributions of this work are summarized as:

- We study OoD detection through the lens of mode connectivity with richer InD and OoD information available, moving beyond the conventional paradigm of relying on a single mode to characterize InD-OoD disparities.
- We reveal that the low-loss path on InD data exhibits markedly varied OoD detection performance, and further show that this OoD performance variation is closely related to low-variance principal directions by a low-dimensional subspace modeling over the path.

- We construct a principal mode by removing low-variance, OoD-sensitive directions and aggregating those dominant directions, achieving stronger OoD detection than mode ensembling and maintaining effective InD recognition.
- Extensive experiments across diverse OoD detection methods and multiple independent modes demonstrate the effectiveness and broad applicability of the proposed method.

2 BACKGROUND

2.1 OUT-OF-DISTRIBUTION DETECTION

OoD detection seeks to identify samples whose distribution deviates from that of the training data (Nguyen et al., 2015; Yang et al., 2024). Generally, OoD detection is formulated as the following binary classification problem with a scoring function $S(\cdot)$:

$$\text{category of } \boldsymbol{x} = \begin{cases} \text{InD}, & S(\boldsymbol{x}) \geq s, \\ \text{OoD}, & S(\boldsymbol{x}) < s. \end{cases} \tag{1}$$

$S(\cdot)$ is devised to characterize the InD-OoD disparities, and determines the category of a test input $\boldsymbol{x}$ by comparing its score $S(\boldsymbol{x})$ to an associated threshold s. $S(\cdot)$ is typically built upon DNN outputs to capture the distinctive patterns that differentiate InD and OoD data. A brief overview of different OoD detection methods can be found in Appendix A.

2.2 MODE CONNECTIVITY

Mode refers to a local optimum achieved during the training of DNNs. The optimization of DNNs is heavily influenced by random factors such as initialization and noise, which are controlled by the random seed. Different seeds produce distinct training trajectories, each converging to an isolated local optimum. These are called independent modes (see Fig.3 in (Fang et al., 2024a)). In the high-dimensional parameter space of DNNs, there exists a vast number of such modes.

Interestingly, prior work has shown that these seemingly isolated modes are not truly independent. **Mode connectivity** reveals that for any two independently trained modes, there exists a continuous path in the parameter space along which every intermediate point maintains high performance on test data (InD data) (Draxler et al., 2018; Garipov et al., 2018; Wortsman et al., 2021; Zhou et al., 2023; Zhao et al., 2025; Theus et al., 2025). This phenomenon suggests that the loss landscapes of DNNs are far from chaotic. Instead, well-performing solutions are connected by simple, low-loss pathways, offering valuable insights into DNN optimization and generalization.

Let $\boldsymbol{\omega}_1 \in \mathbb{R}^D$ and $\boldsymbol{\omega}_2 \in \mathbb{R}^D$ be two independently-trained modes, obtained by minimizing a loss function $\mathcal{L}(\boldsymbol{\omega})$, e.g., cross-entropy for classification, on the same training set w.r.t. different random seeds. D denotes the number of the DNN parameters. We consider a continuous parametric curve $\phi_{\boldsymbol{\theta}}(t) : [0, 1] \to \mathbb{R}^D$ with parameters $\boldsymbol{\theta}$, such that its endpoints satisfy $\phi_{\boldsymbol{\theta}}(0) = \boldsymbol{\omega}_1$ and $\phi_{\boldsymbol{\theta}}(1) = \boldsymbol{\omega}_2$. Two exemplar parameterizations of $\phi_{\boldsymbol{\theta}}$ are described below.

Polygonal chain. A straightforward way to parameterize a curve connecting two endpoint modes $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ is to construct a polygonal chain. Its shape is defined by the intermediate vertices, i.e., the "bends", as learnable parameters $\boldsymbol{\theta}$. A simple case with one bend is given as follows:

$$\phi_{\boldsymbol{\theta}}(t) = \begin{cases} 2(t\boldsymbol{\theta} + (0.5 - t)\boldsymbol{\omega}_1), & 0 \leq t \leq 0.5, \\ 2((t - 0.5)\boldsymbol{\omega}_2 + (1 - t)\boldsymbol{\theta}). & 0.5 \leq t \leq 1. \end{cases} \tag{2}$$

Bézier curve. A more flexible and smooth alternative is the Bézier curve. A quadratic Bézier curve $\phi_{\boldsymbol{\theta}}(t)$ with endpoints $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ is defined by a single learnable intermediate control point $\boldsymbol{\theta}$:

$$\phi_{\boldsymbol{\theta}}(t) = (1 - t)^2\boldsymbol{\omega}_1 + 2t(1 - t)\boldsymbol{\theta} + t^2\boldsymbol{\omega}_2, \quad 0 \leq t \leq 1. \tag{3}$$

Prior work has introduced different techniques for learning such connecting paths, with their specific optimization objectives detailed in (Draxler et al., 2018; Garipov et al., 2018; Wortsman et al., 2021) and intuitive illustrations provided in Fig.1 of (Garipov et al., 2018) and Fig.2 of (Wortsman et al., 2021). Recently, linear low-loss mode-connecting paths are discussed in (Zhou et al., 2023; Zhao et al., 2025; Theus et al., 2025). These studies have advanced deeper insights into the parameter-space properties of DNN optimization.

3 METHODOLOGY

OoD detection has long centered on a single local optimum, or *mode*, extracting its outputs, such as logits (Liu et al., 2020), features (Sun et al., 2022), and gradients (Seleznova et al., 2026), to separate OoD from InD. This single-mode view is inherently limited: InD-OoD disparities arise from an underlying data-model relationship that is not specific to any particular parameter solution. Consequently, any isolated mode captures merely a local and potentially noisy instance of that relationship. Indeed, it has been observed that isolated modes can yield significantly different detection results (Fang et al., 2024a), further underscoring the inadequacy of relying on any single one.

Therefore, we move beyond isolated modes and investigate OoD detection through the lens of *mode connectivity*. To be specific, studies on mode connectivity have shown that any two independently trained modes can be connected by a low-loss path along which every model maintains strong performance on InD data (Garipov et al., 2018). This suggests that such connecting paths encode more complete InD information, and potentially richer InD-OoD disparities, than any single mode, thereby offering opportunities to strengthen OoD detection. While the use of multiple isolated modes has been studied (Fang et al., 2024a; Xu et al., 2024), this mode-path view remains largely underexplored for OoD detection. In this work, we begin by posing two fundamental questions as follows:

`(Q1)` How does OoD detection performance vary along a mode-connecting path?

`(Q2)` How can we leverage this connecting path to boost OoD detection?

In the following, we answer `(Q1)` in Sec.3.1 by empirically evaluating the detection performance along the path, and then address `(Q2)` in Sec.3.2 through a low-dimensional subspace modeling approach on the connecting path for enhanced OoD detection.

3.1 OOD DETECTION ALONG THE MODE-CONNECTING PATH

To address `(Q1)`, we conduct an empirical study that investigates OoD detection performance along a mode-connecting path. To be specific, two independent modes are obtained by training two ResNet50 models ($\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$) on the ImageNet-1K training set w.r.t. two different random seeds, respectively. Then, following (Garipov et al., 2018), a Bézier curve $\phi_{\boldsymbol{\theta}}(t)$ of Eqn.(3) connecting $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ is learned. We sample 20 models by computing $\phi_{\boldsymbol{\theta}}(t)$ at 20 uniformly spaced values of $t \in [0, 1]$, and evaluate their performance.

Figure 2(a) shows the InD accuracy on the ImageNet-1K test set and the OoD detection FPR95 on the iNaturalist dataset (Van Horn et al., 2018) using the Energy score (Liu et al., 2020) of these sampled models. A notable phenomenon is observed: while all intermediate models maintain consistently high InD accuracy along the path, *their OoD detection performance varies substantially*. Particularly, the FPR95 values fluctuate by as many as 10 percentage points, indicating that different models along the same low-loss path can exhibit markedly different capabilities in distinguishing OoD from InD. This observation suggests that, unlike isolated modes, the mode-connecting path encodes diverse inductive biases relevant to OoD detection. More empirical evidence on different OoD datasets, connecting paths and modes to support this phenomenon is provided in Appendix B.

We now turn to `(Q2)` and explore how to leverage this connecting path to boost OoD detection. Two straightforward strategies arise naturally: *ensembling* and *weight averaging*. Specifically, given a set of models sampled along the path, one can either ensemble their outputs for decision-making (Lakshminarayanan et al., 2017; Fort et al., 2019; Fang et al., 2024a), or directly average their parameters into a single merged model (Izmailov et al., 2018; Wortsman et al., 2021). However, our empirical study reveals distinct limitations for both approaches:

- **Extra computational overhead**. Ensembling the outputs of multiple models sampled from the connecting path improves both InD recognition and OoD detection compared to any single sampled model, as shown in Figure 2(b). Yet, incorporating more models does not yield additional gains beyond ensembling the two endpoint modes $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$, while incurring extra computational overhead for storing and forwarding multiple models during inference.
- **Degraded model performance**. Although the connecting path ensures a low-loss region for InD data, weighted averaging the parameters of models sampled from this path may produce a merged model that lies off the path, and potentially in regions with poor performance. Indeed,

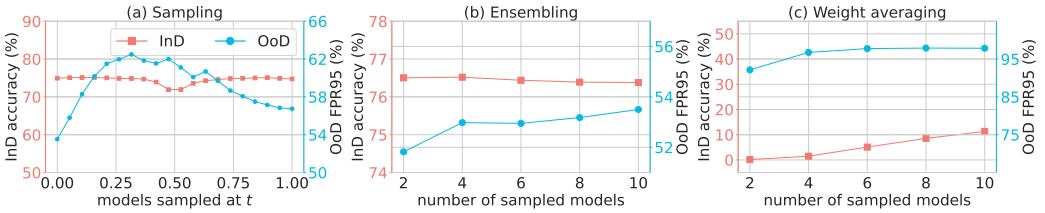


Figure 2: Investigations into InD recognition and OoD detection on a mode-connecting path (higher InD accuracy is better; lower OoD FPR is better). **(a)**: Performance of single models sampled from this path $\phi_{\boldsymbol{\theta}}(t)$ w.r.t. varied $t \in [0, 1]$. **(b)**: Performance of ensembling multiple models sampled from this path. **(c)**: Performance of averaging the weights of multiple sampled models.

> our results in Figure 2(c) show that such a weight-averaged model performs extremely poorly, achieving nearly 10% InD accuracy and over 95% FPR95 on OoD detection, despite averaging as many as 10 models sampled from the path.

Therefore, it remains an open challenge to effectively leverage the rich InD knowledge and the diversified OoD performance along this mode-connecting path beyond ensembling or weight averaging. To this end, we propose a novel perspective to `(Q2)`: low-dimensional subspace modeling over the mode-connecting path.

3.2 Low-dimensional subspace modeling over the mode-connecting path

Given two independent modes $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ and a low-loss path $\phi_{\boldsymbol{\theta}}(t)$ connecting them, each model along this path $\phi_{\boldsymbol{\theta}}(t)$ shares essential knowledge about InD, yet they differ in parametric variations that are largely irrelevant to InD but significantly affect OoD discriminability, as evidenced in Figure 2(a). To uncover the hidden patterns driving this OoD-sensitive variation, we introduce a low-dimensional subspace perspective over this path using PCA-based techniques (Pearson, 1901).

Our goal is to characterize the intrinsic low-dimensional structure of the path itself, rather than to reduce the dimensionality of the ambient parameter space. Since the object of interest is the collection of models sampled along the path, the relevant subspace is naturally defined by the variation across these samples, not by the raw parameter coordinates. Concretely, we seek a small set of parameter-space directions that best explain the dominant patterns from the sampled models.

To this end, we collect K models $\boldsymbol{\omega}^{(i)} \in \mathbb{R}^D$ sampled from $\phi_{\boldsymbol{\theta}}(t)$ and form a parameter matrix $\boldsymbol{\Omega} = [\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}, \cdots, \boldsymbol{\omega}^{(K)}] \in \mathbb{R}^{D \times K}$. We interpret each row of $\boldsymbol{\Omega}$ as the trajectory of a single parameter coordinate across the K sampled models along the path, and analyze coordinate variation among these trajectories to reveal shared patterns of change via PCA. Specifically, we compute the mean trajectory $\boldsymbol{\mu}$ across parameter coordinates, and center the matrix $\boldsymbol{\Omega}$ as $\tilde{\boldsymbol{\Omega}} = \boldsymbol{\Omega} - \mathbf{1}_D \boldsymbol{\mu}^\top$, where $\boldsymbol{\mu} = \frac{1}{D}\sum_{i=1}^{D} \boldsymbol{\Omega}_{i,:} \in \mathbb{R}^K$ and $\mathbf{1}_D$ denotes an all-ones vector of length D. Singular value decomposition is executed on the centered matrix $\tilde{\boldsymbol{\Omega}}$:

$$\tilde{\boldsymbol{\Omega}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top,\ \mathbf{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_D] \in \mathbb{R}^{D \times D},\ \mathbf{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K] \in \mathbb{R}^{K \times K}. \tag{4}$$

Here, the left Singular Vectors (SVs) $\boldsymbol{u}_j$ define orthogonal directions in parameter space, while the right SVs $\boldsymbol{v}_j$ characterize how these directions are expressed across the sampled models. The singular values in $\boldsymbol{\Sigma}$ quantify the strength of parameter variation along the corresponding directions. We jointly flip each SV pair to make the largest-magnitude entry of the left SV positive, ensuring a consistent orientation. Although the ambient parameter dimension satisfies $D \gg K$, the variations represented by the K sampled models lie in a subspace of rank at most K, allowing the path geometry to be characterized by a small number of singular directions. We next examine the resulting singular spectrum and the functional effects of these directions in Figure 3.

Figure 3(a) illustrates the spectral structure of $\boldsymbol{\Omega}$. The first three left SVs $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3$ together explain nearly all parameter variation along the path, accounting for 74.9%, 24.2%, and 0.9%, respectively, with all remaining singular values being on the order of 10^{-7} and thus negligible. This spectral concentration arises naturally from the path parameterization: $\phi_{\boldsymbol{\theta}}(t)$ is a quadratic Bézier curve, so

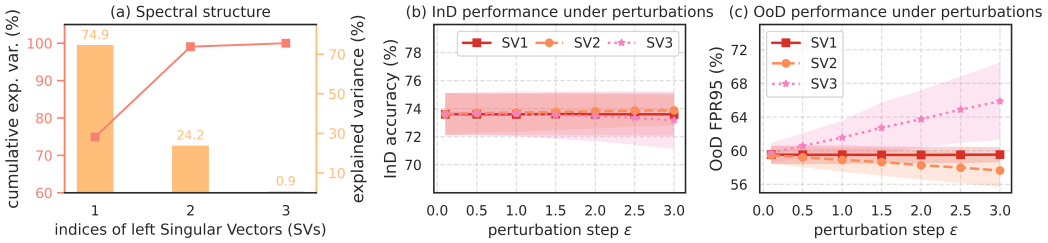


Figure 3: Investigations into the principal directions from a mode-connecting path. **(a)**: The top-3 left SVs in $\mathbf{U}$ from $\mathbf{\Omega}$ preserve nearly all the explained variance. **(b) & (c)**: InD accuracy and OoD FPR95 of the perturbed model $\boldsymbol{\omega}_{j,\varepsilon}$ of Eqn.(5) along SVs $\boldsymbol{u}_j$ w.r.t. varied steps ε.

its samples trace an intrinsically low-dimensional trajectory in the ambient parameter space. Crucially, our analysis goes beyond confirming this expected low-dimensionality and reveals markedly different functional effects across the resulting singular directions, as detailed below.

We then examine how perturbations along these directions affect InD recognition and OoD detection. To be specific, given a model $\boldsymbol{\omega}$ randomly sampled from $\phi_{\boldsymbol{\theta}}(t)$, we apply perturbations of the form

$$\boldsymbol{\omega}_{j,\varepsilon} = \boldsymbol{\omega} + \varepsilon \frac{\boldsymbol{u}_j}{\sqrt{D-1}}, \tag{5}$$

where ε denotes the perturbation step size along $\boldsymbol{u}_j$, and the normalization by $\sqrt{D-1}$ ensures comparable perturbation magnitudes across directions in the high-dimensional parameter space. Here, these left SVs $\boldsymbol{u}_j$ span a set of orthogonal basis. Each $\boldsymbol{u}_j$ corresponds to the normalized direction of the j-th principal component scores, encoding the relative variation pattern among samples along that component. Therefore, adding a perturbation ε along $\boldsymbol{u}_j$ is not introducing arbitrary noise, but instead explores a specific, principal direction derived from the low-loss path geometry. The resulting InD accuracy and OoD FPR95 of the perturbed models $\boldsymbol{\omega}_{j,\varepsilon}$ are shown in Figures 3(b) and 3(c) under repeated experiments, respectively.

Low-variance principal directions are associated with OoD-sensitive variation. According to Figures 3(b) and 3(c), perturbations along the dominant singular directions, i.e., the first two left SVs $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, have negligible impact on OoD detection performance, while leaving InD accuracy nearly unchanged. These directions thereby capture the shared, stable structure for both InD recognition and OoD detection. In contrast, although the third singular direction $\boldsymbol{u}_3$ explains little parameter variation along the path, perturbations along $\boldsymbol{u}_3$ lead to a clear degradation in OoD performance with only marginal change in InD accuracy. This observation suggests that, under controlled directional perturbations, *OoD-sensitive variation is associated with those low-variance principal directions derived from the path.* While this evidence is empirical, it offers a clear operational insight: low-variance principal directions can be discarded to suppress OoD-sensitive variation while largely preserving InD performance.

Aggregating dominant principal directions forms the principal mode. Motivated by the above observations, we construct a *principal mode* by aggregating the dominant high-variance directions that exhibit stable InD and OoD behavior, while suppressing the low-variance directions with greater OoD sensitivity, thereby consolidating the stable structure of the connecting path into a single model. Concretely, the principal mode is defined as the weighted sum of the top-s left singular vectors:

$$\boldsymbol{\omega}^* = \frac{1}{s}\sum_{j=1}^{s} \sigma_j \boldsymbol{u}_j, \tag{6}$$

where $\sigma_1, \ldots, \sigma_s$ are the top-s singular values in $\mathbf{\Sigma}$. Here s is chosen to capture components with substantial explained variance, e.g., the top-2 principal components in the example of Figure 3(a). Algorithm 1 outlines key steps to obtain this principal mode $\boldsymbol{\omega}^*$. An intuitive illustration for the position of $\boldsymbol{\omega}^*$ and the connecting path is visualized in Figure 1.

Distinction between Eqn.(6) and standard PCA reconstruction. Standard rank-s PCA reconstruction minimizes the parameter-space reconstruction error of the sampled models, with the mean

Algorithm 1 Low-dimensional subspace modeling on a low-loss path connecting two modes

Require: A low-loss path $\phi_{\boldsymbol{\theta}}(t)$ connecting two modes.
Ensure: A principal mode $\boldsymbol{\omega}^*$.

1: Sample K models from $\phi_{\boldsymbol{\theta}}(t)$ to form $\boldsymbol{\Omega} = [\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}, \cdots, \boldsymbol{\omega}^{(K)}] \in \mathbb{R}^{D\times K}$.
2: Center $\boldsymbol{\Omega}$: $\tilde{\boldsymbol{\Omega}} = \boldsymbol{\Omega} - \mathbf{1}_D \boldsymbol{\mu}^\top \in \mathbb{R}^{D\times K}$, with $\boldsymbol{\mu} = \frac{1}{D}\sum_{i=1}^{D} \boldsymbol{\Omega}_{i,:} \in \mathbb{R}^K$.
3: Perform SVD: $\tilde{\boldsymbol{\Omega}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_D] \in \mathbb{R}^{D\times D}$, $\mathbf{V} \in \mathbb{R}^{K\times K}$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times K}$ is a rectangular diagonal matrix holding singular values $\sigma_1 \geq \cdots \geq \sigma_K \geq 0$.
4: Compute principal mode $\boldsymbol{\omega}^* = \frac{1}{s}\sum_{j=1}^{s} \sigma_j \boldsymbol{u}_j$, where $\sigma_1, \cdots, \sigma_s$ are the top-s singular values.

added back. However, these sampled models exhibit markedly varied OoD detection results, so faithfully reconstructing them does not necessarily preserve or improve the desired OoD behavior. Instead of reconstructing original samples, Eqn.(6) directly constructs a representative parameter vector $\boldsymbol{\omega}^*$ from the dominant left SVs, retaining directions with stable functional behavior while excluding low-variance directions with higher observed OoD sensitivity. Since $\boldsymbol{\omega}^*$ is not a reconstruction of any particular sampled model, there is no corresponding sample-specific mean to restore.

Theoretical Discussions. To provide a unified theoretical understanding of the principal-mode construction, we characterize both its InD stability and its OoD sensitivity through the spectral geometry of the sampled models. For InD performance, given uniformly low-loss path models, and a smooth InD objective, we show that the loss of $\boldsymbol{\omega}^*$ is bounded by the loss level of the sampled models and a distance term determined by the singular-value spectrum; see Theorem 1 in Appendix C. For OoD behavior, the singular values serve as path-relative directional scales: among equal-norm perturbations, one along a direction with a smaller singular value represents a larger departure from the parameter variations observed along the path. Under an explicit local Gram-curvature alignment hypothesis, Proposition 1 in Appendix C shows that the symmetric response of a smooth OoD surrogate admits an inverse-variance lower bound, providing a conditional interpretation of the directional sensitivity observed in Figure 3. The complete assumptions, formal statements, and proofs are presented in Appendix C. Together, these results provide a theoretical support for our proposed method of retaining the dominant subspace to control InD performance while suppressing the low-variance tail associated with unstable OoD behavior.

4 Experiments

4.1 Setups

Datasets and models. In experiments, we focus primarily on the large-scale ImageNet-1K benchmark (Deng et al., 2009) as the InD dataset. It comprises 1,281,167 training images and 50,000 test images of $224 \times 224 \times 3$ within 1,000 categories. For OoD datasets, we adopt iNaturalist (Van Horn et al., 2018), SUN (Xiao et al., 2010), Places (Zhou et al., 2017) and Texture (Cimpoi et al., 2014), which have been widely adopted in existing methods (Sun et al., 2022). We choose the representative residual network (He et al., 2016) (ResNet50). Besides, additional results on CIFAR10 (Krizhevsky, 2009) and more OoD datasets are provided in Appendix D.

Baseline comparisons. Our method yields a principal mode $\boldsymbol{\omega}^*$ from a low-loss path $\phi_{\boldsymbol{\theta}}$ connecting two independent endpoint modes $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$. To comprehensively evaluate the effectiveness of $\boldsymbol{\omega}^*$, we compare it against $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, as well as their ensemble, in terms of both InD recognition and OoD detection performance.

Regarding OoD detection, we run multiple diverse detection methods on each model. For completeness, we consider 3 representative detection methods covering distinct categories: the logits-based Energy (Liu et al., 2020), the features-based KPCA (Fang et al., 2026) and the gradients-based Grad-PCA (Seleznova et al., 2026). This diverse comparison allows us to verify that the advantage of $\boldsymbol{\omega}^*$ is consistent across different types of detection methods. For the ensemble baseline in OoD detection, we adopt the mode ensemble technique introduced in (Fang et al., 2024a; Lakshminarayanan et al., 2017) for a fair comparison. A detailed elaboration on these baselines can be found in Appendix D.

Table 1: InD recognition and OoD detection results on ImageNet-1K with ResNet50. Averages with standard deviations from 3 mode pairs are reported; best results in **bold**. See Table S1 for full results.

model	InD acc. (%)	OoD datasets								**AVERAGE**	
		iNaturalist		SUN		Places		Texture			
		FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑
Energy (logits-based)											
`edpt-1`	$74.85_{\pm 0.03}$	$57.06_{\pm 1.93}$	$89.41_{\pm 0.41}$	$58.00_{\pm 1.16}$	$86.64_{\pm 0.36}$	$62.60_{\pm 0.12}$	$84.81_{\pm 0.13}$	$53.39_{\pm 1.68}$	$86.43_{\pm 0.43}$	$57.77_{\pm 1.10}$	$86.82_{\pm 0.30}$
`edpt-2`	$75.01_{\pm 0.55}$	$53.84_{\pm 4.84}$	$89.99_{\pm 0.89}$	$57.23_{\pm 1.59}$	$86.98_{\pm 0.69}$	$62.38_{\pm 1.06}$	$84.97_{\pm 0.57}$	$54.09_{\pm 2.49}$	$85.91_{\pm 0.91}$	$56.88_{\pm 2.42}$	$86.96_{\pm 0.73}$
`ensemble`	**76.67**$_{\pm 0.35}$	$51.95_{\pm 1.95}$	$91.19_{\pm 0.34}$	$55.40_{\pm 0.45}$	$87.85_{\pm 0.19}$	$60.36_{\pm 0.52}$	$85.99_{\pm 0.32}$	$49.06_{\pm 0.73}$	$88.17_{\pm 0.25}$	$54.19_{\pm 0.81}$	$88.30_{\pm 0.27}$
Ours	$74.71_{\pm 0.72}$	**35.27**$_{\pm 6.26}$	**93.21**$_{\pm 1.28}$	**44.02**$_{\pm 3.83}$	**89.38**$_{\pm 0.86}$	**53.55**$_{\pm 3.33}$	**86.52**$_{\pm 1.08}$	**42.42**$_{\pm 4.84}$	**88.32**$_{\pm 1.18}$	**43.82**$_{\pm 4.45}$	**89.36**$_{\pm 1.03}$
KPCA (features-based)											
`edpt-1`	$74.85_{\pm 0.03}$	$38.57_{\pm 3.01}$	$91.87_{\pm 0.65}$	$46.54_{\pm 1.73}$	$88.99_{\pm 0.83}$	$60.22_{\pm 1.70}$	$83.92_{\pm 0.89}$	$15.04_{\pm 1.25}$	$96.84_{\pm 0.32}$	$40.09_{\pm 1.79}$	$90.41_{\pm 0.63}$
`edpt-2`	$75.01_{\pm 0.55}$	$40.19_{\pm 3.93}$	$91.28_{\pm 1.32}$	$45.38_{\pm 6.12}$	$89.47_{\pm 2.17}$	$58.15_{\pm 6.13}$	$84.79_{\pm 2.67}$	$17.86_{\pm 3.15}$	$96.41_{\pm 0.58}$	$40.39_{\pm 3.10}$	$90.49_{\pm 1.37}$
`ensemble`	**76.67**$_{\pm 0.35}$	$33.30_{\pm 4.11}$	$93.35_{\pm 0.84}$	$37.98_{\pm 2.15}$	$91.53_{\pm 0.52}$	$52.27_{\pm 2.43}$	$87.06_{\pm 0.66}$	$19.12_{\pm 1.31}$	$96.22_{\pm 0.24}$	$35.67_{\pm 1.90}$	$92.04_{\pm 0.41}$
Ours	$74.71_{\pm 0.72}$	**26.59**$_{\pm 2.50}$	**95.22**$_{\pm 0.38}$	**24.19**$_{\pm 2.65}$	**94.61**$_{\pm 0.55}$	**39.46**$_{\pm 3.95}$	**90.70**$_{\pm 1.10}$	**18.11**$_{\pm 1.41}$	**96.48**$_{\pm 0.29}$	**27.09**$_{\pm 2.13}$	**94.25**$_{\pm 0.54}$
GradPCA (gradients-based)											
`edpt-1`	$74.85_{\pm 0.03}$	$58.26_{\pm 1.67}$	$90.61_{\pm 0.29}$	$78.15_{\pm 0.93}$	$80.86_{\pm 0.37}$	$80.91_{\pm 0.62}$	$78.16_{\pm 0.43}$	$41.91_{\pm 0.27}$	$91.92_{\pm 0.06}$	$64.81_{\pm 0.57}$	$85.39_{\pm 0.21}$
`edpt-2`	$75.01_{\pm 0.55}$	$54.99_{\pm 3.57}$	$90.97_{\pm 0.49}$	$71.56_{\pm 16.83}$	$81.94_{\pm 5.00}$	$75.40_{\pm 15.09}$	$79.17_{\pm 5.22}$	$40.62_{\pm 4.41}$	$91.98_{\pm 0.69}$	$60.64_{\pm 9.89}$	$86.01_{\pm 2.81}$
`ensemble`	**76.67**$_{\pm 0.35}$	**48.46**$_{\pm 3.94}$	**92.41**$_{\pm 0.27}$	$65.29_{\pm 15.62}$	$85.15_{\pm 3.96}$	$71.55_{\pm 14.70}$	$81.93_{\pm 4.35}$	**37.80**$_{\pm 4.45}$	**92.93**$_{\pm 0.39}$	$55.77_{\pm 8.89}$	$88.11_{\pm 2.14}$
Ours	$74.71_{\pm 0.72}$	$55.37_{\pm 1.69}$	$91.01_{\pm 0.23}$	**54.09**$_{\pm 4.02}$	**87.13**$_{\pm 1.74}$	**61.92**$_{\pm 4.55}$	**84.02**$_{\pm 2.15}$	$41.38_{\pm 3.10}$	$92.10_{\pm 0.63}$	**53.19**$_{\pm 3.09}$	**88.56**$_{\pm 1.14}$

Implementation. To obtain isolated modes, we independently train multiple ResNet50 models on ImageNet-1K w.r.t. different random seeds. To learn the mode-connecting path $\phi_{\boldsymbol{\theta}}(t)$, we adopt the classic and easy-to-implement method in (Garipov et al., 2018) with open-source code, covering both the polygonal chain of Eqn.(2) and the Bézier curve of Eqn.(3). On such paths, our low-dimensional subspace modeling is executed to obtain the principal mode $\boldsymbol{\omega}^*$. Detailed implementation settings and computational cost of this full pipeline are provided in Appendix D and can also be referred in our released code at `https://anonymous.4open.science/r/X7kQ/`.

Metrics. To evaluate InD recognition, we report accuracy on the ImageNet-1K test set. For OoD detection, we adopt two standard metrics: *(i)* the False Positive Rate (FPR) of OoD samples when the true positive rate on InD data reaches 95%, and *(ii)* the Area Under the Receiver Operating Characteristic curve (AUROC).

4.2 Main comparisons

Table 1 reports the InD recognition and OoD detection results. Specifically, we train 6 independent modes and group them into 3 pairs: (`mode-1`, `mode-2`), (`mode-3`, `mode-4`) and (`mode-5`, `mode-6`). For each pair, we learn a connecting Bézier curve, and from each curve we derive a principal mode using our low-dimensional subspace modeling approach. For the selected 3 representative OoD detection methods (Energy, KPCA and GradPCA), we report the average results with standard deviations for: *(i)* the two endpoint modes of each path (denoted as `edpt-1` and `edpt-2` in Tab.1), *(ii)* the ensemble of the two endpoint modes (denoted as `ensemble`), and *(iii)* our principal mode. Complete results of each mode pair can be found in Table S1 of Appendix D.

As illustrated, our principal mode $\boldsymbol{\omega}^*$ maintains competitive InD test accuracy when compared with the two single modes (`edpt-1` and `edpt-2`), with a difference of less than one percentage point. At the same time, $\boldsymbol{\omega}^*$ achieves notably stronger OoD detection performance, even surpassing the ensemble of the two modes. It suggests that our low-dimensional parameter subspace modeling effectively removes OoD-sensitive variation without significantly compromising InD performance.

4.3 Sensitivity analysis

Comparisons on different connecting paths. We compare $\boldsymbol{\omega}^*$ derived from two different parameterizations of the same connecting mode pairs: the Bézier curve of Eqn.(3) and the polygonal chain (PolyChain) of Eqn.(2). Table 2 reports the average results with standard deviations over 3 mode pairs. On ImageNet-1K, the Bézier-derived principal mode generally performs better than its PolyChain counterpart, suggesting that the flexibility and smoothness of the Bézier curve can be advantageous in this setting. Full PolyChain results are provided in Table S1 of Appendix D.

Effect of the number of preserved singular vectors. As discussed in Sec.3.2, the number s of preserved left Singular Vectors (SVs) is intended to capture substantial variance. This is validated in the left panel of Figure 4. Retaining only the first left SV, which explains nearly 75% of the

Table 2: Comparisons of $\boldsymbol{\omega}^*$ derived from different paths: Bézier and polygonal chain (PolyChain). OoD detection results are averaged across the selected datasets.

curve	InD acc. (%)	OoD detection *Energy* FPR↓	*Energy* AUROC↑	*KPCA* FPR↓	*KPCA* AUROC↑	*GradPCA* FPR↓	*GradPCA* AUROC↑
Béizer	$74.71_{\pm 0.72}$	$43.82_{\pm 4.45}$	$89.36_{\pm 1.03}$	$27.09_{\pm 2.13}$	$94.25_{\pm 0.54}$	$53.19_{\pm 3.09}$	$88.56_{\pm 1.14}$
PolyChain	$73.44_{\pm 0.01}$	$47.16_{\pm 0.73}$	$87.78_{\pm 0.13}$	$29.91_{\pm 2.95}$	$93.76_{\pm 0.73}$	$55.83_{\pm 1.75}$	$88.55_{\pm 0.08}$

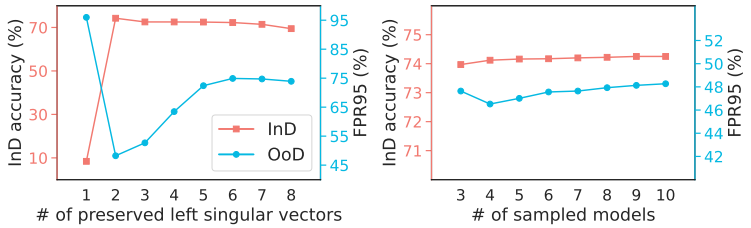


Figure 4: **Left**: Performance of $\boldsymbol{\omega}^*$ w.r.t. the number of preserved left singular vectors s (sampling number fixed at $K = 10$). **Right**: Performance of $\boldsymbol{\omega}^*$ w.r.t. the number of sampled models from the connecting path K (preserved left singular vectors fixed at $s = 2$). The reported FPR95 values are from the Energy method and averaged across the selected OoD datasets.

variance, leads to poor InD accuracy and OoD detection. Including the second SV substantially improves both metrics, whereas incorporating more than 2 SVs coincides with degraded OoD detection performance. These observations suggest that $s = 2$ is a reasonable choice, as both SVs capture significant variance individually and jointly preserve almost all the information.

Effect of the number of sampled models. The right panel of Figure 4 shows the performance of $\boldsymbol{\omega}^*$ w.r.t. the number K of sampled models along the path. The results show that $\boldsymbol{\omega}^*$ remains stable across different K, with consistent InD accuracy and OoD FPR95. This insensitivity suggests that the principal subspace is already well characterized by a small number of samples, as additional models contribute little beyond the dominant variance captured by the first few principal directions. In Table 1, we set $K = 5$ for a balance between computational cost and performance.

5 Conclusion and Discussion

In this work, we study OoD detection through the lens of mode connectivity, focusing on the low-dimensional structure of mode-connecting paths. We first uncover an underexplored phenomenon: low InD loss paths connecting two modes exhibit largely varied OoD detection performance. To better understand and exploit this, we develop a PCA-based low-dimensional subspace modeling approach, and reveal that OoD-sensitive variations are closely associated with low-variance principal directions. Based on this insight, we construct a principal mode by aggregating the dominant principal directions, preserving the low-loss InD structure of the path while removing low-variance directions with higher observed OoD sensitivity. Extensive experiments show that our principal mode maintains effective InD recognition and achieves stronger OoD detection than mode ensembling. By shifting the focus from isolated modes to the low-dimensional structure of connecting paths, our work offers a new parameter-space perspective for OoD detection, suggesting that the intrinsic geometry of the solution manifold is a largely untapped resource for understanding how DNNs behave under distribution shifts.

Meanwhile, our low-dimensional modeling of the mode-connecting path remains data-agnostic, relying on parameter variations along the path without explicitly using data-dependent information. While this perspective proves effective on the datasets and networks considered in our experiments, its applicability to broader architectures and data distributions remains to be further validated. This suggests an important avenue for future research: incorporating data-dependent signals into the low-dimensional parameter subspace modeling process, so that the learned subspace can better reflect the underlying InD-OoD relationship and generalize more reliably across diverse settings.

AI USE STATEMENT

In this work, we used generative AI tools for *helping develop theoretical models, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and designing or providing feedback on experiments*. We have NOT used generative AI tools for *helping develop conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology, implementing methods, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, and interpreting results*. *Generating synthetic data sets* is not applicable to this work.

Additionally, we used generative AI tools for *drafting parts of a research paper, and editing a research paper to improve readability*.

We have reviewed all AI-assisted work. We carefully checked theoretical results and parts of the proof generated from the LLMs to ensure correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Thomas Botschen, Konstantin Kirchheim, and Frank Ortmeier. Out-of-distribution detection with adversarial outlier exposure. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4391–4400. IEEE Computer Society, 2025.

Jiali Cheng and Hadi Amiri. Understanding machine unlearning through the lens of mode connectivity. In *ICML 2025 Workshop on Machine Unlearning for Generative AI*, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Kun Fang, Qinghua Tao, Xiaolin Huang, and Jie Yang. Revisiting deep ensemble for out-of-distribution detection: A loss landscape perspective. *International Journal of Computer Vision*, 132(12):6107–6126, 2024a.

Kun Fang, Qinghua Tao, Kexin Lv, Mingzhen He, Xiaolin Huang, and Jie Yang. Kernel pca for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:134317–134344, 2024b.

Kun Fang, Qinghua Tao, Mingzhen He, Kexin Lv, Runze Yang, Haibo Hu, Xiaolin Huang, Jie Yang, and Longbing Cao. Kernel pca for out-of-distribution detection: Non-linear kernel selection and approximation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):13220–13235, 2026. doi: 10.1109/TPAMI.2026.3705778.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Maximilian Granz, Manuel Heurich, and Tim Landgraf. Weiper: Ood detection using weight perturbations of class projections. *Advances in Neural Information Processing Systems*, 37:35879–35908, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2016.

Alvin Heng, Alexandre H Thiery, and Harold Soh. Out-of-distribution detection with a single unconditional diffusion model. *Advances in Neural Information Processing Systems*, 37:43952–43974, 2024.

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

P Izmailov, AG Wilson, D Podoprikhin, D Vetrov, and T Garipov. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pp. 876–885, 2018.

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Toronto*, 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Aodi Li, Liansheng Zhuang, Xiao Long, Houqiang Li, and Shafei Wang. Exploring mode connectivity in krylov subspace for domain generalization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *International Conference on Machine Learning*, pp. 22528–22538. PMLR, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Guruprasad Raghavan, Bahey Tharwat, Surya Narayanan Hari, Dhruvil Satani, Rex Liu, and Matt Thomson. Engineering flexible machine learning systems by traversing functionally invariant paths. *Nature Machine Intelligence*, 6(10):1179–1196, 2024.

Mariia Seleznova, Hung-Hsu Chou, Claudio Mayrink Verdun, and Gitta Kutyniok. Gradpca: Leveraging ntk alignment for reliable out-of-distribution detection. In *International Conference on Learning Representations*, 2026.

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

Alexander Theus, Alessandro Cabodi, Sotiris Anagnostidis, Antonio Orvieto, Sidak Pal Singh, and Valentina Boeva. Generalized linear mode connectivity for transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Mitchell Wortsman, Maxwell C Horton, Carlos Guestrin, Ali Farhadi, and Mohammad Rastegari. Learning neural network subspaces. In *International Conference on Machine Learning*, pp. 11217–11227. PMLR, 2021.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Chenhui Xu, Fuxun Yu, Zirui Xu, Nathan Inkawhich, and Xiang Chen. Out-of-distribution detection via deep multi-comprehension ensemble. In *International Conference on Machine Learning*, pp. 55465–55489. PMLR, 2024.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Understanding mode connectivity via parameter space symmetry. In *International Conference on Machine Learning*, pp. 77441–77460. PMLR, 2025.

Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. In *International Conference on Learning Representations*, 2020.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Zhanpeng Zhou, Yongyi Yang, Xiaojiang Yang, Junchi Yan, and Wei Hu. Going beyond linear mode connectivity: The layerwise linear feature connectivity. *Advances in neural information processing systems*, 36:60853–60877, 2023.

Jianing Zhu, Yu Geng, Jiangchao Yao, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Bo Han. Diversified outlier exposure for out-of-distribution detection via informative extrapolation. *Advances in neural information processing systems*, 36:22702–22734, 2023.

Appendices are organized as follows.

- Appendix A outlines related work in OoD detection and mode connectivity.
- Appendix B provides additional empirical evidence to support the OoD detection performance along the mode-connecting path.
- Appendix C presents theoretical results to support the principal-mode construction.
- Appendix D details experimental setups, complete detection results, and computational cost.

A Related work

Various OoD detection methods have been developed to characterize the InD-OoD disparities from different perspectives. Most of them focus on outputs from a single mode. In addition to the logits-based (Hendrycks & Gimpel, 2016; Liu et al., 2020), features-based (Sun et al., 2022; Fang et al., 2024b) and gradients-based (Huang et al., 2021; Seleznova et al., 2026) methods outlined in Sec.1, other approaches include introducing outliers during training (Zhu et al., 2023; Botschen et al., 2025), leveraging the diffusion models (Liu et al., 2023; Heng et al., 2024), or exploiting parameter perturbations (Granz et al., 2024), etc. Several works have also considered multiple independently-trained modes (Xu et al., 2024; Fang et al., 2024a). A loss landscape perspective is proposed in (Fang et al., 2024a) to explain the variations in OoD detection performance across different modes, while distinct supervision tasks are explored in (Xu et al., 2024) to enhance model diversities and improve ensemble detection. Importantly, our mode-path view remains largely underexplored.

Mode connectivity has provided useful insights into the optimization landscape of DNNs in the parameter space. Early studies showed that independently trained modes can be connected by non-linear low-loss curves, such as polynomial chains and Bézier curves (Garipov et al., 2018; Wortsman et al., 2021), while later work further discussed linear paths between modes (Zhou et al., 2023; Zhao et al., 2025; Theus et al., 2025). Besides, mode connectivity has also been applied to other fields such as adversarial robustness (Zhao et al., 2020), machine unlearning (Cheng & Amiri, 2025), and domain generalization (Li et al., 2026). Similar concepts, such as functionally invariant paths in the parameter space, have also been explored in (Raghavan et al., 2024). In this work, we take a preliminary step toward extending mode connectivity to OoD detection, focusing in particular on non-linear connecting paths and their low-dimensional subspace structure.

B More evidence on OoD detection along the mode-connecting path

We provide more empirical evidence to support the observed phenomenon of the varied OoD detection performance along a low-loss path connecting two modes in Sec.3.1.

In experiments, we consider two different pairs of independently-trained modes of ResNet50 on ImageNet-1K: (`mode-1`, `mode-2`) and (`mode-3`, `mode-4`), and two different parameterizations of the connecting path: Bézier curve of Eqn.(3) and polygonal chain of Eqn.(2). For OoD detection, we adopt the Energy score (Liu et al., 2020) on the iNaturalist dataset (Van Horn et al., 2018) and an additional SUN dataset (Xiao et al., 2010). On each path, we sample 20 models and evaluate their accuracy on the ImageNet-1K test dataset and FPR95 on the two OoD datasets. The results are demonstrated in Fig.S1.

As illustrated in Fig.S1, the observed phenomenon in Sec.3.1 remains consistent across different pairs of independent modes, path parameterizations, and other OoD datasets. Along the connecting path, despite nearly identical InD accuracy (70%$\sim$75%), the maximum discrepancy in OoD FPR95 can reach nearly 10 percentage points. Therefore, while all intermediate models on the path maintain consistently high InD accuracy, their OoD detection performance varies substantially with a high variance.

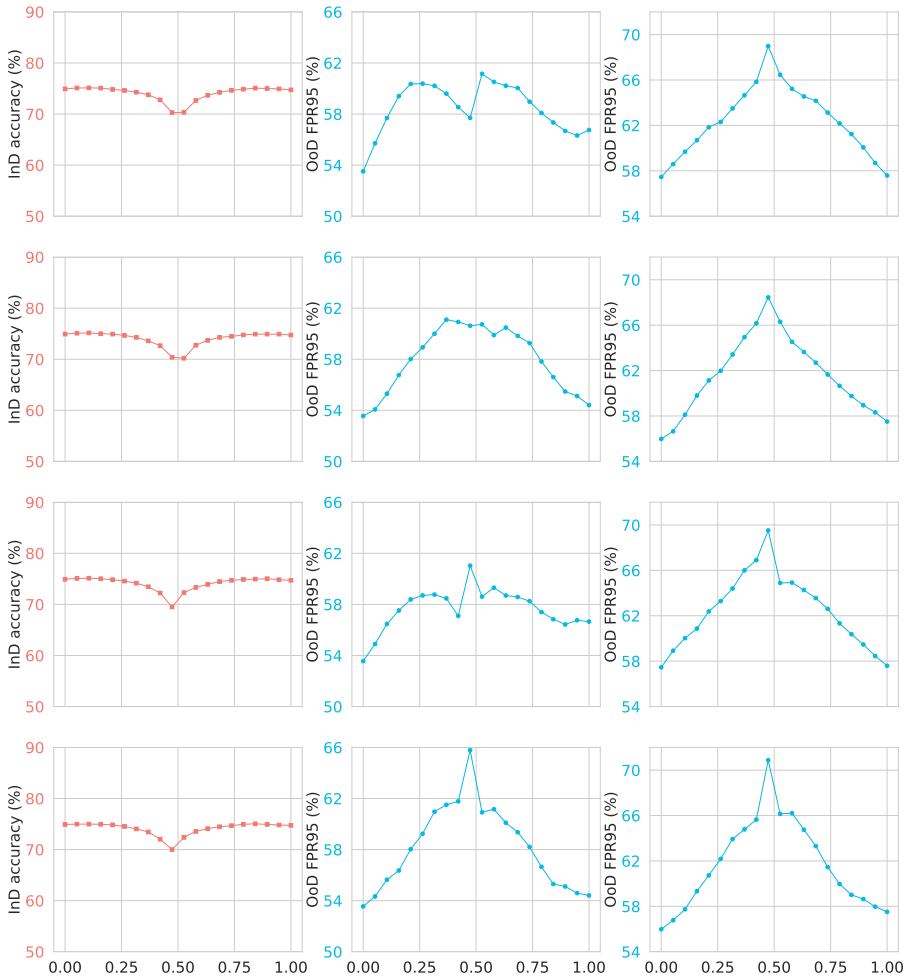


Figure S1: **Columns 1, 2 & 3**: Accuracy on the ImageNet-1K test dataset, FPR95 on the iNaturalist and SUN datasets, respectively. **Rows 1 & 3**: Performance along a Bézier curve and a polygonal chain connecting two endpoint modes (`mode-1` and `mode-2`), respectively. **Rows 2 & 4**: Performance along a Bézier curve and a polygonal chain connecting another two different endpoint modes (`mode-3` and `mode-4`), respectively.

C THEORETICAL DISCUSSIONS

In this section, we provide a unified theoretical understanding on the InD stability and the OoD sensitivity of the principal-mode construction through the spectral geometry of the sampled models.

InD performance. The following Theorem 1 shows that the InD loss of the principal mode $\boldsymbol{\omega}^*$ is controlled by the loss level of the sampled path models $\boldsymbol{\omega}^{(i)}$ and the singular-value spectrum of the parameter matrix $\boldsymbol{\Omega}$.

Theorem 1. *Suppose that the InD losses of the K models $\boldsymbol{\omega}^{(1)}, \cdots, \boldsymbol{\omega}^{(K)}$ sampled from the connecting path are bounded as $\mathcal{L}(\boldsymbol{\omega}^{(i)}) \leq \xi$ with a β-smooth loss function $\mathcal{L}$ on InD data, and that the gradients of these K models are also bounded as $\|\nabla \mathcal{L}(\boldsymbol{\omega}^{(i)})\|_2 \leq G$. The singular values of the centered parameter matrix $\tilde{\boldsymbol{\Omega}}$ are denoted as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ with rank r. The loss of the principal mode $\boldsymbol{\omega}^*$ satisfies*

$$\mathcal{L}(\boldsymbol{\omega}^*) \leq \xi + G\sqrt{\kappa} + \frac{\beta}{2}\kappa, \tag{S1}$$

where $\kappa = \left(\frac{1}{s} + \frac{1}{\sqrt{K}}\right)^2 \sum_{j=1}^{s} \sigma_j^2 + \frac{1}{K} \sum_{j=s+1}^{r} \sigma_j^2$.

Proof. For each sampled model $\boldsymbol{\omega}^{(i)}$, we have:

$$\mathcal{L}(\boldsymbol{\omega}^*) \leq \mathcal{L}(\boldsymbol{\omega}^{(i)}) + \nabla\mathcal{L}(\boldsymbol{\omega}^{(i)})^\top(\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}) + \frac{\beta}{2}\|\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}\|_2^2. \tag{S2}$$

Averaging over the K models, we have:

$$\mathcal{L}(\boldsymbol{\omega}^*) \leq \frac{1}{K}\sum_{i=1}^{K}\mathcal{L}(\boldsymbol{\omega}^{(i)}) + \frac{1}{K}\sum_{i=1}^{K}\nabla\mathcal{L}(\boldsymbol{\omega}^{(i)})^\top(\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}) + \frac{\beta}{2}\frac{1}{K}\sum_{i=1}^{K}\|\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}\|_2^2. \tag{S3}$$

Regarding the first term in the right side of the inequality of (S3), applying $\mathcal{L}(\boldsymbol{\omega}^{(i)}) \leq \xi$, we have $\frac{1}{K}\sum_{i=1}^{K}\mathcal{L}(\boldsymbol{\omega}^{(i)}) \leq \xi$.

Regarding the second term in the right side of the inequality of (S3), applying the Cauchy-Schwarz inequality and the bounded gradients $\|\nabla\mathcal{L}(\boldsymbol{\omega}^{(i)})\|_2 \leq G$, we have:

$$\nabla\mathcal{L}(\boldsymbol{\omega}^{(i)})^\top(\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}) \leq \|\nabla\mathcal{L}(\boldsymbol{\omega}^{(i)})\|_2\|\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}\|_2 \leq G\|\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}\|_2 \tag{S4}$$

Averaging over the K models and applying the Jensen inequality, we have

$$\frac{1}{K}\sum_{i=1}^{K}\nabla\mathcal{L}(\boldsymbol{\omega}^{(i)})^\top(\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}) \leq \frac{1}{K}\sum_{i=1}^{K}G\|\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}\|_2 \leq G\sqrt{\frac{1}{K}\sum_{i=1}^{K}\|\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}\|_2^2}. \tag{S5}$$

We now focus on the term $\frac{1}{K}\sum_{i=1}^{K}\|\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}\|_2^2$.

Recall that the projected matrix is $\boldsymbol{\Omega}_s = \tilde{\boldsymbol{\Omega}}\mathbf{V}_s$, and our solution is:

$$\boldsymbol{\omega}^* = \frac{1}{s}\boldsymbol{\Omega}_s\mathbf{1}_s = \frac{1}{s}\tilde{\boldsymbol{\Omega}}\mathbf{V}_s\mathbf{1}_s = \frac{1}{s}\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top\mathbf{V}_s\mathbf{1}_s = \frac{1}{s}\mathbf{U}\boldsymbol{\Sigma}\begin{bmatrix}\mathbf{I}_s\\ \mathbf{0}\end{bmatrix}\mathbf{1}_s = \frac{1}{s}\sum_{i=1}^{s}\sigma_i\boldsymbol{u}_i. \tag{S6}$$

Similarly, from $\tilde{\boldsymbol{\Omega}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ (recall that $\tilde{\boldsymbol{\Omega}}$ is assumed to be centered in Theorem 1), the i-th column $\boldsymbol{\omega}^{(i)}$ in $\boldsymbol{\Omega}$ can be written as $\boldsymbol{\omega}^{(i)} = \sum_{j=1}^{r}\sigma_j\boldsymbol{v}_{ij}\boldsymbol{u}_j$, and we have

$$\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)} = \frac{1}{s}\sum_{j=1}^{s}\sigma_j\boldsymbol{u}_j - \sum_{j=1}^{r}\sigma_j\boldsymbol{v}_{ij}\boldsymbol{u}_j = \sum_{j=1}^{s}\sigma_j\left(\frac{1}{s} - \boldsymbol{v}_{ij}\right)\boldsymbol{u}_j - \sum_{j=s+1}^{r}\sigma_j\boldsymbol{v}_{ij}\boldsymbol{u}_j. \tag{S7}$$

Note that the singular vectors $\boldsymbol{u}_j$ are mutually-orthogonal, we have

$$\|\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}\|_2^2 = \sum_{j=1}^{s}\sigma_j^2\left(\frac{1}{s} - \boldsymbol{v}_{ij}\right)^2 + \sum_{j=s+1}^{r}\sigma_j^2\boldsymbol{v}_{ij}^2. \tag{S8}$$

Since each right singular vector $\boldsymbol{v}_j$ has unit Euclidean norm, we have $\frac{1}{K}\sum_{i=1}^{K}\boldsymbol{v}_{ij}^2 = \frac{1}{K}$. Moreover, by the Cauchy-Schwarz inequality $|\frac{1}{K}\sum_{i=1}^{K}v_{ij}| \leq \frac{1}{\sqrt{K}}$, we have

$$\begin{aligned}\frac{1}{K}\sum_{i=1}^{K}\|\boldsymbol{\omega}^* - \boldsymbol{\omega}^{(i)}\|_2^2 &= \sum_{j=1}^{s}\sigma_j^2\frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{s} - \boldsymbol{v}_{ij}\right)^2 + \sum_{j=s+1}^{r}\sigma_j^2\frac{1}{K}\sum_{i=1}^{K}\boldsymbol{v}_{ij}^2\\ &\leq \left(\frac{1}{s} + \frac{1}{\sqrt{K}}\right)^2\sum_{j=1}^{s}\sigma_j^2 + \frac{1}{K}\sum_{j=s+1}^{r}\sigma_j^2 \triangleq \kappa.\end{aligned} \tag{S9}$$

Finally, $\mathcal{L}(\boldsymbol{\omega}^*) \leq \xi + G\sqrt{\kappa} + \frac{\beta}{2}\kappa$ with $\kappa = \left(\frac{1}{s} + \frac{1}{\sqrt{K}}\right)^2\sum_{j=1}^{s}\sigma_j^2 + \frac{1}{K}\sum_{j=s+1}^{r}\sigma_j^2$. The proof is finished. □

Theorem 1 establishes a structural relation between the InD loss of the principal mode and its spectrum-dependent distance from the sampled low-loss models, as summarized by κ. Rewriting $\kappa = (\frac{1}{K} + \frac{1}{s^2})\sum_{j=1}^{s}\sigma_j^2 + \frac{1}{K}\sum_{j=s+1}^{r}\sigma_j^2$ shows that rapid spectral decay controls the residual energy outside the retained subspace, but does not alone guarantee that κ is small, since κ also depends on the absolute energy of the retained components. The results should therefore be interpreted as a qualitative stability relation: whenever the spectrum-induced distance κ is small, the smoothness of the InD objective controls the loss increase of $\boldsymbol{\omega}^*$ relative to the sampled low-loss models. The actual preservation of InD performance is verified empirically in our experiments.

OoD behavior. We next provide a conditional local characterization of when lower-variance principal directions can exhibit greater OoD sensitivity. Let $\tilde{\mathbf{\Omega}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the centered parameter matrix of the models sampled along the low-loss connecting path, and define its parameter-space Gram matrix as $\mathbf{G} = \tilde{\mathbf{\Omega}}\tilde{\mathbf{\Omega}}^\top \in \mathbb{R}^{D\times D}$. Suppose that $\mathrm{rank}(\tilde{\mathbf{\Omega}}) = r$. The nonzero eigen-decomposition of $\mathbf{G}$ is $\mathbf{G} = \sum_{j=1}^{r} \lambda_j \boldsymbol{u}_j \boldsymbol{u}_j^\top$, where $\lambda_j = \sigma_j^2$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$. Thus, λ_j measures the total squared parameter variation of the sampled path along $\boldsymbol{u}_j$. Directions with large λ_j are strongly represented by the path, whereas directions with small λ_j constitute its low-variance tail.

Because our empirical perturbation analysis examines principal directions with substantial variance, we restrict the following local characterization to those top-q ones:

$$\mathcal{S}_q = \mathrm{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q\}, \tag{S10}$$

and define the corresponding truncated Gram matrix as

$$\mathbf{G}_q = \sum_{j=1}^{q} \lambda_j \boldsymbol{u}_j \boldsymbol{u}_j^\top. \tag{S11}$$

This restriction avoids extending the local condition below to trailing directions with numerically negligible singular values.

For each $j \leq q$, consider a perturbation of fixed Euclidean norm along $\boldsymbol{u}_j$:

$$\boldsymbol{\delta}_j = \eta \boldsymbol{u}_j, \qquad \|\boldsymbol{\delta}_j\|_2 = |\eta|, \qquad \eta = \frac{\varepsilon}{\sqrt{D-1}}. \tag{S12}$$

Although these perturbations have the same Euclidean norm across directions, their magnitudes relative to the variations exhibited by the path differ. Specifically,

$$\boldsymbol{\delta}_j^\top \mathbf{G}_q^\dagger \boldsymbol{\delta}_j = \frac{\eta^2}{\lambda_j}, \qquad j \leq q, \tag{S13}$$

where $\mathbf{G}_q^\dagger$ is the pseudoinverse of $\mathbf{G}_q$. Hence, within the analyzed subspace, a perturbation along a direction with smaller λ_j represents a greater path-normalized departure from the parameter variations observed among the sampled models.

To relate this geometric departure to OoD behavior, we introduce a sufficient local alignment condition within $\mathcal{S}_q$. Intuitively, the Gram geometry of the sampled models characterizes the parameter variations that are consistently realized by models on the connected low-loss path. These variations are therefore empirically compatible with the shared low-loss and stable InD behavior of the sampled models. In contrast, a perturbation with a larger variance-normalized departure moves the model toward variations that are more weakly represented by the path and hence less constrained by its shared low-loss behavior. Since the training objective primarily constrains InD classification rather than OoD-related quantities such as feature geometry, confidence scores, or gradient statistics, these quantities may change more sharply along such weakly supported directions. This motivates the following explicit modeling hypothesis, which formalizes, rather than derives from the path geometry alone, a local regime in which a larger path-normalized departure is reflected in higher OoD-risk curvature.

Let $\mathcal{R}_{\mathrm{ood}}(\boldsymbol{\omega})$ denote a differentiable scalar surrogate for the OoD detection risk, whose increase corresponds to degraded OoD detection performance, and let $\mathbf{H}_{\mathrm{ood}}$ denote its local Hessian around a reference model $\boldsymbol{\omega}_0$. We formalize the restricted Gram-curvature alignment as

$$\boldsymbol{z}^\top \mathbf{H}_{\mathrm{ood}} \boldsymbol{z} \geq \alpha \boldsymbol{z}^\top \mathbf{G}_q^\dagger \boldsymbol{z}, \qquad \forall\, \boldsymbol{z} \in \mathcal{S}_q, \quad \alpha > 0. \tag{S14}$$

Consequently, for each $j \leq q$,

$$\boldsymbol{u}_j^\top \mathbf{H}_{\mathrm{ood}} \boldsymbol{u}_j \geq \frac{\alpha}{\lambda_j}. \tag{S15}$$

Thus, within the analyzed subspace, a direction with smaller path variance admits a larger inverse-variance lower bound on its local OoD-risk curvature.

To isolate the second-order effect without requiring $\boldsymbol{\omega}_0$ to be stationary for the OoD-risk surrogate, we consider the symmetric local response

$$\Delta\mathcal{R}^{(j)}_{\mathrm{ood,sym}}(\eta) = \frac{1}{2}\left[\mathcal{R}_{\mathrm{ood}}(\boldsymbol{\omega}_0 + \boldsymbol{\delta}_j) + \mathcal{R}_{\mathrm{ood}}(\boldsymbol{\omega}_0 - \boldsymbol{\delta}_j)\right] - \mathcal{R}_{\mathrm{ood}}(\boldsymbol{\omega}_0). \tag{S16}$$

A second-order Taylor expansion around $\boldsymbol{\omega}_0$ gives

$$\Delta\mathcal{R}^{(j)}_{\mathrm{ood,sym}}(\eta) = \frac{1}{2}\boldsymbol{\delta}_j^\top \mathbf{H}_{\mathrm{ood}}\boldsymbol{\delta}_j + o(\eta^2), \tag{S17}$$

where the first-order terms cancel between the two perturbation directions. Under the restricted Gram-curvature alignment condition above, this symmetric local response admits an inverse-variance lower bound within $\mathcal{S}_q$. We summarize this implication in the following Proposition 1.

Proposition 1. *Let $j \le q$ and $\mathbf{G}_q\boldsymbol{u}_j = \lambda_j\boldsymbol{u}_j$, where $\mathbf{G}_q$ is the truncated Gram matrix on $\mathcal{S}_q$ and λ_j is the eigenvalue associated with $\boldsymbol{u}_j$. Let $\boldsymbol{\delta}_j = \eta\boldsymbol{u}_j$, where $\eta = \varepsilon/\sqrt{D-1}$ and $\|\boldsymbol{\delta}_j\|_2 = |\eta|$. Assume that*

$$\boldsymbol{z}^\top \mathbf{H}_{\mathrm{ood}}\boldsymbol{z} \ge \alpha \boldsymbol{z}^\top \mathbf{G}_q^\dagger \boldsymbol{z}, \qquad \forall\, \boldsymbol{z} \in \mathcal{S}_q, \quad \alpha > 0, \tag{S18}$$

where $\mathbf{H}_{\mathrm{ood}}$ is the local Hessian of the OoD-risk surrogate around $\boldsymbol{\omega}_0$. Then, as $\eta \to 0$, the corresponding symmetric local response satisfies

$$\begin{aligned}\Delta\mathcal{R}^{(j)}_{\mathrm{ood,sym}}(\eta) &= \frac{1}{2}\boldsymbol{\delta}_j^\top \mathbf{H}_{\mathrm{ood}}\boldsymbol{\delta}_j + o(\eta^2) \\ &\ge \frac{\alpha\eta^2}{2\lambda_j} + o(\eta^2) \\ &= \frac{\alpha\varepsilon^2}{2(D-1)\lambda_j} + o(\eta^2).\end{aligned} \tag{S19}$$

Consequently, within the analyzed subspace and for perturbations of the same Euclidean norm, directions with smaller singular values admit a larger inverse-variance lower bound on their symmetric local OoD-risk response. This proposition is a conditional local characterization under the stated alignment assumption, rather than a universal consequence of low path variance alone.

D Additional experimental setups and results

We provide detailed setups and complete results for the experiments in the main text.

D.1 Baselines

In experiments, we select 3 representative OoD detection methods, *logits-based* Energy (Liu et al., 2020), *features-based* KPCA (Fang et al., 2026), and *gradients-based* GradPCA (Seleznova et al., 2026), to comprehensively evaluate the performance of our principal mode $\boldsymbol{\omega}^*$. For each method, its ensemble version is achieved by aggregating the corresponding DNN outputs, as suggested in (Fang et al., 2024a). The individual detection methods and their ensemble variants are detailed below.

Notations Let the neural network for a classification task be denoted as $f(\cdot;\boldsymbol{\omega}) : \mathcal{X} \to \mathbb{R}^C$ with parameters $\boldsymbol{\omega} \in \mathbb{R}^D$. The output of $f(\boldsymbol{x};\boldsymbol{\omega})$ is a C-dimensional logit vector corresponding to the C classes. For brevity, we omit $\boldsymbol{\omega}$ and write $f(\boldsymbol{x})$. We further denote the penultimate-layer features extracted from f as $\boldsymbol{z}$.

Energy adopts the energy function on the logits from a DNN, since energy is better aligned with the input probability density. The detection scoring function of Energy is defined as

$$S(\boldsymbol{x}) = \log \sum_{i=1}^{C} \exp(f_i(\boldsymbol{x})), \tag{S20}$$

where $f_i(\boldsymbol{x})$ denotes the i-th element in the C-dimension output logits.

Ensemble on Energy adopts the output logits from K independently-trained DNNs $f^{(1)}, \cdots, f^{(K)}$ with the following detection scoring function:

$$S(\boldsymbol{x}) = \log \sum_{i=1}^{C} \exp\left(\frac{1}{K}\sum_{j=1}^{K} f_i^{(j)}(\boldsymbol{x})\right). \tag{S21}$$

KPCA operates on the DNN features and seeks a non-linear feature subspace to separate InD and OoD data. In experiments, we apply KPCA to the penultimate-layer features $\boldsymbol{z}$ and adopt its Nyström approximation for OoD detection, following the proposed method in (Fang et al., 2026). The detection score is based on the KPCA reconstruction error:

$$e_{\text{rec}}(\boldsymbol{x}) = \|\mathbf{U}_q\mathbf{U}_q^\top \hat{\boldsymbol{z}} - \hat{\boldsymbol{z}}\|_2, \tag{S22}$$

where $\hat{\boldsymbol{z}}$ denotes the centered features, and $\mathbf{U}_q$ is the matrix of the top-q principal directions obtained from the feature covariance matrix of the training data.

Ensemble on KPCA leverages the averaged features across multiple DNNs $f^{(1)}, \cdots, f^{(K)}$. To be specific, the training features extracted from each model are averaged to calculate the covariance matrix, from which the principal directions are derived. During inference, the features of new samples obtained from the K DNNs are also averaged, and the reconstruction error based on the averaged features serves as the detection score.

GradPCA leverages the low-rank structure of gradients to distinguish InD from OoD data. Particularly, GradPCA (Seleznova et al., 2026) operates on the gradients of the last linear layer of a DNN w.r.t. the maximum logit value. A projection matrix $\mathbf{P}$ onto the desired gradient subspace is obtained by applying PCA to the class-wise gradient matrix of the training data. The detection score of GradPCA is then defined as the cosine of the angle between the gradient vector $\boldsymbol{g}$ of a new sample $\boldsymbol{x}$ and its projection onto this subspace:

$$S(\boldsymbol{x}) = \frac{\|\mathbf{P}\boldsymbol{g}\|}{\|\boldsymbol{g}\|}. \tag{S23}$$

Ensemble on GradPCA extends GradPCA by employing an ensemble of K DNNs. To be specific, for a given input, the logits from all K models are first averaged. The scalar output used for gradient computation is then taken as the maximum value of the averaged logits. Subsequently, the same procedure as standard GradPCA is followed. By combining multiple models, this ensemble reduces the variance of gradient estimates and improves the robustness of the low-rank subspace characterization, leading to more reliable OoD detection.

D.2 Detailed setups

Modes In the main results of Table 1, we train 6 independent modes of ResNet50 on ImageNet-1K, and group them into 3 pairs: (`mode-1`, `mode-2`), (`mode-3`, `mode-4`) and (`mode-5`, `mode-6`). For the first 5 modes, we directly adopt the released checkpoints from (Fang et al., 2024a), and additionally train the `mode-6` following the instructions of (Fang et al., 2024a) for a fair setup: The ResNet50 model of `mode-6` is trained in parallel on 4 NVIDIA GTX 4090 GPUs using the stochastic gradient descent (SGD) optimizer on the ImageNet-1K training dataset. The training runs for 90 epochs with an initial learning rate of 0.1, a batch size of 256, a weight decay of 1×10^{-4}, and the learning rate is reduced by 10% every 30 epochs.

Mode-connecting paths We adopt the method in (Garipov et al., 2018) to train Bézier curves and polynomial chains with one intermediate vertex on two endpoint modes. Each training experiment is executed in parallel on 4 NVIDIA GTX 4090 GPUs using the SGD optimizer on the ImageNet-1K training dataset. The training runs for 90 epochs with an initial learning rate of 0.01, a batch size of 256, a weight decay of 3×10^{-4}, and the learning rate follows a cosine decay schedule.

Low-dimensional subspace modeling approach For each mode-connecting path, the default settings of our low-dimensional subspace modeling approach to yield results in Tables 1, 2 and S1 are sampling $K = 5$ models and reducing to $s = 2$ principal directions. Besides, after obtaining the principal mode $\boldsymbol{\omega}^*$, we perform a forward pass over the training set to recalibrate the running statistics of the Batch Normalization layers. The same recalibration is applied to the weight-averaging baseline in Figure 2(c) for a correct implementation.

All the checkpoints of modes, connecting paths, and principal modes will be released.

Table S1: InD recognition and OoD detection results with ResNet50 on ImageNet-1K.

model	InD acc. (%)	OoD datasets								**AVERAGE**	
		iNaturalist		SUN		Places		Texture			
		FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑
Energy (logits-based)											
`mode-1`	74.85	55.84	89.69	58.46	86.55	62.50	84.95	54.02	86.61	57.71	86.95
`mode-2`	74.72	55.66	89.71	57.25	87.30	61.99	85.15	55.51	85.62	57.60	86.94
`ensemble`	**76.43**	52.76	91.07	55.71	87.91	60.31	86.11	49.89	**88.14**	54.67	88.31
Ours (Bézier)	74.16	41.26	91.89	46.12	**89.10**	55.18	**86.13**	**45.46**	87.82	47.01	**88.73**
Ours (PolyChain)	73.43	**37.82**	**92.19**	**44.49**	88.84	**54.63**	85.54	49.61	84.90	**46.64**	87.87
`mode-3`	74.87	56.06	89.60	56.68	87.04	62.58	84.76	51.49	86.74	56.70	87.04
`mode-4`	74.67	57.51	89.27	58.80	86.18	63.58	84.33	55.55	85.19	58.86	86.24
`ensemble`	**76.50**	53.36	90.93	55.60	87.64	60.91	85.63	48.74	**87.94**	54.65	88.03
Ours (Bézier)	74.45	35.78	93.31	46.33	88.69	55.76	85.69	**44.96**	87.47	45.71	88.79
Ours (PolyChain)	74.05	**34.19**	**93.62**	**44.39**	**89.08**	**54.48**	**86.00**	**44.96**	87.30	**44.51**	**89.00**
`mode-5`	74.82	59.28	88.94	58.87	86.34	62.73	84.71	54.66	85.94	58.89	86.48
`mode-6`	75.65	48.35	90.99	55.63	87.45	61.57	85.42	51.22	86.93	54.19	87.70
`ensemble`	**77.07**	49.72	91.57	54.88	88.01	59.87	86.23	48.55	88.44	53.25	88.56
Ours (Bézier)	75.52	**28.78**	**94.44**	**39.60**	**90.35**	**49.72**	**87.75**	**36.84**	**89.66**	**38.74**	**90.55**
Ours (PolyChain)	73.45	39.53	92.14	47.14	87.97	58.56	84.27	45.46	86.36	47.67	87.68
KPCA (features-based)											
`mode-1`	74.85	35.59	92.61	45.13	89.69	58.32	84.77	**14.84**	**96.84**	38.47	90.98
`mode-2`	74.72	44.58	89.75	50.84	87.49	63.17	82.50	16.42	96.61	43.75	89.09
`ensemble`	**76.43**	36.77	92.60	40.15	91.17	53.95	86.61	20.57	95.94	37.86	91.58
Ours (Bézier)	74.16	**29.47**	**94.79**	24.29	94.41	39.90	90.41	19.43	96.29	**28.27**	**93.98**
Ours (PolyChain)	73.43	34.72	93.83	**22.53**	**94.91**	**37.38**	**90.98**	28.30	94.44	30.73	93.54
`mode-3`	74.87	38.50	91.71	46.02	89.21	60.73	84.00	**13.90**	**97.15**	39.79	90.52
`mode-4`	74.67	36.99	92.07	46.53	89.13	59.95	84.15	15.69	96.86	39.79	90.55
`ensemble`	**76.50**	28.76	94.25	37.92	91.29	53.37	86.75	18.01	96.32	34.52	92.15
Ours (Bézier)	74.45	**25.23**	95.34	26.78	94.19	43.17	89.77	18.26	96.33	28.36	93.91
Ours (PolyChain)	74.05	26.02	**95.42**	**20.37**	**95.67**	**36.16**	**91.84**	23.97	95.36	**26.63**	**94.57**
`mode-5`	74.82	41.61	91.34	48.47	88.08	61.60	82.99	**16.37**	96.52	42.01	89.73
`mode-6`	75.65	38.99	92.01	38.76	91.79	51.32	87.73	21.47	95.76	37.64	91.83
`ensemble`	**77.07**	34.36	93.19	35.86	92.12	49.48	87.82	18.78	96.39	34.62	92.38
Ours (Bézier)	75.52	**25.07**	**95.52**	**21.49**	**95.23**	**35.31**	**91.91**	16.63	**96.81**	**24.63**	**94.87**
Ours (PolyChain)	73.45	36.67	93.50	25.11	94.32	41.68	89.79	25.98	95.07	32.36	93.17
GradPCA (gradients-based)											
`mode-1`	74.85	58.85	90.29	77.23	80.85	80.34	78.33	42.00	91.88	64.61	85.34
`mode-2`	74.72	57.78	90.50	80.02	80.06	82.39	77.34	44.13	91.47	66.08	84.84
`ensemble`	**76.43**	**52.30**	**92.12**	64.54	86.14	71.66	82.68	**42.00**	**92.56**	57.63	88.38
Ours (Bézier)	74.16	53.64	90.85	**52.77**	86.99	**59.95**	83.95	42.29	91.59	**52.16**	88.35
Ours (PolyChain)	73.43	61.40	89.99	53.10	**88.11**	61.88	**84.78**	50.73	91.28	56.78	**88.54**
`mode-3`	74.87	59.55	90.67	79.09	80.50	81.57	77.67	41.61	91.99	65.46	85.21
`mode-4`	74.67	56.22	90.93	82.49	78.16	85.72	75.11	42.06	91.70	66.62	83.98
`ensemble`	**76.50**	**48.65**	**92.45**	81.27	80.79	86.19	77.25	**38.26**	**92.90**	63.59	85.85
Ours (Bézier)	74.45	57.02	90.90	58.60	85.47	67.12	81.90	43.92	91.92	56.67	87.55
Ours (PolyChain)	74.05	55.96	90.90	**53.29**	**87.29**	**62.16**	**83.86**	43.85	91.86	**53.81**	**88.48**
`mode-5`	74.82	56.37	90.87	78.12	81.24	80.81	78.48	42.13	91.89	64.36	85.62
`mode-6`	75.65	50.97	91.48	52.18	87.61	58.08	85.05	35.67	92.76	49.23	89.22
`ensemble`	**77.07**	**44.42**	**92.66**	**50.05**	88.53	**56.80**	85.85	**33.14**	**93.34**	**46.10**	**90.10**
Ours (Bézier)	75.52	55.46	91.27	50.89	**88.94**	58.69	**86.20**	37.93	92.80	50.74	89.80
Ours (PolyChain)	73.45	59.24	90.73	54.60	87.72	63.35	84.36	50.44	91.69	56.91	88.63

D.3 Complete detection results

Table S1 provides the full results corresponding to Tables 1 and 2, including the performance of each mode pair and the corresponding principal modes under the two path parameterizations: the Bézier curve and the polygonal chain.

Notably, `mode-6` achieves substantially better GradPCA detection performance than the other five modes, making it an atypical outlier and consequently producing an unusually strong ensemble baseline for its corresponding mode pair. To further evaluate the general performance under GradPCA, we conduct additional experiments on three different mode pairs. Averaged across these pairs, the ensemble obtains an FPR95 of $63.81\%_{\pm 3.57\%}$ and an AUROC of $86.26\%_{\pm 0.96\%}$, whereas our principal mode achieves $55.43\%_{\pm 2.59\%}$ and $88.12\%_{\pm 0.37\%}$, respectively. These results suggest that the exceptional result associated with `mode-6` is an isolated case.

Results on CIFAR10 Given that the large-scale and challenging ImageNet-1K benchmark already serves as a strong baseline, we include additional experiments on CIFAR-10 (Krizhevsky, 2009)

Table S2: InD recognition and OoD detection results with ResNet18 on CIFAR10.

model	InD acc. (%)	OoD datasets SVHN FPR↓	SVHN AUROC↑	iSUN FPR↓	iSUN AUROC↑	LSUN FPR↓	LSUN AUROC↑	Texture FPR↓	Texture AUROC↑	Places FPR↓	Places AUROC↑	**AVERAGE** FPR↓	**AVERAGE** AUROC↑
`mode-1`	94.76	43.05	90.85	33.57	94.10	20.00	96.03	45.74	89.85	43.45	89.74	37.16	92.11
`mode-2`	94.87	42.85	91.54	38.82	92.26	18.00	96.50	41.26	90.48	40.53	90.28	36.29	92.21
`ensemble`	**95.36**	38.55	93.57	**29.13**	**95.04**	16.51	96.91	38.65	92.20	38.08	91.55	32.19	93.85
Ours (Bézier)	94.62	13.81	97.42	36.95	94.28	3.22	99.29	**33.69**	**92.45**	27.16	**94.71**	22.97	**95.63**
Ours (PolyChain)	94.51	**10.79**	**97.77**	37.99	94.07	**2.78**	**99.38**	34.01	92.27	**27.00**	94.69	**22.52**	**95.63**
`mode-3`	94.79	44.76	92.14	48.26	90.27	18.30	96.49	46.91	88.96	43.16	89.84	40.28	91.54
`mode-4`	94.82	39.44	91.86	26.85	95.10	19.27	96.22	49.66	88.44	43.91	89.71	35.83	92.27
`ensemble`	**95.45**	37.12	93.34	**25.65**	**95.67**	15.49	97.03	42.41	90.85	38.69	91.26	31.87	93.63
Ours (Bézier)	94.61	19.58	96.76	35.69	93.94	7.79	98.54	39.49	**91.15**	32.58	93.81	27.03	**94.84**
Ours (PolyChain)	94.69	**17.87**	**96.89**	38.76	93.26	**7.46**	**98.59**	**39.04**	90.93	**31.69**	**93.89**	**26.96**	94.71
`mode-5`	94.55	52.12	89.59	33.12	93.75	21.19	96.16	49.27	88.32	43.59	89.22	39.86	91.41
`mode-6`	94.64	37.25	93.84	36.27	93.11	24.15	95.42	44.72	89.92	42.46	89.77	36.97	92.41
`ensemble`	**95.27**	28.56	95.13	25.78	95.37	18.60	96.56	40.18	91.49	38.03	91.08	30.23	93.93
Ours (Bézier)	94.32	7.22	**98.47**	21.34	96.30	4.21	99.09	32.20	93.15	30.95	93.98	19.19	96.20
Ours (PolyChain)	94.39	**7.02**	98.46	**19.74**	**96.50**	**3.45**	**99.18**	**30.44**	**93.49**	**30.67**	**93.99**	**18.27**	**96.32**

and more OoD datasets for further validation, though these are not as comprehensive as the main ImageNet-1K evaluation. To be specific, for CIFAR10 as InD, we evaluate on 5 OoD datasets including SVHN (Netzer et al., 2011), iSUN (Xu et al., 2015), LSUN (Yu et al., 2015), Texture (Cimpoi et al., 2014) and Places (Zhou et al., 2017), using ResNet18 following the settings in (Sun et al., 2022). Similarly, we train 6 independent modes with the SGD optimizer on the CIFAR10 training set on one NVIDIA GTX 4090 GPU. Each mode is trained for 100 epochs with an initial learning rate of 0.1, a batch size of 256, a weight decay of 5×10^{-4}, and a cosine decay learning rate schedule. Besides, we train 3 Bézier curves based on the 6 modes, each for 100 epochs with an initial learning rate of 0.01, the same batch size and weight decay, and also a cosine decay schedule. The low-dimensional subspace modeling approach for $\boldsymbol{\omega}^*$ follows that on ImageNet-1K, i.e., sampling $K = 5$ models and reducing to $s = 2$ principal directions.

Table S2 demonstrates results on CIFAR10 with InD recognition and OoD detection under the Energy score. As demonstrated, our principal mode achieves stronger detection results than the mode ensembling, and simultaneously maintains effective InD accuracy. Regarding the path parameterization, PolyChain yields slightly lower average FPR95 than the Bézier curve across the three mode pairs, while their InD accuracy and AUROC remain comparable. Together with the ImageNet-1K results in Table 2, where the Bézier curve performs better across the evaluated detectors, these results indicate that neither parameterization is uniformly superior and that their relative performance depends on the experimental setting.

D.4 Computational cost

The full pipeline of producing the principal mode $\boldsymbol{\omega}^*$ consists of three stages: `(a)` training two independent modes; `(b)` learning a connecting path between them; and `(c)` constructing the principal mode from the path according to Sec.3.2. In this section, we record and analyze the computational cost of each stage. Experiments are executed on the same machine equipped with 4 NVIDIA GTX 4090 GPUs for a fair comparison, using the ImageNet-1K dataset with ResNet50. The reported runtime is averaged over multiple training epochs.

The required GPU hours per epoch in stages `(a)` and `(b)` are shown in Table S3. After stage `(b)`, constructing the principal mode of stage `(c)` incurs only a small additional overhead. Specifically, our method samples models along the learned path, averages leading singular components from SVD, and finally performs a single forward pass over the training set to recompute statistics of the batch normalization layers. This stage `(c)` takes *nearly 3 minutes*, which is negligible compared with the cost of stages `(a)` and `(b)`.

Table S3: Computational cost of stages `(a)` and `(b)` with ResNet50 on ImageNet-1K.

stage	GPU hours per epoch	total epochs
`(a)`	$\approx 0.30 \times 2$	90×2
`(b)`	≈ 0.14	90

Overall, the dominant training cost of our method comes from learning the connecting path between two independently-trained modes, rather than constructing the principal mode itself. Nevertheless,

our method produces *a single principal mode*, whose inference cost (runtime, memory, and computational complexity) is identical to that of a standard single model. Consequently, it is directly compatible with existing post-hoc OoD detection methods without introducing any additional inference overhead. Besides, compared with standard ensembles, the proposed principal mode requires only a single forward pass during deployment, whereas ensemble methods incur inference costs that grow linearly with the number of constituent models. More broadly, we view this work as an exploration of how parameter-space geometry can be exploited for OoD detection, and hope it provides useful insights for future research.