WHAT DRIVES THE INLIER-MEMORIZATION EFFECT? A THEORY OF OUTLIER DETECTION VIA EARLY TRAINING DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Outlier detection (OD) aims to identify anomalous instances by learning the underlying structure of normal data (inliers), and is particularly challenging in fully unsupervised settings where no information about anomalies is available during training. Recent advances have leveraged the inlier-memorization (IM) effect, a phenomenon in which deep models memorize inlier patterns earlier than those of outliers, as a powerful signal for distinguishing outliers. However, despite its empirical success, the theoretical understanding of the IM effect remains limited. In this work, we present a theoretical study of the IM effect. Focusing on a simple autoencoder with a fixed decoder and an encoder trained by full-batch gradient descent, we show that, under suitable regularity conditions, the IM effect emerges over a range of training iterations. Our results characterize when such an interval exists and which outliers can be separated from inliers throughout the interval. Further, we characterize how data distribution and weight initialization affect the strength and persistence of the IM effect, and provide empirical support for these theoretical predictions through simulation studies. Guided by these insights, we study compact representations and EMA-based initialization as two practical strategies for enhancing the IM effect. Experiments on real-world data further evaluate their individual and combined effects on the performance of IM-based outlier detection methods.

1 INTRODUCTION

Outlier detection (OD) aims to identify anomalous instances (outliers) that deviate from the underlying structure of normal data (inliers) (Chandola et al., 2009; Chalapathy & Chawla, 2019). It plays a crucial role in a wide range of applications, including fraud detection (Hilal et al., 2022), cybersecurity (Ahmed et al., 2016), medical diagnosis (Fernando et al., 2021), and data preprocessing for downstream learning tasks. Among various settings, unsupervised outlier detection (UOD), where both inliers and outliers are contained in the training data and no information about anomalousness is accessible, is the most general and particularly challenging setting.

To address this issue, a recent line of work exploits the *inlier-memorization (IM) effect* (Kim et al., 2024): when a deep unsupervised model is trained on data containing both inliers and outliers, it prioritizes reducing the loss of inliers before that of outliers in early training steps. Thus, the per-sample loss of an under-fitted model effectively serves as an outlier score, and this insight motivates various methodologies (Kim et al., 2024; Huang et al., 2024; Zhang et al., 2024; Cho et al., 2025; Kang et al., 2026). These approaches show strong empirical performance across diverse datasets, highlighting the practical importance of the IM effect in UOD.

Despite these successes, the IM effect itself remains largely underexplored from a theoretical perspective. Existing works primarily rely on empirical observations or heuristic arguments (for example, attributing the IM effect to the higher density and prevalence of inliers in the data distribution), but lack a rigorous understanding of how long the effect persists and how strong the resulting separation is, i.e., its persistence and strength, and how these depend on key factors such as the data distribution and model initialization. This gap limits both the interpretability of the IM effect and the principled design of more reliable IM-based outlier detection methods.

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.

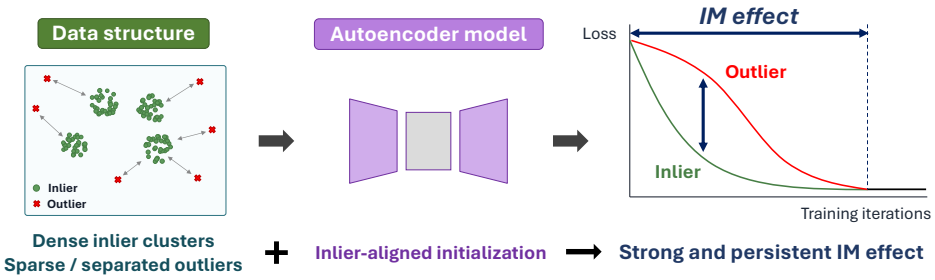


Figure 1: *Inlier-memorization (IM) effect and the factors that influence its strength (how large the loss gap is) and persistence (how long the loss gap persists).*

To understand the IM effect theoretically, we study how the reconstruction errors of inliers and outliers evolve during autoencoder training. Specifically, we address the following two questions: *(i) when does a training interval exhibiting the IM effect exist, and (ii) which outliers can be reliably distinguished from inliers throughout this interval?* To rigorously answer these questions, we show that, under suitable regularity conditions, the IM effect emerges over a nonempty interval of training iterations, and characterize which outliers can be distinguished from inliers throughout this interval in terms of their geometric separation from the inlier cluster structure. We then examine the implications of these conditions to determine how data geometry and initialization influence the strength and persistence of the IM effect. We use controlled simulations to test these theoretical implications. Figure 1 provides a schematic overview of the phenomenon and its governing factors.

Building on these theoretical and empirical insights, we study two simple, theory-motivated modifications to existing IM-based methods. First, we examine whether embedding representations can induce more favorable data geometry and thereby strengthen the IM effect. Second, we study whether EMA-based initialization can favor the learning of inlier structure during early training and thereby prolong the IM effect. These modifications can be incorporated into existing methods with only minimal changes. Experiments on real-world data evaluate them individually and in combination, showing that their effects on detection performance are dataset-dependent. Despite this variability, the two modifications improve average performance overall, with the best combined variant achieving the highest average AUROC and AUPRC among all methods compared across 57 benchmarks.

Our contributions can be summarized as follows:

- ◇ We show that, under suitable regularity conditions, the IM effect emerges over a nonempty interval, characterize the outliers separated within it, and analyze how data geometry and initialization affect its strength and persistence, with simulations supporting these qualitative predictions.
- ◇ Inspired by the theoretical analysis, we study compact representations and EMA-based initialization as simple modifications to existing IM-based detectors and evaluate their empirical effects on detection performance.

2 Related Work

Recent advances have shown that deep learning provides a powerful framework for UOD. Classical approaches include density-based methods such as LOF (Breunig et al., 2000), cluster-based methods such as CBLOF (He et al., 2003), robust statistics-based approaches such as MCD (Fauconnier & Haesbroeck, 2009), and tree-based methods such as Isolation Forest (IF) (Liu et al., 2008). Boundary-based methods such as OCSVM (Schölkopf et al., 2001) and SVDD (Tax & Duin, 2004) learn a separating boundary for normal data, with deep extensions including DeepSVDD (Ruff et al., 2018) and DeepSAD (Ruff et al., 2020).

Self-supervised learning has also been widely applied to OD tasks (Tack et al., 2020; Golan & El-Yaniv, 2018). Representative approaches learn representations by contrastive training (Chen et al., 2020), by maximizing mutual information between masked and unmasked inputs (Shenkar & Wolf, 2022), or by learning neural transformations that generate diverse views (Qiu et al., 2021).

A wide range of deep learning techniques with diverse modeling perspectives have been applied to UOD. For example, RDA (Zhou & Paffenroth, 2017) enhances autoencoder-based outlier detection

by improving robustness to outliers during representation learning, while DSEBM (Zhai et al., 2016) detects anomalies using deep energy-based models that assign low energy to normal samples and high energy to anomalous ones. AnoGAN (Schlegl et al., 2017) employs generative adversarial networks to learn the manifold of normal data, and GAAL (Liu et al., 2019) generates informative potential outliers to better define the decision boundary. More recently, DTE (Livernoche et al., 2023) leverages diffusion-based models for likelihood-based detection, and ROBOD (Ding et al., 2022) improves robustness via a hyper-ensemble framework.

Recently, the IM effect has attracted attention in the outlier detection literature. Following its introduction in ODIM (Kim et al., 2024), several works have developed methods based on this phenomenon. For instance, ALTBI (Cho et al., 2025) enhances the IM effect via adaptive mini-batch scheduling and trimmed loss optimization, while IMBoost (Kang et al., 2026) improves performance by incorporating active learning. Relatedly, EntropyStop (Huang et al., 2024) and GradStop (Zhang et al., 2024) exploit training dynamics for label-free early stopping, using loss entropy and gradient cohesion, respectively. These studies provide insights into how training dynamics relate to outlier detection performance, but their focus is on stopping criteria rather than a theoretical characterization of the IM effect. A rigorous understanding of when the IM effect emerges, how long it persists, and which outliers can be separated throughout this period therefore remains limited.

3 A Theoretical Analysis of the Inlier-Memorization Effect

3.1 Preliminaries and Problem Setup

Inlier-Memorization Effect. When training deep unsupervised models on datasets containing both inliers and outliers, it is often observed that the model tends to fit inlier samples earlier than outliers (Kim et al., 2024; Huang et al., 2024; Zhang et al., 2024). Consequently, after a certain number of training iterations, the reconstruction or likelihood errors of inliers become small, whereas those of outliers remain relatively large. We refer to this phenomenon as the *inlier-memorization (IM) effect*. Intuitively, inliers are frequent and share a common structure, so lowering their loss first is the more effective direction early in training. Importantly, the IM effect is a general phenomenon arising during the training of deep unsupervised models, rather than being specific to a particular architecture. It has been empirically observed across a range of models, including autoencoders, variational autoencoders (VAEs, Kingma & Welling, 2013; Rezende et al., 2014), and normalizing flows (Kobyzev et al., 2020). For theoretical simplicity, we focus our analysis on a simple autoencoder. Extending the analysis to other generative models is a promising direction for future work.

Data Geometry. Suppose that we observe n unlabeled samples $\{x_i\}_{i=1}^n \subset \mathbb{R}^p$ with $\|x_i\|_2 \le 1$. We write $[m] := \{1, \dots, m\}$ for a positive integer m. Let $\mathcal{I} \subset [n]$ and $\mathcal{O} = [n] \setminus \mathcal{I}$ denote the (unobserved) inlier and outlier index sets, respectively. Let $\mu_1, \dots, \mu_K \in \mathbb{R}^p$ be the cluster centers. For each $i \in [n]$, define $g(i) := \operatorname{argmin}_{k\in[K]} \|x_i - \mu_k\|_2$. We write $\widetilde{x}_i := \mu_{g(i)}$ for the cluster center assigned to x_i and $e_i := x_i - \widetilde{x}_i$ for its deviation from that center. We then define the nearest-center distance as $d_i := \|e_i\|_2 = \|x_i - \widetilde{x}_i\|_2$. For each $k \in [K]$, let $\Lambda_k := \{i : g(i) = k\}$ and $n_k := |\Lambda_k|$, and assume $n_k > 0$. We write $n_{\min} := \min_k n_k$ and $n_{\max} := \max_k n_k$. Let $\delta := \min_{a\neq b} \|\mu_a - \mu_b\|_2$ denote the minimum separation between cluster centers. For the inlier index set $\mathcal{I}$, define $\varepsilon_c := \max_{i\in\mathcal{I}} d_i$, which is the inlier radius. We further define $\rho := \max_{k\in[K]} |\mathcal{O} \cap \Lambda_k|/n_k \in [0, 1]$, which is the largest contamination fraction, or outlier ratio, among the nearest-center clusters.

Model Setup. We consider a bias-free single-hidden-layer autoencoder $F_W : \mathbb{R}^p \to \mathbb{R}^p$ defined as

$$F_W(x) := A\phi(Wx), \tag{1}$$

where $W \in \mathbb{R}^{H\times p}$ is the *trainable* encoder weight matrix (H denotes the number of hidden nodes), $A \in \mathbb{R}^{p\times H}$ is the *fixed* decoder weight matrix, and ϕ is an activation function. Following standard autoencoder training (Hinton & Salakhutdinov, 2006), we minimize the reconstruction error:

$$\mathcal{L}(W) := \frac{1}{2}\sum_{i=1}^{n} \|x_i - F_W(x_i)\|_2^2 = \frac{1}{2}\|X - \mathcal{F}_W(X)\|_2^2 \tag{2}$$

where $X := [x_1^\top, \dots, x_n^\top]^\top \in \mathbb{R}^{np}$ and $\mathcal{F}_W(X) := [F_W(x_1)^\top, \dots, F_W(x_n)^\top]^\top \in \mathbb{R}^{np}$. For theoretical simplicity, only the encoder W is trained, while the decoder A remains fixed. We note

that such a simplification is commonly adopted in theoretical analyses of single-hidden-layer neural networks (Arora et al., 2019; Oymak & Soltanolkotabi, 2019; Li et al., 2020a).

Training via Gradient Descent. We minimize the objective function eq. (2) using gradient descent (GD) with initialization $W_0 \in \mathbb{R}^{H\times p}$ and learning rate $\eta > 0$. Let $w := \mathrm{vec}(W) \in \mathbb{R}^{Hp}$ denote the vectorized parameter, and we use W and w interchangeably when no confusion arises. Starting from $w_0 = \mathrm{vec}(W_0)$, GD updates w_τ at each iteration $\tau \in \mathbb{Z}_{\geq 0}$ according to

$$w_{\tau+1} = w_\tau - \eta \nabla_w \mathcal{L}(W_\tau) = w_\tau - \eta J(W_\tau)^\top r_\tau,$$

where $J(W_\tau) := \frac{\partial}{\partial w}\mathcal{F}_{W_\tau}(X) \in \mathbb{R}^{np\times Hp}$ and $r_\tau := \mathcal{F}_{W_\tau}(X) - X \in \mathbb{R}^{np}$ denote the Jacobian of the model outputs with respect to w and the concatenated residual at iteration τ, respectively.

3.2 Theoretical Analysis

Outline. We establish an interval during which every reconstruction remains close to its assigned cluster center. This center-tracking behavior gives rise to the IM effect: because inliers lie close to their assigned centers, they have small reconstruction errors, whereas outliers sufficiently far from their assigned cluster centers retain larger reconstruction errors. This center-tracking behavior may hold only during an intermediate stage of training, since continued training may increasingly fit sample-specific deviations rather than shared cluster structure.

To formalize this mechanism, our argument proceeds in two steps. Let $\widetilde{X} := [\widetilde{x}_1^\top, \ldots, \widetilde{x}_n^\top]^\top \in \mathbb{R}^{np}$ denote the stacked *cluster-centered* inputs. We first analyze gradient descent on the reference objective $\widetilde{\mathcal{L}}(\widetilde{W}) := \frac{1}{2}\|\mathcal{F}_{\widetilde{W}}(\widetilde{X}) - X\|_2^2$, which uses the assigned cluster centers as inputs while retaining the original observations X as reconstruction targets. Let $\widetilde{W}_\tau$ denote the resulting reference trajectory, initialized at $\widetilde{W}_0 = W_0$. We then show that the actual trajectory W_τ remains close to $\widetilde{W}_\tau$ for a controlled number of iterations. This comparison transfers the assigned-center proximity guarantee from the reference dynamics to the observed-data dynamics, yielding a bound that combines the reference-dynamics error with the discrepancy between the two trajectories.

This argument first yields Theorem 3.1, which establishes an interval $\mathcal{T}_\nu$ over which every reconstruction remains closer to its assigned center than to any other center. Within this interval, Theorem 3.2 shows that any outlier far from the inlier cluster structure has a larger reconstruction error than every inlier, while Corollary 3.3 replaces the iteration-dependent separation threshold with a single geometric condition valid throughout $\mathcal{T}_\nu$.

Main Results. Let $\nu_0 := \max_{k\in[K]} \|\frac{1}{n_k}\sum_{i\in\Lambda_k}(x_i - \mu_k)\|_2$ denote the largest cluster-wise mean deviation, and fix $\nu \geq \nu_0$. For $\tau \geq t_\nu$, define the reconstruction-to-center bound

$$B_\nu(\tau) := a + \mathfrak{D}(\tau) + \nu(1 + q^{(\tau - t_\nu)/2}).$$

Here, t_ν denotes the iteration at which the cluster-centered dynamics reach the residual scale ν, and $q \in (0,1)$ is their contraction factor. The constant a bounds the error caused by replacing each input with its assigned center, $\mathfrak{D}(\tau)$ accounts for the drift between the actual and cluster-centered trajectories, and the final term bounds the distance of the reference reconstructions from their assigned centers. Because the drift term increases while the reference term decreases, their trade-off creates an intermediate interval in which $B_\nu(\tau) < \delta/2$, under the conditions in Appendix A.3.2.

Theorem 3.1 (Interval of Cluster-Consistent Reconstructions)**.** *Under the regularity, step-size, and interval conditions stated in Appendix A.3.2, there exists a nonempty interval $\mathcal{T}_\nu$ of consecutive training iterations such that every $\tau \in \mathcal{T}_\nu$ satisfies*

$$\max_{i\in[n]} \|F_{W_\tau}(x_i) - \widetilde{x}_i\|_2 \leq B_\nu(\tau) < \frac{\delta}{2}.$$

Thus, throughout $\mathcal{T}_\nu$, every reconstruction remains within $\delta/2$ of its assigned center. Since any two distinct centers are separated by at least δ, the assigned center $\widetilde{x}_i$ is the unique nearest center to $F_{W_\tau}(x_i)$ for every $i \in [n]$ and $\tau \in \mathcal{T}_\nu$. We therefore refer to $\mathcal{T}_\nu$ as the *cluster-consistent interval*. The conditions in Appendix A.3.2 further clarify how the data geometry, initialization, and optimization dynamics govern the existence and duration of this interval.

Cluster consistency alone, however, does not establish the IM effect, because it does not compare the reconstruction errors of inliers and outliers. The next result shows how compact inlier geometry and sufficient outlier separation turn cluster consistency into the IM effect.

Theorem 3.2 (Inlier–Outlier Reconstruction-Error Separation). *Suppose that the bound $\|F_{W_\tau}(x_i) - \widetilde{x}_i\|_2 \le B_\nu(\tau)$ in Theorem 3.1 holds for every $i \in [n]$ at iteration τ. Then every outlier $i \in \mathcal{O}$ satisfying $d_i > 2B_\nu(\tau) + \varepsilon_c$ has a larger reconstruction error than every inlier:*

$$\max_{j\in\mathcal{I}} \|x_j - F_{W_\tau}(x_j)\|_2 < \|x_i - F_{W_\tau}(x_i)\|_2.$$

At a fixed iteration τ, the reconstruction error of any inlier is at most $\varepsilon_c + B_\nu(\tau)$, whereas that of an outlier i is at least $d_i - B_\nu(\tau)$. This yields the iteration-dependent separation condition $d_i > 2B_\nu(\tau) + \varepsilon_c$ in Theorem 3.2. Since $2B_\nu(\tau) < \delta$ throughout $\mathcal{T}_\nu$, the simpler, iteration-independent condition $d_i > \delta + \varepsilon_c$ guarantees separation over the entire interval, leading to the following corollary.

Corollary 3.3 (Uniform separation). *For the interval $\mathcal{T}_\nu$ in Theorem 3.1, every outlier $i \in \mathcal{O}$ satisfying $d_i > \delta + \varepsilon_c$ has a larger reconstruction error than every inlier at every $\tau \in \mathcal{T}_\nu$:*

$$\max_{j\in\mathcal{I}} \|x_j - F_{W_\tau}(x_j)\|_2 < \|x_i - F_{W_\tau}(x_i)\|_2.$$

Thus, the condition $d_i > \delta + \varepsilon_c$ identifies outliers whose reconstruction errors remain uniformly larger than those of all inliers throughout the cluster-consistent interval.

Remark 3.1 (Averaging Scores Across Training Iterations). *In the unsupervised setting, we know neither the optimal stopping iteration nor which outliers satisfy $d_i > 2B_\nu(\tau) + \varepsilon_c$ at any given iteration. This motivates aggregating reconstruction errors over a range of training iterations. For any finite, nonempty set $\mathcal{S} \subseteq \mathbb{Z}_{\ge 0}$ of training iterations, define $s_i(\mathcal{S}) := \frac{1}{|\mathcal{S}|}\sum_{\tau\in\mathcal{S}} \|x_i - F_{W_\tau}(x_i)\|_2^2$. Ideally, $\mathcal{S}$ would be contained in, or at least overlap substantially with, $\mathcal{T}_\nu$, but this interval is unknown in practice. Iterations in $\mathcal{S} \cap \mathcal{T}_\nu$ at which the separation condition holds contribute positive margins. When these dominate the remaining contributions, averaging retains the separation while reducing sensitivity to the choice of a single stopping iteration. This can be interpreted as a theoretical explanation for the within-model score ensemble used by ALTBI (Cho et al., 2025).*

Remark 3.2 (Possible Extensions). *For theoretical tractability, we analyze an autoencoder in which only the encoder is trained while the decoder is fixed. Existing work on jointly trained autoencoders (Nguyen et al., 2021) and VAE training dynamics (Wang & Huang, 2025) suggests possible routes toward extending our results to end-to-end autoencoder training and broader generative models, respectively. Another direction is to extend the analysis from full-batch gradient descent to mini-batch SGD and adaptive-learning-rate optimizers such as Adam (Kingma & Ba, 2014). Although joint encoder–decoder training with mini-batch Adam lies outside our theoretical setting, simulations under this practical setting exhibit qualitatively similar behavior, as shown in Figures B.1 to B.3 and Table B.2. We leave formal theoretical extensions to these settings for future work.*

3.3 Impact of Data Geometry and Initialization on the IM Effect

Recall that $B_\nu(\tau)$ combines a time-independent input-replacement term, an increasing trajectory-drift term, and a decreasing reference-error term. The duration over which $B_\nu(\tau) < \delta/2$ governs the persistence of the IM effect, whereas the separation threshold $2B_\nu(\tau) + \varepsilon_c$ at each iteration governs its strength. We next summarize these dependencies; formal details are provided in Appendix A.3, and their predicted qualitative effects are examined through controlled simulations in Section 4.1.

Persistence of the IM Effect. The persistence of the IM effect corresponds to the length of $\mathcal{T}_\nu$: the effect lasts longer when $B_\nu(\tau)$ falls below $\delta/2$ earlier and remains below it for more iterations. The bound highlights four factors that favor a longer interval. (i) A smaller ε_c makes the inlier clusters more compact and tightens the bound on the cluster-wise mean deviation ν_0, potentially allowing a smaller reference-error level ν. (ii) More balanced cluster sizes increase $n_{\min}$ and decrease $n_{\max}$, promoting faster reference contraction and tighter trajectory-drift bounds. (iii) A smaller ρ limits the contribution of outliers to the cluster-wise mean deviation and can likewise permit a smaller ν. (iv) A smaller $\widetilde{R}_0 := \|\mathcal{F}_{W_0}(\widetilde{X}) - X\|_2$, which measures the initial reconstruction error under cluster-centered inputs, can lead to an earlier entry time t_ν and a tighter drift bound. Taken together, these factors tend to widen $\mathcal{T}_\nu$, although their precise effects interact through t_ν and $B_\nu(\tau)$.

Strength of the IM Effect. At iteration τ, an outlier is guaranteed to be separated from every inlier when $d_i > 2B_\nu(\tau) + \varepsilon_c$. Thus, a smaller separation threshold certifies a broader set of outliers and corresponds to a stronger IM effect. A smaller ε_c lowers this threshold directly and can also tighten $B_\nu(\tau)$ through more favorable data geometry. More balanced cluster sizes can promote faster contraction and tighter drift bounds, a smaller outlier rate can permit a lower reference-error level, and a smaller initial cluster-center reconstruction error can yield earlier entry and tighter trajectory-drift bounds. These effects can indirectly strengthen the IM effect by reducing $B_\nu(\tau)$.

4 EXPERIMENTS

4.1 SIMULATION STUDIES

We examine how data geometry and initialization affect the IM effect. Specifically, we investigate four key factors discussed in Section 3.3: (i) the within-cluster variation ε_c, (ii) cluster-size balance $(n_{\min} : n_{\max})$, and (iii) the outlier rate ρ, and (iv) the initialization quality, quantified by the initial cluster-center reconstruction error $\widetilde{R}_0 := \|\mathcal{F}_{W_0}(\widetilde{X}) - X\|_2$.

Data Construction. We generate $n = 2{,}000$ samples in $\mathbb{R}^2$ with $K = 3$ cluster centers. The centers μ_1, μ_2, and μ_3 are located in $[-10, 10]^2$, with a minimum pairwise separation of at least 16. Unless otherwise specified, we use balanced clusters, so that $n_{\min} = n_{\max}$ up to integer rounding. Each inlier receives a Gaussian perturbation $\xi_i \sim \mathcal{N}(0, \widetilde{\varepsilon}_c^2 I)$, where I denotes the identity matrix, and we control the within-cluster variation ε_c defined in Section 3.1 through $\widetilde{\varepsilon}_c$. For notational simplicity, we write ε_c instead of $\widetilde{\varepsilon}_c$ whenever the context is clear in this section. For the default contamination setting, we assign the same cluster-wise outlier fraction, $|\mathcal{O} \cap \Lambda_k|/n_k = 0.05$ for every k, up to integer rounding, so that $\rho = 0.05$. The outliers are sampled uniformly from $[-18, 18]^2$, subject to a minimum distance of 10 from every cluster center. All features are min–max scaled to $[-1, 1]$. The specific values of the factor-related quantities are set separately for each experiment.

Model Training and Evaluation Metrics. We use an autoencoder with a single layer of hidden size $H = 32$, whose activation function is $\tanh$, with Kaiming uniform initialization (He et al., 2015). We fix the decoder and train the encoder only using full-batch SGD with learning rate 0.1. At each iteration τ, we compute the per-sample reconstruction error $\|x_i - F_{W_\tau}(x_i)\|_2^2$ as the outlier score and evaluate the corresponding AUROC, denoted by AUROC(τ). We define the *IM window* as the interval between the first and the last epoch at which AUROC$(\tau) \geq 0.9$, which we use as an empirical proxy for the theoretical interval $\mathcal{T}_\nu$. More details are given in Appendix B.2.

Results. We present the results showing how ε_c, $(n_{\min}, n_{\max})$, ρ, and $\widetilde{R}_0$ each affect the IM effect. In Appendix B.3, we repeat them under a practical training scheme (i.e., learning both the encoder and decoder with mini-batches and the Adam optimizer) and observe similar results.

(i) **Effect of within-cluster variation ε_c**: We vary $\varepsilon_c \in \{1.0, 2.5, 4.0\}$ while keeping all other factors fixed. Figure 2 shows the data distributions and AUROC trajectories of the trained autoencoder. Consistent with the analysis in Section 3.3, tighter clusters (smaller ε_c) yield a stronger and more sustained IM effect: the peak AUROC increases from 0.949 ($\varepsilon_c = 4.0$) to 0.975 ($\varepsilon_c = 1.0$), and the IM window widens from 3,660 to 9,030, with the onset T_1 becoming earlier ($100 \to 80$).

(ii) **Effect of cluster-size balance**: With fixed $\varepsilon_c = 1.5$, we vary the cluster-size ratio $n_{\max} : n_{\min}$ across three levels: balanced (1:1), moderate (11:1), and heavy imbalance (27:1). The smallest cluster is assigned to a fixed spatial position across settings for fair comparison. Figure 3 shows that heavier imbalance degrades the persistence of the IM effect: the IM window narrows from 11,670 epochs (balanced) to 9,240 epochs (heavily imbalanced), while the peak AUROC rises from 0.959 to 0.967. This is consistent with Section 3.3, where more balanced clusters favor a more persistent IM effect.

(iii) **Effect of outlier rate ρ**: We fix $\varepsilon_c = 2.0$ and vary $\rho \in \{0.05, 0.10, 0.15\}$. Figure 4 shows that a higher ρ substantially shortens the IM window from 6,930 iterations ($\rho = 0.05$) to 2,795 iterations ($\rho = 0.15$). This persistence trend is consistent with the analysis in Section 3.3, which indicates that a smaller ρ can permit a wider IM interval. We also observe stronger empirical

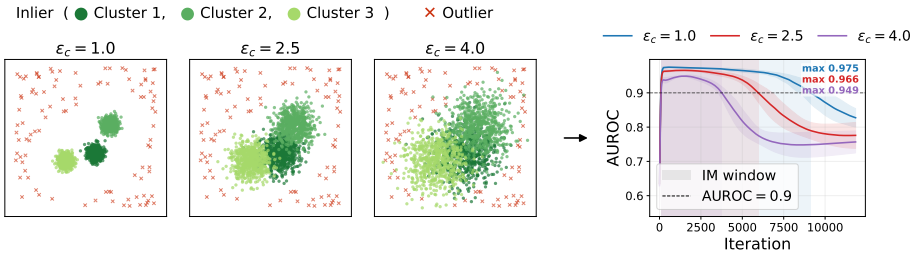


Figure 2: *Effect of within-cluster variation* ε_c. (Left) Data distributions for $\varepsilon_c \in \{1.0, 2.5, 4.0\}$. (Right) AUROC trajectories; shaded regions indicate the IM window, curves are means over five decoder initializations, and bands show the standard deviation.

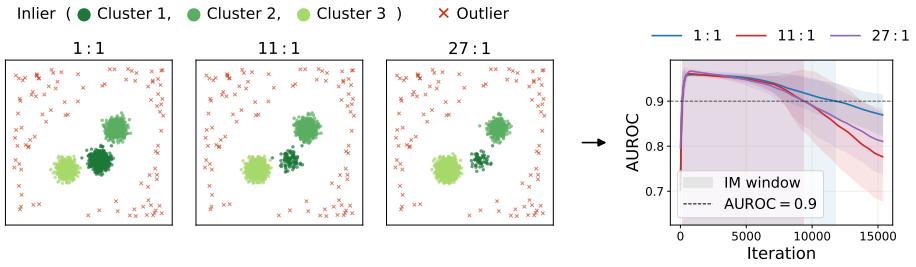


Figure 3: *Effect of cluster-size balance* $n_{\max} : n_{\min}$. (Left) Data distributions for $(n_{\max} : n_{\min}) \in \{(1:1), (11:1), (27:1)\}$. (Right) AUROC trajectories; shaded regions indicate the IM window, curves are means over ten decoder initializations, and bands show the standard deviation.

separation at lower outlier rates, with the peak AUROC increasing from 0.952 at $\rho = 0.15$ to 0.969 at $\rho = 0.05$.

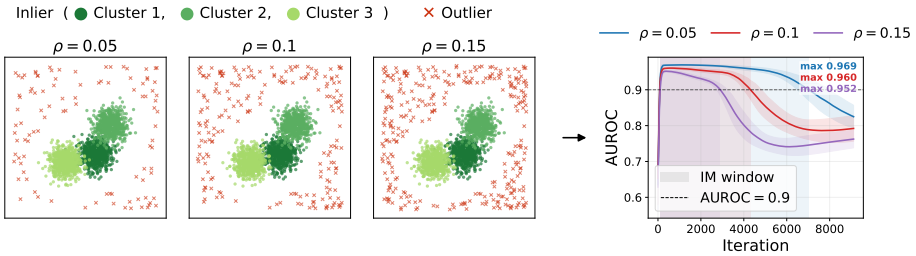


Figure 4: *Effect of outlier rate* ρ. (Left) Data distributions for $\rho \in \{0.05, 0.10, 0.15\}$. (Right) AUROC trajectories; shaded regions indicate the IM window, curves are means over five decoder initializations, and bands show the standard deviation.

(iv) **Effect of initialization quality** $\widetilde{R}_0$: We fix $\varepsilon_c = 4.0$ and examine how the initial cluster-center reconstruction error affects the IM window. Table 1 reports the IM window $[T_1, T_2]$ and the best AUROC for three initializations with different values of $\widetilde{R}_0$. Consistent with the analysis in Section 3.3, a smaller $\widetilde{R}_0$ is associated with an earlier onset and a wider IM window, while the best AUROC remains similar across the three initializations.

Table 1: *Effect of initialization quality* $\widetilde{R}_0$ on the IM window and detection performance (best AUROC).

$\widetilde{R}_0$	T_1	T_2	IM window length $(T_2 - T_1)$	Best AUROC
30.1	82	6,057	5,975	0.952
33.9	148	3,789	3,641	0.948
36.5	171	1,414	1,243	0.950

4.2 THEORY-DRIVEN GUIDELINES FOR REAL-WORLD DATA ANALYSIS

Our theoretical analysis in Section 3 explains the emergence of the IM effect and characterizes how its strength and persistence depend on data distribution and initialization. Building on these insights, we study two simple, theory-motivated strategies for improving IM-based solvers: (i) data preprocessing to shape the data distribution, and (ii) initialization schemes that favor inlier-oriented training dynamics.

Using Compact Representations. While several properties of the data distribution, such as cluster sizes $(n_{\min}, n_{\max})$, are intrinsic, preprocessing may reduce the inlier radius ε_c and the cluster-wise mean deviation ν_0. As discussed in Section 3.3, these geometric quantities influence the strength and persistence of the IM effect.

A natural approach is to transform the data into a representation space where inlier clusters become more compact, removing nuisance variation that does not carry the cluster structure. One can leverage pre-trained models used in the same domain, such as TabPFN (Hollmann et al., 2023; Ye et al., 2025) for tabular data, ViT (Dosovitskiy et al., 2021) for images, and BERT (Devlin et al., 2019) for text. When such models are unavailable, an alternative is to use self-supervised representations (Gui et al., 2024), such as contrastive learning (van den Oord et al., 2018). Prior work shows that contrastive learning preserves cluster structure while reducing within-cluster variability (Huang et al., 2023).

Initialization for Capturing Cluster Structure. As discussed in Section 3.3, a smaller initial cluster-center reconstruction error $\widetilde{R}_0$ can favor an earlier onset and a more persistent IM effect. This suggests that favorable initializations should capture coarse inlier structure without prematurely fitting sample-specific noise.

One practical approach is to refine the initialization by applying an exponential moving average (EMA) (Tarvainen & Valpola, 2017) to the model parameters during the early training phase. While EMA is commonly used to stabilize optimization, here it is intended to help suppress the influence of outliers in the early stage. By averaging parameters over successive updates, EMA may bias the model toward consistently learned structures, which are dominated by inliers, and mitigate the effect of irregular gradient signals induced by outliers.

Empirical Validation on Real-World Data. We evaluate these two strategies using two representative IM-based methods, ODIM (Kim et al., 2024) and ALTBI (Cho et al., 2025). Given the scalability limitations of TabPFN (Hollmann et al., 2023), we focus on eleven tabular datasets from ADBench (Han et al., 2022) with $1{,}000 < n < 50{,}000$ and $p \leq 62$. To assess the individual and combined effects of the two strategies, we compare four configurations for each method: (i) Original uses raw inputs without EMA; (ii) +EMA applies EMA initialization to the model trained on raw inputs; (iii) +TabPFN uses TabPFN embeddings without EMA; and (iv) +EMA+TabPFN combines TabPFN embeddings with EMA initialization. Performance is measured by AUROC, and further implementation details are provided in Appendices B.1 and B.2.

Figure 5 summarizes the individual and combined effects of the two strategies on ODIM and ALTBI. The effects are dataset-dependent: compared with Original, +EMA+TabPFN improves AUROC on seven of the eleven datasets for ODIM and decreases it on the remaining four. The mean AUROC increases from 0.640 to 0.649 for ODIM and from 0.642 to 0.671 for ALTBI.

To explore whether the performance changes are associated with data geometry, we conduct a post-hoc diagnostic motivated by Corollary 3.3. We define the performance change as $\Delta := \text{AUROC}(\text{+EMA + TabPFN}) - \text{AUROC}(\text{Original})$. Using outlier labels, we fit a Bayesian Gaussian mixture model with diagonal covariance matrices to the inliers, allowing at most 20 components and retaining those whose mixture weights exceed 0.02. For each fit, let $\hat{K}$ and $\hat{\mu}_1, \ldots, \hat{\mu}_{\hat{K}}$ denote the number and centers of the retained components. We define $\hat{\delta} := \min_{a \neq b} \|\hat{\mu}_a - \hat{\mu}_b\|_2$ and use the 95th percentile of inlier distances to the nearest retained center as $\hat{\varepsilon}_c$. We then compute the geometric diagnostic $\sigma := \text{median}_{i \in \mathcal{O}}\{\min_{k \in [\hat{K}]} \|x_i - \hat{\mu}_k\|_2\}/(\hat{\delta} + \hat{\varepsilon}_c)$. We report σ averaged over three independent mixture fits. Outlier labels are used only for evaluation and this post-hoc analysis, never by the detectors. The denominator is a quantile-based proxy for the threshold $\delta + \varepsilon_c$, so a smaller σ indicates a smaller typical outlier-to-center distance relative to this threshold. This diagnostic is intended to explore geometric associations rather than verify the sufficient conditions of Corollary 3.3.

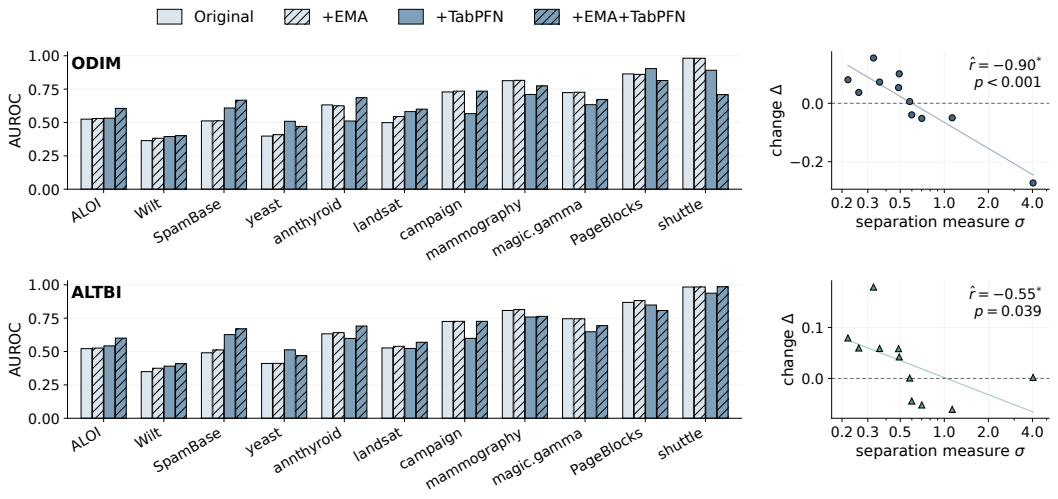


Figure 5: *Effects of TabPFN preprocessing and EMA initialization*. (Top) ODIM and (Bottom) ALTBI. (Left) AUROC on the eleven datasets. (Right) The separation measure σ versus the change Δ in AUROC. Lines are least-squares fits, $\hat{r}$ is the sample Pearson correlation between $\log \sigma$ and Δ, and p is the one-sided p-value for $H_0 : r = 0$ against $H_1 : r < 0$, where r is the population correlation. Table B.3 presents the results on each dataset.

As shown in Figure 5 along with the full results in Table B.3 in Appendix B.3, we observe that smaller σ tends to accompany larger gains in these eleven datasets. The sample Pearson correlation $\hat{r}$ between $\log \sigma$ and Δ is more pronounced for ODIM ($\hat{r} = -0.90$, $p < 0.001$) than for ALTBI ($\hat{r} = -0.55$, $p = 0.039$), where p is the one-sided p-value for $H_0 : r = 0$ against $H_1 : r < 0$.

Overall Benchmark Performance. Additional analyses in Appendix B.3 extend the evaluation beyond the eleven tabular datasets. On the 47 tabular datasets, EMA initialization is applied to all datasets, while TabPFN representations are used only for the eleven datasets considered above. On the three image benchmarks, we similarly examine EMA initialization and ViT representations. Tables B.4 and B.5 report these component-wise results, and Figures B.4 and B.5 summarize the corresponding AUROC and AUPRC performance.

We further evaluate the methods on the full set of 57 ADBench datasets. Here, the original ODIM and ALTBI already use ViT and BERT representations for image and text data, respectively; thus, the full-benchmark modification consists of EMA initialization across all datasets and TabPFN replacement for the selected eleven tabular datasets. Across all 57 datasets, these modifications increase the average AUROC of ALTBI from 0.757 to 0.766 and of ODIM from 0.751 to 0.757 (Tables B.6 and B.7). The resulting ALTBI variant achieves the highest average AUROC and AUPRC among the methods compared in Tables B.8 and B.9.

5 Concluding Remarks

In this paper, we analyzed the IM effect in a fixed-decoder autoencoder trained by gradient descent, establishing conditions for a nonempty training interval in which sufficiently distant outliers have larger reconstruction errors than all inliers. Our analysis further clarifies how data geometry and initialization influence the strength and persistence of the IM effect. Simulations support these qualitative insights under both the theoretical setting and more practical joint-training settings.

Guided by these insights, we studied compact representations and EMA-based initialization for existing IM-based detectors. Although their effects are dataset-dependent, the empirical results are broadly consistent with the theoretical insights, and the resulting ALTBI variant incorporating both strategies achieves the highest average AUROC and AUPRC among the methods evaluated across 57 ADBench datasets. Future work includes extending the theory to deeper architectures, joint training, stochastic optimization, and broader generative models.

AI USE STATEMENT

Generative AI tools were mainly used to aid or polish writing. The research questions, the assumptions, and the theorem statements were developed by the authors without them. In particular, all numerical results reported in this paper, including the simulations and the real-data experiments, were obtained from experiments run by ourselves. We take responsibility for the final content of this study.

ETHICS STATEMENT

This work uses publicly released benchmark data `ADBench` and introduces none of its own. Outlier labels are never seen by the detectors and are used only for evaluation and for the diagnostics of Section 4.2. No human subjects were involved, and no personally identifiable information was collected or used.

REPRODUCIBILITY STATEMENT

Our theoretical claims are stated with its assumptions and proved in Appendix A. Appendix B.1 describes the datasets and Appendix B.2 the architectures, optimizers, and protocols for both the simulations and the real-data experiments. All real-world datasets are taken from the public repository.

REFERENCES

Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu. A survey of network anomaly detection techniques. *Journal of network and computer applications*, 60:19–31, 2016.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pp. 322–332. PMLR, 2019. URL http://proceedings.mlr.press/v97/arora19a.html.

Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *International Conference on Learning Representations*, 2020.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, may 2000. ISSN 0163-5808. doi: 10.1145/335191.335388.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020.

Seoyoung Cho, Jaesung Hwang, Kwan-Young Bak, and Dongha Kim. ALTBI: constructing improved outlier detection models via optimization of inlier-memorization effect. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 11544–11552. AAAI Press, 2025. doi: 10.1609/AAAI.V39I11.33256. URL https://doi.org/10.1609/aaai.v39i11.33256.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*

2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Xueying Ding, Lingxiao Zhao, and Leman Akoglu. Hyperparameter sensitivity in deep outlier detection: Analysis and a scalable hyper-ensemble solution. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/3e9113e2bc2e700baa7d765470f140e1-Abstract-Conference.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Cecile Fauconnier and Gentiane Haesbroeck. Outliers detection with the minimum covariance determinant estimator in practice. *Statistical Methodology*, 6(4):363–379, 2009.

Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM computing surveys (CSUR)*, 54(7): 1–37, 2021.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *arXiv preprint arXiv:1805.10917*, 2018.

Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 1:59–63, 2012.

Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. DROCC: deep robust one-class classification. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3711–3721. PMLR, 2020.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):9052–9071, 2024. doi: 10.1109/TPAMI.2024.3415112. URL https://doi.org/10.1109/TPAMI.2024.3415112.

Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. In *Neural Information Processing Systems (NeurIPS)*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern recognition letters*, 24(9-10):1641–1650, 2003.

Waleed Hilal, S Andrew Gadsden, and John Yawney. Financial fraud: a review of anomaly detection techniques and recent advances. *Expert systems With applications*, 193:116429, 2022.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647. URL https://www.science.org/doi/abs/10.1126/science.1127647.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Weiran Huang, Mingyang Yi, Xuyang Zhao, and Zihao Jiang. Towards the generalization of contrastive self-supervised learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=XDJwuEYHhme.

Yihong Huang, Yuang Zhang, Liping Wang, Fan Zhang, and Xuemin Lin. Entropystop: Unsupervised deep outlier detection with loss entropy. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 1143–1154. ACM, 2024. doi: 10.1145/3637528.3671943. URL https://doi.org/10.1145/3637528.3671943.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8580–8589, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Minseo Kang, Seunghwan Park, and Dongha Kim. Memorize early, then query: Inlier-memorization-guided active outlier detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (17):15000–15008, Mar. 2026. doi: 10.1609/aaai.v40i17.38522. URL https://ojs.aaai.org/index.php/AAAI/article/view/38522.

Dongha Kim, Jaesung Hwang, Jongjin Lee, Kunwoong Kim, and Yongdai Kim. ODIM: outlier detection via likelihood of under-fitted generative models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=R8nbccD7kv.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43 (11):3964–3979, 2020.

Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 157–166, 2005.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4313–4324. PMLR, 26–28 Aug 2020a. URL https://proceedings.mlr.press/v108/li20j.html.

Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. Copod: copula-based outlier detection. In *2020 IEEE international conference on data mining (ICDM)*, pp. 1118–1123. IEEE, 2020b.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George Chen. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pp. 413–422. IEEE, 2008.

Yezheng Liu, Zhe Li, Chong Zhou, Yuanchun Jiang, Jianshan Sun, Meng Wang, and Xiangnan He. Generative adversarial active learning for unsupervised outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1517–1528, 2019.

Victor Livernoche, Vineet Jain, Yashar Hezaveh, and Siamak Ravanbakhsh. On diffusion modeling for anomaly detection. *CoRR*, abs/2305.18593, 2023. doi: 10.48550/ARXIV.2305.18593.

Thanh Van Nguyen, Raymond K. W. Wong, and Chinmay Hegde. Benefits of jointly training autoencoders: An improved neural tangent kernel analysis. *IEEE Trans. Inf. Theory*, 67(7): 4669–4692, 2021. doi: 10.1109/TIT.2021.3065212. URL https://doi.org/10.1109/TIT.2021.3065212.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *CoRR*, abs/1902.04674, 2019. URL http://arxiv.org/abs/1902.04674.

Tomáš Pevnỳ. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102:275–304, 2016.

Chen Qiu, Timo Pfrommer, Marius Kloft, Stephan Mandt, and Maja Rudolph. Neural transformation learning for deep anomaly detection beyond images. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 8703–8714. PMLR, 2021. URL http://proceedings.mlr.press/v139/qiu21a.html.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 427–438, 2000.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/rezende14.html.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402. PMLR, 10–15 Jul 2018.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2020.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In Marc Niethammer, Martin Styner, Stephen R. Aylward, Hongtu Zhu, Ipek Oguz, Pew-Thian Yap, and Dinggang Shen (eds.), *Information Processing in Medical Imaging - 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*, Lecture Notes in Computer Science, pp. 146–157. Springer, 2017. doi: 10.1007/978-3-319-59050-9_12. URL https://doi.org/10.1007/978-3-319-59050-9_12.

Bernhard Schölkopf, John Platt, John Shawe-Taylor, Alexander Smola, and Robert Williamson. Estimating support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 07 2001. doi: 10.1162/089976601750264965.

Tom Shenkar and Lior Wolf. Anomaly detection for tabular data with internal contrastive learning. In *International conference on learning representations*, 2022.

Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. In *Proceedings of the IEEE foundations and new directions of data mining workshop*, pp. 172–179. IEEE Press, 2003.

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

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 1195–1204, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54:45–66, 2004.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Li Wang and Wei Huang. On the convergence analysis of over-parameterized variational autoencoders: a neural tangent kernel perspective. *Mach. Learn.*, 114(1):15, 2025. doi: 10.1007/S10994-024-06683-Z. URL https://doi.org/10.1007/s10994-024-06683-z.

Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A closer look at tabpfn v2: Strength, limitation, and extension. *CoRR*, abs/2502.17361, 2025. doi: 10.48550/ARXIV.2502.17361. URL https://doi.org/10.48550/arXiv.2502.17361.

Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *International conference on machine learning*, pp. 1100–1109. PMLR, 2016.

Yuang Zhang, Liping Wang, Yihong Huang, Yuanxing Zheng, Fan Zhang, and Xuemin Lin. Gradstop: Exploring training dynamics in unsupervised outlier detection through gradient. *arXiv preprint arXiv:2412.08501*, 2024.

Chong Zhou and Randy C. Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS*, pp. 665–674. ACM, 2017. doi: 10.1145/3097983.3098052. URL https://doi.org/10.1145/3097983.3098052.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.

A THEORETICAL STUDIES

We present here the assumptions, formal statements, and proofs of Theorem 3.1 and Corollary 3.3, which establish the early inlier fitting behavior of the autoencoder. The proofs proceed in three steps.

First, in Appendix A.2, we introduce the cluster-consistent subspace and study gradient descent on the cluster-centered inputs $\widetilde{X}$. Second, in Appendix A.3, we compare the observed-input and cluster-centered trajectories using aggregate nearest-center deviations. Finally, we turn the resulting center-reconstruction bound into a pointwise reconstruction separation theorem and prove that its integer-valued training window is nonempty. We note that our proof follows a similar structure to that of Li et al. (2020a).

A.1 PRELIMINARIES AND TECHNICAL SETUP

We retain the stacked-vector notation from Section 3.1. mat is the inverse of column-wise vectorization. Replacing each x_i by its assigned center $\widetilde{x}_i$ defines $\widetilde{X}$ and $\mathcal{F}_W(\widetilde{X})$. The deviation vector is $E := X - \widetilde{X}$, with aggregate norm $R_E := \|E\|_2 = (\sum_i d_i^2)^{1/2}$. For a block vector $v = (v_1^\top, \dots, v_n^\top)^\top$, write $\|v\|_{2,\infty} := \max_i \|v_i\|_2$ and write $\|v\|_2$ for its Euclidean norm. For matrices, let $\|\cdot\|$ denote the operator norm and $\|\cdot\|_F$ denote the Frobenius norm.

We use the following normalization assumption. The upper bound is set to one for convenience; any finite bound can be carried through the proof.

Assumption A.1 (Bounded Sample Norms)**.** $\|x_i\|_2 \le 1, \forall i \in [n]$.

A.2 CLUSTER-CENTERED FRAMEWORK

A.2.1 NEAREST-CENTER GEOMETRY AND SUBSPACE PROJECTIONS

Definition A.1 (Support Subspace)**.** *Define*

$$S_+ := \left\{ v = (v_1^\top, \dots, v_n^\top)^\top \in \mathbb{R}^{np} : \exists u_1, \dots, u_K \in \mathbb{R}^p \text{ such that } v_i = u_k \text{ whenever } i \in \Lambda_k \right\}.$$

We also define $S_- := S_+^\perp$, and let Π_{S_+} and Π_{S_-} denote the orthogonal projections onto S_+ and S_-, respectively.

Remark A.1. *S_+ consists of vectors whose components are identical within each cluster. That is, all entries corresponding to samples in the same cluster share the same vector value.*

Lemma A.1 (Explicit Expression for Π_{S_+})**.** *For any $v = (v_1^\top, \dots, v_n^\top)^\top \in \mathbb{R}^{np}$ and any cluster $k \in [K]$, the orthogonal projection of v onto S_+ is given by*

$$(\Pi_{S_+} v)_i = \frac{1}{n_k} \sum_{j \in \Lambda_k} v_j, \quad \forall i \in \Lambda_k.$$

Proof. Define $Pv \in \mathbb{R}^{np}$ by $(Pv)_i := \frac{1}{n_k} \sum_{j \in \Lambda_k} v_j, \forall i \in \Lambda_k$. Clearly $Pv \in S_+$. It remains to show that $v - Pv \in S_+^\perp$. Fix $u \in S_+$. Then there exist $u_1, \dots, u_K \in \mathbb{R}^p$ such that $u_i = u_k$ whenever $i \in \Lambda_k$. Hence

$$\begin{aligned}\langle v - Pv, u \rangle &= \sum_{k=1}^K \sum_{i \in \Lambda_k} \left\langle v_i - \frac{1}{n_k} \sum_{j \in \Lambda_k} v_j, u_k \right\rangle = \sum_{k=1}^K \left\langle \sum_{i \in \Lambda_k} v_i - \sum_{i \in \Lambda_k} \frac{1}{n_k} \sum_{j \in \Lambda_k} v_j, u_k \right\rangle \\ &= \sum_{k=1}^K \left\langle \sum_{i \in \Lambda_k} v_i - \sum_{j \in \Lambda_k} v_j, u_k \right\rangle = 0.\end{aligned}$$

Therefore $v - Pv \in S_+^\perp$, so $Pv = \Pi_{S_+} v$. □

Lemma A.2 (Cluster-Wise Mean Deviation)**.** *For each $k \in [K]$, define*

$$\bar{e}_k := \frac{1}{n_k} \sum_{j \in \Lambda_k} e_j, \quad \nu_0 := \max_{k \in [K]} \|\bar{e}_k\|_2.$$

Then $(\Pi_{S_+} E)_i = \bar{e}_k$ for every $i \in \Lambda_k$, and hence $\|\Pi_{S_+} E\|_{2,\infty} = \nu_0$.

Proof. The identity follows immediately by applying Lemma A.1 to E. Taking the maximum block norm gives the second claim. □

Lemma A.3 (Interpretable Bound on the Mean Deviation)**.** *Let*

$$a_{\mathcal{O}} := \max_{k\in[K]} \frac{1}{n_k} \sum_{i\in\mathcal{O}\cap\Lambda_k} d_i.$$

Then $\nu_0 \le \varepsilon_c + a_{\mathcal{O}}$. Moreover, with $d_{\mathcal{O},\max} := \max_{i\in\mathcal{O}} d_i$ (with value zero when $\mathcal{O} = \varnothing$), we have $a_{\mathcal{O}} \le \rho d_{\mathcal{O},\max}$.

Proof. For every k, the triangle inequality and $\|e_i\|_2 = d_i$ give

$$\begin{aligned}\|\bar{e}_k\|_2 &\le \frac{1}{n_k} \sum_{i\in\mathcal{I}\cap\Lambda_k} d_i + \frac{1}{n_k} \sum_{i\in\mathcal{O}\cap\Lambda_k} d_i \\ &\le \varepsilon_c + a_{\mathcal{O}}.\end{aligned}$$

The final claim follows from $n_k^{-1} \sum_{i\in\mathcal{O}\cap\Lambda_k} d_i \le (|\mathcal{O}\cap\Lambda_k|/n_k) d_{\mathcal{O},\max} \le \rho d_{\mathcal{O},\max}$. □

A.2.2 Jacobian Structure on the Cluster-Centered Inputs

The reference loss $\widetilde{\mathcal{L}}(W) := \frac{1}{2}\|\mathcal{F}_W(\widetilde{X}) - X\|_2^2$ replaces inputs by their centers while retaining the observed reconstruction targets. The observed-input loss is defined in eq. (2).

Let $\widetilde{J}(W) := \partial\mathcal{F}_W(\widetilde{X})/\partial w$; $J(W)$ is the observed-input Jacobian from eq. (2). Both have dimensions $np \times Hp$. In addition, for two parameter matrices W_1, W_2, define their corresponding averaged Jacobians by

$$\widetilde{J}(W_1, W_2) := \int_0^1 \widetilde{J}(W_2 + t(W_1 - W_2))dt$$

and

$$J(W_1, W_2) := \int_0^1 J(W_2 + t(W_1 - W_2))dt.$$

The following lemma shows that the averaged Jacobians yield an exact mean-value representation of the change in the autoencoder outputs between two parameter values.

Lemma A.4 (Mean-Value Jacobian Representation)**.** *For any $W_1, W_2 \in \mathbb{R}^{H\times p}$, we have*

$$\mathcal{F}_{W_1}(\widetilde{X}) - \mathcal{F}_{W_2}(\widetilde{X}) = \widetilde{J}(W_1, W_2)(\mathrm{vec}(W_1) - \mathrm{vec}(W_2))$$

and

$$\mathcal{F}_{W_1}(X) - \mathcal{F}_{W_2}(X) = J(W_1, W_2)(\mathrm{vec}(W_1) - \mathrm{vec}(W_2))$$

Proof. Integrate the derivative of the output map along the segment from W_2 to W_1. The fundamental theorem of calculus directly gives both identities. □

We define the cluster-center map by

$$\mathcal{F}_W(\mu) := \begin{bmatrix} F_W(\mu_1) \\ \vdots \\ F_W(\mu_K) \end{bmatrix} \in \mathbb{R}^{Kp},$$

which stacks the autoencoder outputs evaluated at the cluster centers. We further define its Jacobian matrix with respect to the parameter vector w by

$$J_c(W) := \frac{\partial\mathcal{F}_W(\mu)}{\partial w} \in \mathbb{R}^{Kp\times Hp}.$$

The following assumption imposes standard local regularity conditions on the Jacobian at the cluster centers, ensuring stable and well-conditioned optimization dynamics near the initialization.

Assumption A.2 (Cluster-Center Jacobian). *Fix an initialization $W_0 \in \mathbb{R}^{H\times p}$ and a radius $R_{\mathrm{loc}} > 0$. Define*

$$\mathcal{D}(W_0, R_{\mathrm{loc}}) := \{W \in \mathbb{R}^{H\times p} : \|W - W_0\|_F \le R_{\mathrm{loc}}\}.$$

Assume that there exist constants $\alpha_c, \beta_c, L_c > 0$ such that, for every $W \in \mathcal{D}(W_0, R_{\mathrm{loc}})$,

(i) for every unit vector $z \in \mathbb{R}^{Kp}$, $\alpha_c \le \|J_c(W)^\top z\|_2 \le \beta_c$;

(ii) for all $W_1, W_2 \in \mathcal{D}(W_0, R_{\mathrm{loc}})$, $\|J_c(W_1) - J_c(W_2)\| \le L_c\|W_1 - W_2\|_F$.

Remark A.2 (Milder Assumptions). *An advantage of our theoretical framework is that, in contrast to existing works that analyze training dynamics of neural networks, it does not rely on strong overparameterization or restrictive architectural assumptions. These analyses typically consider extremely wide networks and carefully structured output layers to guarantee perfect fitting (Jacot et al., 2018; Arora et al., 2019; Li et al., 2020a;a; Oymak & Soltanolkotabi, 2019; Nguyen et al., 2021). On the other hand, our analysis does not require perfect fitting, but only that inliers achieve smaller reconstruction error than outliers at some stage of training. This enables milder assumptions, requiring only a bounded spectral norm of the decoder matrix A and, on the width, one hidden node per cluster rather than a width polynomial in n (Remark A.3), thereby narrowing the gap between theory and practice.*

Remark A.3 (Scope of the sufficient conditions). *The result does not require an NTK-style width polynomial in n, but the cluster-center Jacobian condition implies the explicit rank requirement $H \ge K$. That is, Assumption A.2(i) requires $J_c(W)$ to have full row rank. Since $J_c(W) \in \mathbb{R}^{Kp\times Hp}$, it necessarily implies $H \ge K$. We can state this implication explicitly: the analysis does not require a width polynomial in n, but it does require at least one hidden block per reference center. The formal guarantee is for a fixed decoder, encoder-only full-batch gradient descent, and a local regularity region, whereas the real-data experiments use jointly trained models and stochastic optimization, beyond the setting covered by the theorem. However, we numerically show that our theory-driven strategies can be applied well to the practical setting.*

Under Assumption A.2, the cluster-representative Jacobian inherits the same structured regularity properties, summarized in the following proposition.

Proposition A.5 (Jacobian Structure of Cluster-Representative Inputs). *Assume Assumption A.2. Fix*

$$\alpha := \sqrt{n_{\min}}\alpha_c, \beta := \sqrt{n_{\max}}\beta_c, L := \sqrt{n_{\max}}L_c.$$

Then, for every $W \in \mathcal{D}(W_0, R_{\mathrm{loc}})$,

(i) $\mathrm{range}(\widetilde{J}(W)) \subset S_+$;

(ii) for every unit vector $u \in S_+$, $\alpha \le \|\widetilde{J}(W)^\top u\|_2 \le \beta$;

(iii) for all $W_1, W_2 \in \mathcal{D}(W_0, R_{\mathrm{loc}})$, $\|\widetilde{J}(W_1) - \widetilde{J}(W_2)\| \le L\|W_1 - W_2\|_F$.

Proof. Define the duplication operator $D : \mathbb{R}^{Kp} \to \mathbb{R}^{np}$ by $(Dz)_i := z_k$ whenever $i \in \Lambda_k$. Then $D(\mathbb{R}^{Kp}) = S_+$, $\mathcal{F}_W(\widetilde{X}) = D\mathcal{F}_W(\mu)$, and $\widetilde{J}(W) = DJ_c(W)$. This proves $\mathrm{range}(\widetilde{J}(W)) \subset S_+$.

Now fix a unit vector $u \in S_+$. Since $u \in S_+$, there exist vectors $u_1, \dots, u_K \in \mathbb{R}^p$ such that $u_i = u_k$ for all $i \in \Lambda_k$. Then $\|u\|_2^2 = \sum_{k=1}^K n_k\|u_k\|_2^2 = 1$ and

$$D^\top u = \begin{bmatrix} n_1 u_1 \\ \vdots \\ n_K u_K \end{bmatrix}.$$

Hence, we get $\|D^\top u\|_2^2 = \sum_{k=1}^K n_k^2\|u_k\|_2^2$. Since $\sum_{k=1}^K n_k\|u_k\|_2^2 = 1$, we also get $n_{\min} \le \|D^\top u\|_2^2 \le n_{\max}$ and $\widetilde{J}(W)^\top u = J_c(W)^\top D^\top u$. Let $z := D^\top u/\|D^\top u\|_2$. Then $\|z\|_2 = 1$, so by Assumption A.2,

$$\alpha_c\|D^\top u\|_2 \le \|\widetilde{J}(W)^\top u\|_2 \le \beta_c\|D^\top u\|_2.$$

Using the bound on $\|D^\top u\|_2$ yields $\sqrt{n_{\min}}\alpha_c \le \|\widetilde{J}(W)^\top u\|_2 \le \sqrt{n_{\max}}\beta_c$. Finally, we have

$$\widetilde{J}(W_1) - \widetilde{J}(W_2) = D(J_c(W_1) - J_c(W_2)).$$

Since $\|D\| = \sqrt{n_{\max}}$, we obtain

$$\|\widetilde{J}(W_1) - \widetilde{J}(W_2)\| \le \sqrt{n_{\max}}\|J_c(W_1) - J_c(W_2)\| \le \sqrt{n_{\max}} L_c \|W_1 - W_2\|_F.$$

This completes the proof. □

A.2.3 Gradient Descent Dynamics on Cluster-Centered Inputs

Let $(\widetilde{W}_\tau)_{\tau\ge 0}$ denote the matrix-form gradient descent sequence for the cluster-centered loss $\widetilde{\mathcal{L}}(W)$, initialized at $\widetilde{W}_0 = W_0$, and generated by

$$\widetilde{w}_{\tau+1} = \widetilde{w}_\tau - \eta \widetilde{J}(\widetilde{W}_\tau)^\top \widetilde{r}_\tau, \quad \text{where} \quad \widetilde{r}_\tau := \mathcal{F}_{\widetilde{W}_\tau}(\widetilde{X}) - X,$$

and $\widetilde{W}_\tau := \mathrm{mat}(\widetilde{w}_\tau)$. Let $R_0 := \|\widetilde{r}_0\|_2, \quad \widehat{R}_0 := \widetilde{R}_0 + R_E$ and for any $\nu > 0$, we also define the iteration point

$$t_\nu := T_\nu^{\mathrm{exact}} := \left\lceil \frac{2}{\eta\alpha^2} \log_+ \left(\frac{\widehat{R}_0}{\nu} \right) \right\rceil, \quad q := 1 - \eta\alpha^2.$$

Here $\log_+(x) := \max\{\log x, 0\}$ for $x > 0$, with $\log_+(0) := 0$.

Assumption A.3 (Radius condition)**.** $R_{\mathrm{loc}} \ge \frac{2\beta}{\alpha^2}\widehat{R}_0$.

Remark A.4. *Recall that $\widetilde{R}_0 = \|\mathcal{F}_{W_0}(\widetilde{X}) - \widetilde{X}\|_2$. Since $X = \widetilde{X} + E$, the triangle inequality gives the exact data-adaptive upper bound*

$$R_0 \le \widetilde{R}_0 + \|E\|_2 = \widehat{R}_0.$$

The case $\widehat{R}_0 = 0$ is immediate; hence below we assume $\widehat{R}_0 > 0$.

The following lemma summarizes the key gradient descent dynamics under the cluster-centered loss.

Lemma A.6 (Gradient Descent Dynamics on Cluster Representatives)**.** *Assume Assumption A.2 and Assumption A.3. If $0 < \eta \le \min\left\{ \frac{\alpha^2}{8\beta^4}, \frac{\alpha^2}{2L\beta^2\widehat{R}_0} \right\}$, then we have for all $\tau \ge 0$ that $\Pi_{S_-}(\widetilde{r}_\tau) = \Pi_{S_-}(\widetilde{r}_0)$, $\|\Pi_{S_+}(\widetilde{r}_\tau)\|_2^2 \le (1-\eta\alpha^2)^\tau \|\Pi_{S_+}(\widetilde{r}_0)\|_2^2$, and $\|\widetilde{W}_\tau - W_0\|_F \le \frac{2\beta}{\alpha^2}\widehat{R}_0$. Moreover, if $\|\Pi_{S_+}(E)\|_{2,\infty} \le \nu$, then every $\tau \ge T_\nu^{\mathrm{exact}}$ satisfies the sharper post-entry estimate*

$$\|\mathcal{F}_{\widetilde{W}_\tau}(\widetilde{X}) - \widetilde{X}\|_{2,\infty} \le \nu + q^{(\tau - T_\nu^{\mathrm{exact}})/2}\nu \le 2\nu.$$

Proof. Write $\overline{r}_\tau := \Pi_{S_+}(\widetilde{r}_\tau), \overline{e}_\tau := \Pi_{S_-}(\widetilde{r}_\tau), \widetilde{r}_\tau = \overline{r}_\tau + \overline{e}_\tau$. By Lemma A.4 and the gradient descent update, we have $\widetilde{r}_{\tau+1} = \widetilde{r}_\tau - \eta C_\tau \widetilde{r}_\tau$, where $C_\tau := \widetilde{J}(\widetilde{W}_{\tau+1}, \widetilde{W}_\tau)\widetilde{J}(\widetilde{W}_\tau)^\top$.

Since $\widetilde{J}(W) = DJ_c(W)$, $\mathrm{range}(\widetilde{J}(W)) \subset S_+$ for every W; hence $\widetilde{J}(\widetilde{W}_\tau)^\top v = 0$ for all $v \in S_-$. Hence $\widetilde{J}(\widetilde{W}_\tau)^\top \overline{e}_\tau = 0$, and $C_\tau \overline{e}_\tau = 0$. Therefore $\Pi_{S_-}(\widetilde{r}_{\tau+1}) = \Pi_{S_-}(\widetilde{r}_\tau - \eta C_\tau \widetilde{r}_\tau) = \Pi_{S_-}(\widetilde{r}_\tau)$. By induction,

$$\Pi_{S_-}(\widetilde{r}_\tau) = \Pi_{S_-}(\widetilde{r}_0) \quad \text{for all } \tau \ge 0.$$

Moreover, since $C_\tau \overline{e}_\tau = 0$, we have $\overline{r}_{\tau+1} = \overline{r}_\tau - \eta C_\tau \overline{r}_\tau$, which results in

$$\|\overline{r}_{\tau+1}\|_2^2 = \|\overline{r}_\tau\|_2^2 - 2\eta\langle \overline{r}_\tau, C_\tau \overline{r}_\tau \rangle + \eta^2 \|C_\tau \overline{r}_\tau\|_2^2. \tag{A.1}$$

Since $\alpha \le \beta$, the step-size bound gives $q \in (0,1)$. We prove by simultaneous induction that, for every $s \ge 0$,

$$\widetilde{W}_s \in \mathcal{D}(W_0, R_{\mathrm{loc}}), \quad \|\overline{r}_s\|_2 \le q^{s/2}\|\overline{r}_0\|_2. \tag{A.2}$$

The claims hold at $s = 0$. If they hold through time τ, then

$$\begin{aligned} \|\widetilde{W}_{\tau+1} - W_0\|_F &\le \eta\beta \sum_{s=0}^{\tau} \|\overline{r}_s\|_2 \\ &\le \eta\beta\|\overline{r}_0\|_2 \sum_{s=0}^{\tau} q^{s/2} \le \frac{\eta\beta R_0}{1-\sqrt{q}} \le \frac{2\beta}{\alpha^2} R_0 \le \frac{2\beta}{\alpha^2}\widehat{R}_0 \le R_{\mathrm{loc}}. \end{aligned} \tag{A.3}$$

Here $1-\sqrt{1-x} \geq x/2$ for $x \in [0,1]$ was used with $x = \eta\alpha^2$. Thus $\widetilde{W}_{\tau+1}$, and hence the segment from $\widetilde{W}_\tau$ to $\widetilde{W}_{\tau+1}$, lies in the local ball.

Also,

$$\|\widetilde{W}_{\tau+1} - \widetilde{W}_\tau\|_F \leq \eta\beta\|\overline{r}_\tau\|_2 \leq \eta\beta\widehat{R}_0.$$

Integrating the Jacobian Lipschitz bound along this segment yields

$$\|\widetilde{J}(\widetilde{W}_{\tau+1}, \widetilde{W}_\tau) - \widetilde{J}(\widetilde{W}_\tau)\| \leq \int_0^1 Lt\|\widetilde{W}_{\tau+1} - \widetilde{W}_\tau\|_F \, dt \leq \frac{L\eta\beta}{2}\widehat{R}_0 \leq \frac{\alpha^2}{4\beta}.$$

For a unit vector $u \in S_+$, Proposition A.5(ii) and eq. (A.3) now imply

$$\begin{aligned}
\langle u, C_\tau u\rangle &= \|\widetilde{J}(\widetilde{W}_\tau)^\top u\|_2^2 + \left\langle (\widetilde{J}(\widetilde{W}_{\tau+1}, \widetilde{W}_\tau) - \widetilde{J}(\widetilde{W}_\tau))^\top u, \widetilde{J}(\widetilde{W}_\tau)^\top u \right\rangle \\
&\geq \alpha^2 - \frac{\alpha^2}{4\beta}\,\beta = \frac{3\alpha^2}{4}, \\
\|C_\tau u\|_2 &\leq \left(\beta + \frac{\alpha^2}{4\beta}\right)\beta \leq 2\beta^2.
\end{aligned}$$

Applying the two homogeneous bounds to $u = \overline{r}_\tau/\|\overline{r}_\tau\|_2$ (with the zero case immediate) in eq. (A.1) gives

$$\begin{aligned}
\|\overline{r}_{\tau+1}\|_2^2 &\leq \left(1 - \frac{3}{2}\eta\alpha^2 + 4\eta^2\beta^4\right)\|\overline{r}_\tau\|_2^2 \\
&\leq (1-\eta\alpha^2)\|\overline{r}_\tau\|_2^2 = q\|\overline{r}_\tau\|_2^2,
\end{aligned}$$

where the second inequality uses $\eta \leq \alpha^2/(8\beta^4)$. Together with eq. (A.3), this closes the induction. Thus

$$\|\Pi_{S_+}(\widetilde{r}_\tau)\|_2^2 \leq q^\tau\|\Pi_{S_+}(\widetilde{r}_0)\|_2^2, \quad \|\widetilde{W}_\tau - W_0\|_F \leq \frac{2\beta}{\alpha^2}\widehat{R}_0.$$

Finally, since the inputs $\widetilde{x}_i$ are constant inside each cluster, we have $\mathcal{F}_W(\widetilde{X}) \in S_+$ for every W. Also $\widetilde{X} \in S_+$. Hence $\mathcal{F}_{\widetilde{W}_\tau}(\widetilde{X}) - \widetilde{X} \in S_+$. Since $\widetilde{r}_\tau = \mathcal{F}_{\widetilde{W}_\tau}(\widetilde{X}) - X = (\mathcal{F}_{\widetilde{W}_\tau}(\widetilde{X}) - \widetilde{X}) - E$, we get $\mathcal{F}_{\widetilde{W}_\tau}(\widetilde{X}) - \widetilde{X} = \Pi_{S_+}(\widetilde{r}_\tau) + \Pi_{S_+}(E)$. Therefore $\|\mathcal{F}_{\widetilde{W}_\tau}(\widetilde{X}) - \widetilde{X}\|_{2,\infty} \leq \|\Pi_{S_+}(\widetilde{r}_\tau)\|_2 + \|\Pi_{S_+}(E)\|_{2,\infty}$. The definition of t_ν and $q \leq e^{-\eta\alpha^2}$ give $q^{t_\nu/2}\widehat{R}_0 \leq \nu$. Thus, for every $\tau \geq t_\nu$,

$$\|\Pi_{S_+}(\widetilde{r}_\tau)\|_2 \leq q^{\tau/2}\widehat{R}_0 \leq q^{(\tau-t_\nu)/2}\nu. \tag{A.4}$$

Combining this with the projection identity and $\|\Pi_{S_+}E\|_{2,\infty} \leq \nu$ proves the post-entry bound. □

A.3 From the Reference Dynamics to the Observed Data

We compare the reference iterates with GD on the observed inputs.

A.3.1 Stability under nearest-center replacement

Assumption A.4 (Activation regularity)**.** *The activation is twice differentiable and satisfies $|\phi'(t)| \leq B_1$ and $|\phi''(t)| \leq B_2$ for every $t \in \mathbb{R}$.*

Let

$$R_\mu := \max_{k\in[K]}\|\mu_k\|_2, \quad R_W := \|W_0\|_F + R_{\mathrm{loc}},$$

and define the input perturbation bounds $b_i := \|A\|B_1(\|W_0 e_i\|_2 + R_{\mathrm{loc}}d_i)$, $\quad a := \max_i b_i$.

$$\Delta_{F,2} := \Big(\sum_i b_i^2\Big)^{1/2}, \quad \Delta_J := \|A\|\big(B_1 + B_2 R_W R_\mu\big)R_E. \tag{A.5}$$

The term $W_0 e_i$ retains the direction of each deviation. $\Delta_{F,2}$ and Δ_J bound the resulting changes in outputs and Jacobians.

Proposition A.7 (Stability under Nearest-Center Replacement)**.** *Under Assumption A.4, every* $W \in \mathcal{D}(W_0, R_{\mathrm{loc}})$ *satisfies*

$$\|F_W(x_i) - F_W(\widetilde{x}_i)\|_2 \le b_i \quad \textit{for every } i \in [n], \tag{A.6}$$

$$\|\mathcal{F}_W(X) - \mathcal{F}_W(\widetilde{X})\|_2 \le \Delta_{F,2}, \tag{A.7}$$

$$\|J(W) - \widetilde{J}(W)\| \le \Delta_J. \tag{A.8}$$

Proof. Fix $W \in \mathcal{D}(W_0, R_{\mathrm{loc}})$, so $\|W\| \le R_W$. The Lipschitz property of ϕ gives, for every i,

$$\|F_W(x_i) - F_W(\widetilde{x}_i)\|_2 \le \|A\|B_1\|We_i\|_2 \le b_i.$$

This is exactly eq. (A.6); squaring and summing over i and using the definition of $\Delta_{F,2}$ proves eq. (A.7).

For the Jacobian bound, fix $U \in \mathbb{R}^{H\times p}$ with $\|U\|_F = 1$, and put $D_i = \mathrm{diag}(\phi'(Wx_i))$ and $\widetilde{D}_i = \mathrm{diag}(\phi'(W\widetilde{x}_i))$. The directional derivative of the ith output block is $J_i(W)[U] = AD_iUx_i$, and therefore

$$(J(W) - \widetilde{J}(W))[U]_i = AD_iUe_i + A(D_i - \widetilde{D}_i)U\widetilde{x}_i.$$

Assumption A.4 and $\|W\| \le R_W$ give

$$\begin{aligned}\|D_i\| &\le B_1,\\ \|D_i - \widetilde{D}_i\| &= \max_{h\in[H]} |\phi'(w_h^\top x_i) - \phi'(w_h^\top \widetilde{x}_i)|\\ &\le B_2\|We_i\|_\infty \le B_2R_Wd_i,\end{aligned}$$

where $w_h^\top$ is row h of W. Since $\|Ue_i\|_2 \le d_i$ and $\|U\widetilde{x}_i\|_2 \le R_\mu$, it follows blockwise that

$$\|(J(W) - \widetilde{J}(W))[U]_i\|_2 \le \|A\|(B_1 + B_2R_WR_\mu)d_i.$$

Consequently,

$$\|(J(W) - \widetilde{J}(W))[U]\|_2^2 \le \|A\|^2(B_1 + B_2R_WR_\mu)^2\sum_{i=1}^n d_i^2 = \Delta_J^2.$$

Taking the supremum over $\|U\|_F = 1$ proves eq. (A.8) and completes the proof. □

Let $(W_\tau)_{\tau\ge0}$ be GD on the observed-input loss, as in eq. (2), and recall the cluster-centered sequence $(\widetilde{W}_\tau)_{\tau\ge0}$ from Lemma A.6. Define

$$\beta_X := \beta + \Delta_J, \quad \kappa := L\widehat{R}_0 + \beta\beta_X, \quad \gamma := \Delta_J\widehat{R}_0 + \beta_X\Delta_{F,2}. \tag{A.9}$$

Because $\beta_X \ge \beta > 0$, we have $\kappa \ge \beta^2 > 0$. For an integer horizon $T \ge 0$, also let

$$\mathfrak{d}(T) := \frac{\gamma}{\kappa}\left((1+\eta\kappa)^T - 1\right), \quad \mathfrak{D}(T) := \beta_c\,\mathfrak{d}(T), \quad T \in \mathbb{Z}_{\ge0}. \tag{A.10}$$

Here $\mathfrak{d}$ bounds the drift in parameter space and $\mathfrak{D}$ the resulting drift in the outputs, so $\mathfrak{D}$ is the term that enters B_ν in Section 3.2. For each sample $i \in [n]$, we further define the *sample-wise path error*

$$\zeta_T^{(i)} := b_i + \mathfrak{D}(T). \tag{A.11}$$

Write $S_{\mathrm{loc}} := R_{\mathrm{loc}} - (2\beta/\alpha^2)\widehat{R}_0$ for the radius remaining after the reference trajectory's displacement.

Proposition A.8 (Stability of the GD Trajectory)**.** *Under the assumptions of Lemma A.6 and Proposition A.7, if*

$$\mathfrak{d}(T) \le S_{\mathrm{loc}}, \tag{A.12}$$

then, for every $0 \le \tau \le T$*,*

$$\|W_\tau - \widetilde{W}_\tau\|_F \le \mathfrak{d}(\tau), \tag{A.13}$$

$$\max_{k\in[K]} \|F_{W_\tau}(\mu_k) - F_{\widetilde{W}_\tau}(\mu_k)\|_2 \le \mathfrak{D}(\tau), \tag{A.14}$$

$$\|F_{W_\tau}(x_i) - F_{\widetilde{W}_\tau}(\widetilde{x}_i)\|_2 \le \zeta_\tau^{(i)} \quad \textit{for every } i \in [n]. \tag{A.15}$$

Proof. Set $d_\tau^W := \|W_\tau - \widetilde{W}_\tau\|_F$ and $p_\tau := \|\mathcal{F}_{W_\tau}(X) - \mathcal{F}_{\widetilde{W}_\tau}(\widetilde{X})\|_2$. Let $r_\tau = \mathcal{F}_{W_\tau}(X) - X$ and recall $\widetilde{r}_\tau = \mathcal{F}_{\widetilde{W}_\tau}(\widetilde{X}) - X$. We prove simultaneously that, for $0 \le \tau \le T$,

$$d_\tau^W \le \mathfrak{d}(\tau) \le \mathfrak{d}(T), \quad W_\tau \in \mathcal{D}(W_0, R_{\mathrm{loc}}). \tag{A.16}$$

The claim is immediate at $\tau = 0$. Suppose it holds at time $\tau < T$. Lemma A.6 ensures that $\widetilde{W}_\tau$ lies in the local ball and, by its invariant and contracting orthogonal residual components, $\|\widetilde{r}_\tau\|_2 \le R_0 \le \widehat{R}_0$.

Because the local ball is convex, the segment from W_τ to $\widetilde{W}_\tau$ also lies in it. Proposition A.7, Proposition A.5, and the averaged-Jacobian representation give

$$\|J(W_\tau) - \widetilde{J}(\widetilde{W}_\tau)\| \le \Delta_J + L d_\tau^W, \quad \|J(W_\tau)\| \le \Delta_J + \beta = \beta_X, \tag{A.17}$$

$$\begin{aligned} p_\tau &\le \|\mathcal{F}_{W_\tau}(X) - \mathcal{F}_{W_\tau}(\widetilde{X})\|_2 + \|\mathcal{F}_{W_\tau}(\widetilde{X}) - \mathcal{F}_{\widetilde{W}_\tau}(\widetilde{X})\|_2 \\ &\le \Delta_{F,2} + \beta d_\tau^W. \end{aligned} \tag{A.18}$$

Subtracting the two GD updates exactly and inserting $J(W_\tau)^\top \widetilde{r}_\tau$ gives

$$d_{\tau+1}^W \le d_\tau^W + \eta \|J(W_\tau) - \widetilde{J}(\widetilde{W}_\tau)\| \|\widetilde{r}_\tau\|_2 + \eta \|J(W_\tau)\| \|r_\tau - \widetilde{r}_\tau\|_2.$$

Since $r_\tau - \widetilde{r}_\tau$ is exactly the output difference defining p_τ, substituting eqs. (A.17) and (A.18) yields

$$d_{\tau+1}^W \le (1 + \eta\kappa) d_\tau^W + \eta\gamma.$$

The envelope in eq. (A.10) obeys the same recurrence with initial value zero. Hence the induction hypothesis gives $d_{\tau+1}^W \le \mathfrak{d}(\tau+1) \le \mathfrak{d}(T)$. Using Lemma A.6 and the path-radius condition now gives

$$\|W_{\tau+1} - W_0\|_F \le d_{\tau+1}^W + \|\widetilde{W}_{\tau+1} - W_0\|_F \le \mathfrak{d}(T) + \frac{2\beta}{\alpha^2}\widehat{R}_0 \le R_{\mathrm{loc}}.$$

This proves both parts of eq. (A.16) at time $\tau + 1$ and closes the simultaneous induction. In particular, eq. (A.13) holds.

The mean-value representation at the K centers and $\|J_c(W)\| \le \beta_c$ therefore give

$$\|\mathcal{F}_{W_\tau}(\mu) - \mathcal{F}_{\widetilde{W}_\tau}(\mu)\|_2 \le \beta_c d_\tau^W.$$

Each center block is bounded by the full norm, which proves eq. (A.14). Finally, fix i and insert $F_{W_\tau}(\widetilde{x}_i)$. Since $\widetilde{x}_i = \mu_{g(i)}$ is one of the K centers, the first term is controlled by eq. (A.6) and the second by eq. (A.14):

$$\begin{aligned} \|F_{W_\tau}(x_i) - F_{\widetilde{W}_\tau}(\widetilde{x}_i)\|_2 &\le \|F_{W_\tau}(x_i) - F_{W_\tau}(\widetilde{x}_i)\|_2 + \|F_{W_\tau}(\mu_{g(i)}) - F_{\widetilde{W}_\tau}(\mu_{g(i)})\|_2 \\ &\le b_i + \beta_c d_\tau^W \le b_i + \mathfrak{D}(\tau) = \zeta_\tau^{(i)}, \end{aligned}$$

which proves eq. (A.15). □

A.3.2 Post-Entry Center Memorization and Reconstruction Separation

Fix $\nu \ge \nu_0$ with $\nu > 0$. Combining the input perturbation, trajectory drift, and reference residual gives the uniform bound

$$B_\nu(t) := a + \mathfrak{D}(t) + \nu\left(1 + q^{(t-t_\nu)/2}\right), \quad t \ge t_\nu. \tag{A.19}$$

All its constants were defined in the preceding steps of the proof.

Theorem A.9 (Reconstruction-to-center error)**.** *Suppose Assumptions A.1, A.2, A.3, and A.4 hold. Let*

$$0 < \eta \le \min\left\{\frac{\alpha^2}{8\beta^4}, \frac{\alpha^2}{2L\beta^2\widehat{R}_0}\right\}. \tag{A.20}$$

Fix $\nu \ge \nu_0$ with $\nu > 0$ and $T \ge t_\nu$. If

$$\mathfrak{d}(T) \le S_{\mathrm{loc}}, \tag{A.21}$$

then every $t \in [t_\nu, T] \cap \mathbb{Z}$ and every $i \in [n]$ satisfy

$$\|F_{W_t}(x_i) - \mu_{g(i)}\|_2 \le B_\nu(t). \tag{A.22}$$

Proof. Insert the reference reconstruction at the assigned center. Proposition A.8 bounds the difference from the observed reconstruction, and Lemma A.6 bounds the remaining reference error. Their sum is at most $B_\nu(t)$ because $b_i \le a$.

□

For Theorem 3.1, impose the assumptions and step-size bound of Theorem A.9, with $\widehat{R}_0 > 0$, $S_{\mathrm{loc}} > 0$, and $c := \delta/2 - a - \nu > 0$. For the prescribed integer $\ell \ge 1$, set

$$s_c := \left\lceil \frac{2\log_+(2\nu/c)}{\log(1/q)} \right\rceil, \quad t_0 := t_\nu + s_c, \quad t_+ := t_0 + \ell. \tag{A.23}$$

The sufficient interval condition is

$$\gamma\big((1+\eta\kappa)^{t_+} - 1\big) < \kappa \min\{S_{\mathrm{loc}}, c/(2\beta_c)\}. \tag{A.24}$$

It ensures that the decreasing reference residual and increasing path error are both small enough on $[t_0, t_+]$. Define the interval by

$$\mathcal{T}_\nu := \{\tau \in \mathbb{Z}_{\ge t_\nu} : \mathfrak{d}(\tau) \le S_{\mathrm{loc}},\ 2B_\nu(\tau) < \delta\}.$$

Proof of Theorem 3.1. The step-size condition implies $0 < q < 1$. The definition of s_c gives

$$\nu q^{s_c/2} \le c/2.$$

Indeed, this is immediate if $\nu \le c/2$; otherwise it follows by taking logarithms and using the ceiling in eq. (A.23). For any $t \in [t_0, t_+]$, the decreasing reference term therefore satisfies $\nu q^{(t-t_\nu)/2} \le c/2$. Meanwhile, eq. (A.24) and monotonicity of $\mathfrak{d}$ imply

$$\mathfrak{d}(t) \le \mathfrak{d}(t_+) < S_{\mathrm{loc}}, \quad \mathfrak{D}(t) < c/2.$$

Theorem A.9 applies throughout this interval, and

$$B_\nu(t) < a + c/2 + \nu + c/2 = \delta/2.$$

For any other center μ_b, $b \ne g(i)$, the reverse triangle inequality gives

$$\|F_{W_t}(x_i) - \mu_b\|_2 \ge \delta - \|F_{W_t}(x_i) - \mu_{g(i)}\|_2 > \|F_{W_t}(x_i) - \mu_{g(i)}\|_2.$$

Thus each reconstruction has the same unique nearest center as its assigned input.

To establish the interval claim, extend the bound to real $t \ge t_\nu$. Both $(1+\eta\kappa)^t$ and $q^{(t-t_\nu)/2}$ have nonnegative second derivatives, so $B_\nu(t)$ is convex. Its strict sublevel set $\{t : 2B_\nu(t) < \delta\}$ is an interval. Since $\mathfrak{d}$ is nondecreasing, intersecting this interval with $\{t \ge t_\nu : \mathfrak{d}(t) \le S_{\mathrm{loc}}\}$ and then with the integers gives consecutive integers. The explicit subinterval already constructed proves nonemptiness and the lower bound of $\ell + 1$ iterations. If $\gamma = 0$, the radius condition holds at every horizon and the interval may be unbounded; the proof remains valid without division by γ. □

Proof of Corollary 3.3. For $j \in \mathcal{I}$, the definition of ε_c in Section 3.1 and eq. (A.22) give $\|x_j - F_{W_t}(x_j)\|_2 \le \varepsilon_c + B_\nu(t)$.

For an outlier satisfying $d_i > \delta + \varepsilon_c$ and $t \in \mathcal{T}_\nu$,

$$\|x_i - F_{W_t}(x_i)\|_2 \ge d_i - B_\nu(t).$$

Since $2B_\nu(t) < \delta$ on $\mathcal{T}_\nu$,

$$d_i - B_\nu(t) > \delta + \varepsilon_c - B_\nu(t) > \varepsilon_c + B_\nu(t),$$

so the outlier's reconstruction error strictly exceeds that of every inlier. The same argument proves Theorem 3.2 under its direct sufficient condition.

□

B Detailed Experimental Analysis

B.1 Data Description

Real Benchmark Datasets. The broader benchmark study evaluates a total of 57 datasets from `ADBench` (Han et al., 2022), including 47 tabular datasets, 5 image datasets, and 5 text datasets. The tabular datasets cover diverse application domains, including healthcare, finance, physics, and astronautics. Table B.1 summarizes the basic information of all datasets. For additional details, we refer the reader to the official `ADBench` GitHub repository[1].

All input features, including both raw features and learned representations, are preprocessed using min-max scaling. For tabular datasets without learned representations, we directly use the scaled raw features as inputs. For image and text datasets, we use the pre-computed embedding features provided in `ADBench`: image datasets are represented by features from a pre-trained Vision Transformer (ViT, Dosovitskiy et al., 2021) and text datasets by BERT embeddings (Devlin et al., 2019). For several tabular datasets, we also consider representations extracted using TabPFN-v2 (Ye et al., 2025) from the official GitHub repository.[2] Since TabPFN requires a supervised setup, we treat the first column as the target variable and the remaining columns as input features, and extract the 192-dimensional representation of the resulting model. When the first column has negligible variance, the second column is used instead. To ensure computational feasibility and representation quality, we restrict the use of TabPFN representation to tabular datasets of moderate size, $1{,}000 < n < 50{,}000$ and $p \leq 62$. The tabular datasets used for TabPFN representation are `ALOI`, `annthyroid`, `campaign`, `landsat`, `magic.gamma`, `mammography`, `PageBlocks`, `shuttle`, `SpamBase`, `Wilt`, and `yeast`.

B.2 Implementation Details

As described in Appendix B.1, all datasets are preprocessed using min-max scaling so that each feature lies in $[0, 1]$. All ODIM and ALTBI experiments are run on a single NVIDIA H100 GPU. The simulations are run on a single NVIDIA RTX 3090 using `PyTorch`. Each reported result is averaged over independent runs with multiple random initializations: five decoder initializations for the simulations, ten for experiment (ii), and three for both ODIM and ALTBI. The remainder of this subsection details the architecture and optimization choices for the simulation study (Section 4.1) and the real-world experiments (Section 4.2).

Simulation Study. We train a single-hidden-layer autoencoder $F_W(x) = A \tanh(Wx)$ with hidden size $H = 32$, except in the cluster-balance experiment (ii), which uses $H = 12$. Both the encoder weight W and the decoder weight A are initialized using the Kaiming uniform scheme (He et al., 2015). The decoder is then fixed and only the encoder is trained. Training uses full-batch gradient descent on $\mathcal{L}(W)/n$ with a constant learning rate of 0.1 and $40{,}000$ iterations, without momentum, weight decay or a schedule. The IM window is computed as the first contiguous interval of training epochs τ during which $\mathrm{AUROC}(\tau) \geq 0.9$, since under this optimizer the trajectory crosses 0.9 again much later.

For experiment (iv), the three initializations with different $\widetilde{R}_0$ are obtained by sampling three random initializations using the Kaiming uniform scheme (He et al., 2015).

Real-World Experiments: Backbones and Baselines. For the two IM-based backbones, ODIM (Kim et al., 2024) and ALTBI (Cho et al., 2025), we follow the original source code published by their authors, including the network architectures, optimizers, batch sizes, and stopping criteria. Since ODIM and ALTBI share the same implementation details, we adopt the configuration provided in the official ODIM GitHub repository.[3] Unless otherwise specified, we use min-max scaled raw input features for tabular data, ViT-based representations followed by min-max scaling for image data, and BERT-based representations followed by min-max scaling for text data. At the beginning of each training run, we maintain an exponential moving average (EMA) of the network parameters, $\bar{W}_t = \alpha \bar{W}_{t-1} + (1 - \alpha) W_t$, with decay $\alpha = 0.999$ over the first m optimizer updates, where

[1] https://github.com/Minqi824/ADBench
[2] https://github.com/PriorLabs/TabPFN
[3] https://github.com/jshwang0311/ODIM

Table B.1: *Description of* `ADBench` *benchmark datasets.*

Number	Dataset Name	#Samples	#Features	#Outlier	%Outlier	Category
1	ALOI	49534	27	1508	3.04	Image
2	annthyroid	7200	6	534	7.42	Healthcare
3	backdoor	95329	196	2329	2.44	Network
4	breastw	683	9	239	34.99	Healthcare
5	campaign	41188	62	4640	11.27	Finance
6	cardio	1831	21	176	9.61	Healthcare
7	Cardiotocography	2114	21	466	22.04	Healthcare
8	celeba	202599	39	4547	2.24	Image
9	census	299285	500	18568	6.2	Sociology
10	cover	286048	10	2747	0.96	Botany
11	donors	619326	10	36710	5.93	Sociology
12	fault	1941	27	673	34.67	Physical
13	fraud	284807	29	492	0.17	Finance
14	glass	214	7	9	4.21	Forensic
15	Hepatitis	80	19	13	16.25	Healthcare
16	http	567498	3	2211	0.39	Web
17	InternetAds	1966	1555	368	18.72	Image
18	Ionosphere	351	32	126	35.9	Oryctognosy
19	landsat	6435	36	1333	20.71	Astronautics
20	letter	1600	32	100	6.25	Image
21	Lymphography	148	18	6	4.05	Healthcare
22	magic.gamma	19020	10	6688	35.16	Physical
23	mammography	11183	6	260	2.32	Healthcare
24	mnist	7603	100	700	9.21	Image
25	musk	3062	166	97	3.17	Chemistry
26	optdigits	5216	64	150	2.88	Image
27	PageBlocks	5393	10	510	9.46	Document
28	pendigits	6870	16	156	2.27	Image
29	Pima	768	8	268	34.9	Healthcare
30	satellite	6435	36	2036	31.64	Astronautics
31	satimage-2	5803	36	71	1.22	Astronautics
32	shuttle	49097	9	3511	7.15	Astronautics
33	skin	245057	3	50859	20.75	Image
34	smtp	95156	3	30	0.03	Web
35	SpamBase	4207	57	1679	39.91	Document
36	speech	3686	400	61	1.65	Linguistics
37	Stamps	340	9	31	9.12	Document
38	thyroid	3772	6	93	2.47	Healthcare
39	vertebral	240	6	30	12.5	Biology
40	vowels	1456	12	50	3.43	Linguistics
41	Waveform	3443	21	100	2.9	Physics
42	WBC	223	9	10	4.48	Healthcare
43	WDBC	367	30	10	2.72	Healthcare
44	Wilt	4819	5	257	5.33	Botany
45	wine	129	13	10	7.75	Chemistry
46	WPBC	198	33	47	23.74	Healthcare
47	yeast	1484	8	507	34.16	Biology
48	CIFAR10	5263	512	263	5	Image
49	FashionMNIST	6315	512	315	5	Image
50	MNIST-C	10000	512	500	5	Image
51	MVTec-AD	5354	512	1258	23.5	Image
52	SVHN	5208	512	260	5	Image
53	Agnews	10000	768	500	5	NLP
54	Amazon	10000	768	500	5	NLP
55	Imdb	10000	768	500	5	NLP
56	Yelp	10000	768	500	5	NLP
57	20news	11905	768	591	4.96	NLP

$m = 170$ for ODIM and 40 for ALTBI. The averaged parameters $\bar{W}_m$ are then used as the initial weights for the main training, after which the standard (non-EMA) optimizer dynamics resume. A decay of 0.999 corresponds to an effective averaging window of order 10^3 iterations, which is long enough that the warm-up phase emphasizes the slowly-changing inlier-driven component of the gradient and suppresses the irregular signal contributed by outliers, consistent with the role of $\widetilde{R}_0$ identified in Theorem 3.1. Since the average retains a fraction 0.999^m of the initialization, the warm-up length matters: a warm-up of ten updates retains 0.99 of it and leaves the initialization essentially unchanged.

The modified ODIM and ALTBI with the two proposed strategies (denoted ODIM* and ALTBI*, respectively, in Figure B.4 and Figure B.5) replace the input with a pre-trained representation (TabPFN for tabular data, ViT for image data) and warm up the parameters with the EMA scheme.

The 22 baselines cover classical machine-learning detectors, deep autoencoder-based methods, and diffusion-based detectors. These methods include traditional machine-learning approaches implemented in `ADBench`, such as kNN Ramaswamy et al. (2000), LOF Breunig et al. (2000), OCSVM Schölkopf et al. (2001), CBLOF He et al. (2003), PCA Shyu et al. (2003), FeatureBagging Lazarevic & Kumar (2005), IForest (Liu et al., 2008), MCD Fauconnier & Haesbroeck (2009), HBOS Goldstein & Dengel (2012), LODA Pevnỳ (2016), COPOD Li et al. (2020b), and ECOD Li et al. (2022). We also consider two deep learning-based UOD methods, DAGMM Zong et al. (2018) and DeepSVDD (Ruff et al., 2018), which are available in `ADBench`, as well as more recent approaches beyond `ADBench`, including DROCC Goyal et al. (2020), ICL (Shenkar & Wolf, 2022), GOAD Bergman & Hoshen (2020), and DTE Livernoche et al. (2023). For fair comparison, the results of the 22 baseline methods are directly taken from the supplementary materials of Livernoche et al. (2023), without re-running the corresponding models. In contrast, we implement ODIM, ALTBI, and their modified versions ourselves.

B.3 Detailed Experiment Results

Extension to Practical Training Settings (Section 4.1). The four experiments of Section 4.1 adopt the fixed-decoder, encoder-only full-batch gradient-descent setup analyzed in our theory. To examine whether the same qualitative relationships persist under a more practical training setup, we repeat all four of them with joint encoder-decoder training and mini-batch Adam (Kingma & Ba, 2014). The factor-wise trends reported below remain qualitatively similar, providing empirical support for the broader relevance of our theoretical insights beyond the formal setting, although these results do not constitute a formal extension of the theorem. We present the results showing how ε_c, $(n_{\min}, n_{\max})$, ρ, and $\widetilde{R}_0$ each affect the IM effect under joint training with the Adam optimizer.

(i) **Effect of within-cluster variation** ε_c: We vary $\varepsilon_c \in \{1.0, 2.5, 4.0\}$ while keeping all other terms fixed. Figure B.1 shows the data distributions and AUROC trajectories of the trained autoencoder. As suggested by Theorem 3.1, tighter clusters (smaller ε_c) yield a stronger and more sustained IM effect: the peak AUROC increases from 0.944 ($\varepsilon_c = 4.0$) to 0.975 ($\varepsilon_c = 1.0$), and the IM window widens from 193 to 372 epochs. This result confirms that smaller within-cluster variation makes the IM effect more stable and more pronounced.

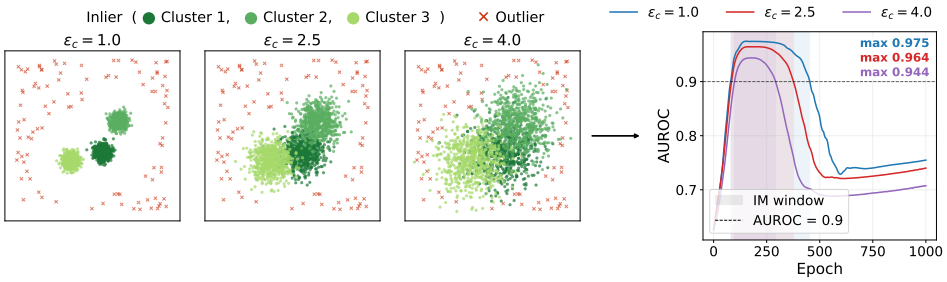


Figure B.1: *Effect of within-cluster variation* ε_c. (Left) Data distributions for $\varepsilon_c \in \{1.0, 2.5, 4.0\}$. (Right) AUROC trajectories; shaded regions indicate the IM window.

(ii) **Effect of cluster-size balance**: With fixed $\varepsilon_c = 1.5$, we vary the cluster-size ratio $n_{\max} : n_{\min}$ across three levels: balanced (1:1), moderate (11:1), and heavy imbalance (27:1). The smallest cluster is assigned to a fixed spatial position across settings for fair comparison. Figure B.2 shows that heavier imbalance degrades the persistence of the IM effect: the IM window narrows from 596 epochs (balanced) to 462 epochs (heavy imbalance). This is consistent with Theorem 3.1, which suggests that imbalanced cluster sizes lead to a shorter duration of the IM effect.

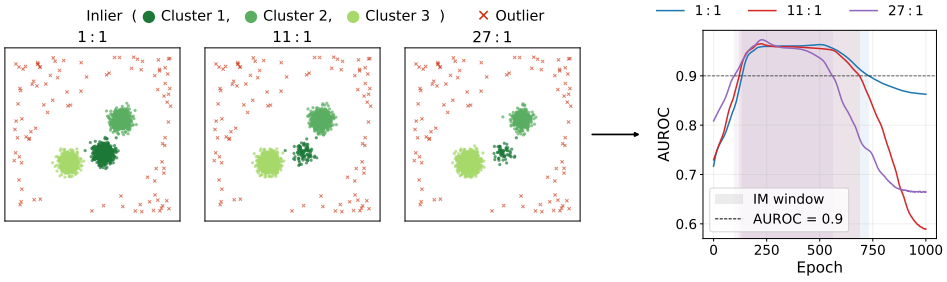


Figure B.2: *Effect of cluster-size balance* $n_{\max} : n_{\min}$. (Left) Data distributions for $(n_{\max} : n_{\min}) \in \{(1:1), (11:1), (27:1)\}$. (Right) AUROC trajectories; shaded regions indicate the IM window.

(iii) **Effect of outlier rate** ρ: We fix $\varepsilon_c = 2.0$ and vary $\rho \in \{0.03, 0.10, 0.25\}$. Figure B.3 shows that a higher ρ substantially shortens the IM window from 391 epochs ($\rho = 0.03$) to 94 epochs ($\rho = 0.25$). This is consistent with our theoretical analysis, which suggests that a smaller ρ leads to a prolonged IM effect. Interestingly, we also observe that the strength of the IM effect improves as ρ decreases, as the peak AUROC increases from 0.929 to 0.968. While this behavior is not directly explained by the current theory, it suggests that a lower ρ may further facilitate the distinction between inliers and outliers in practice.

(iv) **Effect of initialization quality** $\widetilde{R}_0$: We fix $\varepsilon_c = 3.0$ and examine how the initial reconstruction quality affects the IM window. As discussed in Section 3.2, a low $\widetilde{R}_0 = \|\mathcal{F}_{W_0}(\widetilde{X}) - \widetilde{X}\|_2$ indicates a

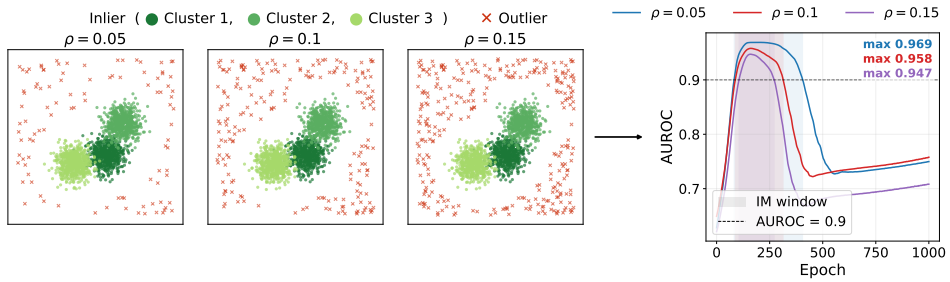


Figure B.3: *Effect of outlier rate ρ.* (Left) Data distributions for $\rho \in \{0.03, 0.10, 0.25\}$. (Right) AUROC trajectories; shaded regions indicate the IM window.

high-quality initial parameter W_0. Table B.2 reports the IM window $[T_1, T_2]$ and the best AUROC for three initializations with different $\widetilde{R}_0$. A smaller $\widetilde{R}_0$ produces a later T_2 and thus a wider IM window at a similar best AUROC, supporting our theoretical finding that a smaller $\widetilde{R}_0$ lets the autoencoder capture the inlier cluster structure early, thereby prolonging the IM effect.

Interestingly, although not predicted by the theory, T_1 also decreases as $\widetilde{R}_0$ becomes smaller, suggesting that better initialization accelerates the onset of the IM effect in practice.

Table B.2: *Effect of initialization quality $\widetilde{R}_0$* on the IM window and detection performance (best AUROC).

$\widetilde{R}_0$	T_1	T_2	IM window length $(T_2 - T_1)$	Best AUROC
0.40	37	509	472	0.961
0.55	56	433	377	0.962
0.60	94	456	362	0.964

Detailed Results for the Eleven Tabular Datasets (Section 4.2). Table B.3 provides the numerical values underlying Figure 5 for both ODIM and ALTBI. For each dataset and method, "AUROC" is the score obtained with the original min-max-scaled tabular input and without EMA, while $\Delta := \mathrm{AUROC}(+\mathrm{EMA} + \mathrm{TabPFN}) - \mathrm{AUROC}(\mathrm{Original})$ is the change after applying both proposed strategies.

Table B.3: *Full results for Figure 5.* Δ is the AUROC change after applying EMA warm-up and TabPFN preprocessing together. Positive changes are bolded.

		ODIM		ALTBI	
Dataset	σ	AUROC	Δ	AUROC	Δ
ALOI	0.22	0.525	**+0.081**	0.521	**+0.079**
Wilt	0.26	0.364	**+0.037**	0.349	**+0.060**
SpamBase	0.33	0.512	**+0.155**	0.491	**+0.179**
yeast	0.36	0.398	**+0.073**	0.411	**+0.059**
annthyroid	0.49	0.632	**+0.054**	0.633	**+0.058**
landsat	0.49	0.499	**+0.101**	0.527	**+0.042**
campaign	0.58	0.729	**+0.006**	0.726	**+0.001**
mammography	0.60	0.813	−0.040	0.808	−0.044
magic.gamma	0.70	0.724	−0.052	0.746	−0.052
PageBlocks	1.13	0.864	−0.050	0.868	−0.061
shuttle	4.04	0.982	−0.273	0.983	**+0.002**

AUROC Comparison (Section 4.2). Figure B.4 compares the original and modified ODIM and ALTBI variants with the 22 reference baselines in AUROC. It reports the average across the 47 tabular datasets and the individual results on `CIFAR10`, `FashionMNIST`, and `SVHN`.

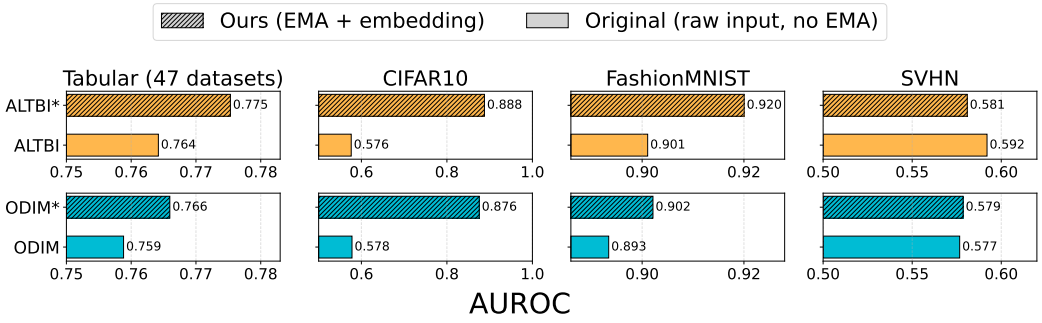


Figure B.4: *Average over 47 tabular datasets (leftmost) and results on `CIFAR10`, `FashionMNIST`, and `SVHN`.* Hatched bars (ALTBI* and ODIM*) are the proposed variants combining the applicable pre-trained representation with EMA warm-up, while the plain bars (ALTBI and ODIM) use each benchmark's original input representation without EMA.

AUPRC Comparison (Section 4.2). Figure B.5 presents the performance comparison under the AUPRC metric. On the tabular datasets (leftmost panel), ALTBI* and ODIM* remain among the top-performing methods, although the gap to machine-learning-based baselines such as MCD and IForest narrows compared to AUROC. This result reflects the well-known sensitivity of AUPRC to the outlier proportion, since on a subset of `ADBench` datasets with extremely low contamination rates, isolation- and density-based detectors can retain a relative advantage. On the image datasets, the improvement from the pre-trained representation is even more pronounced under AUPRC than under AUROC. That is, ALTBI* and ODIM* dominate all baselines on CIFAR10 and FashionMNIST by a large margin.

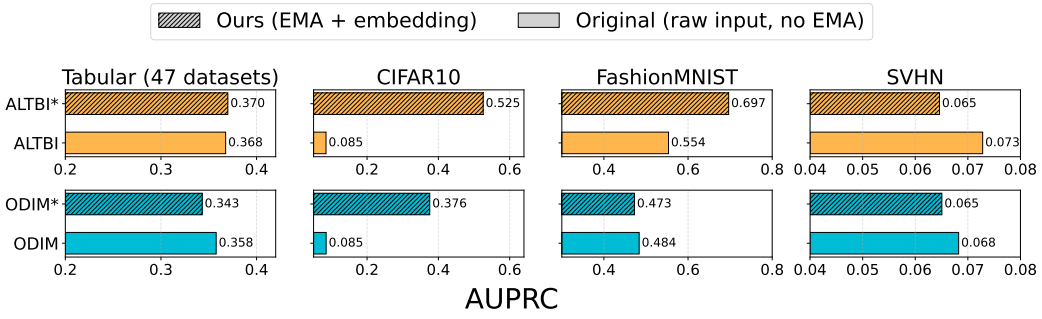


Figure B.5: *Unsupervised outlier detection performance (AUPRC) on the `ADBench` benchmark (Han et al., 2022).* (Left → Right) average over 47 tabular datasets and results on CIFAR10, FashionMNIST, and SVHN. Hatched bars (ALTBI* and ODIM*) denote the proposed variants that combine the EMA-based initialization with a pre-trained representation.

Additional Results (Section 4.2). Table B.4 reports the AUROC of IM-based methods, ALTBI and ODIM, on the 47 tabular ADBench datasets under three configurations: (i) the vanilla raw input training without the proposed techniques, corresponding to the original implementations of ODIM and ALTBI, (ii) with the EMA warm-up applied to early iterations (+EMA), and (iii) with both the EMA warm-up and a TabPFN embedding (+EMA +TabPFN), where the TabPFN representation is applied only to the subset of datasets described in Appendix B.1. The EMA warm-up alone improves the AUROC of both methods by 0.6 to 0.7 points, supporting the role of small $\widetilde{R}_0$ in widening the IM window. Using the TabPFN representation gives a further AUROC improvement for ALTBI, while for ODIM the two strategies yield comparable AUROC. Because the representation changes only eleven of the 47 datasets, these averages understate its effect where it is applied, and Figure 5 reports the results for each of those datasets.

Table B.4: *Comparison of performance on 47 tabular ADBench datasets* (Han et al., 2022) in terms of the two proposed strategies. Best results are **bolded**.

Method	ALTBI	ODIM
Raw (Vanilla)	0.764	0.759
+EMA	0.770	**0.766**
+EMA +TabPFN (ours)	**0.775**	**0.766**

Table B.5 reports the analogous comparison on the three image datasets. As in the tabular case, 'Raw (Vanilla)' denotes training on raw inputs without EMA warm-up, '+EMA' applies EMA to raw inputs, and '+EMA +ViT' further combines EMA with the ViT representation. Using EMA alone does not provide a significant improvement.

In contrast, '+EMA +ViT' improves the performance of ALTBI on CIFAR10 and FashionMNIST, and consistently improves the performance of ODIM on these datasets, while yielding comparable AUROC on SVHN. This is because, on CIFAR10 and FashionMNIST, the ViT embedding tightens the inlier cluster structure, allowing the EMA warm-up to further enlarge the IM window. In contrast, on SVHN, the ViT embedding does not produce sufficiently compact inlier clusters for the EMA warm-up to yield additional gains.

Table B.5: *Per-component ablation on the three image benchmarks*, in AUROC averaged over the ten normal-class settings. '+ViT' applies the original ALTBI (Cho et al., 2025) or ODIM (Kim et al., 2024) method on a ViT embedding (Dosovitskiy et al., 2021), and '+EMA +ViT' additionally applies the EMA warm-up. Best results are **bolded**.

Method	ALTBI			ODIM		
	CIFAR10	FashionMNIST	SVHN	CIFAR10	FashionMNIST	SVHN
Raw (Vanilla)	0.576	0.901	**0.592**	0.578	0.893	0.577
+EMA	0.576	0.902	**0.592**	0.579	0.892	0.576
+EMA +ViT (ours)	**0.888**	**0.920**	0.581	**0.876**	**0.902**	**0.579**

Full-Benchmark Results (Section 4.2). Table B.6 and Table B.7 compare the IM-based methods (ALTBI and ODIM) on all 57 ADBench datasets, under AUROC and AUPRC. In particular, we compare three variants: Original (the original ALTBI and ODIM using ADBench's default inputs for each data type, i.e., raw features for tabular data and ViT/BERT embeddings for image/text data), +EMA (applying EMA warm-up on the same inputs), and +EMA +TabPFN (additionally replacing a subset of tabular inputs with TabPFN embeddings). For average AUROC, ALTBI rises from 0.757 (Original) to 0.762 (+EMA) and 0.766 (Ours), while ODIM changes from 0.751 to 0.756 and 0.757.

Tables B.8 and B.9 compare the modified ALTBI variant with 22 existing baselines and with the internally implemented ODIM and ALTBI methods. All ODIM variants are reported in Tables B.6 and B.7. The external baseline results are taken directly from (Livernoche et al., 2023), whereas our ALTBI and ODIM results are averaged over three random initializations. Because the original ALTBI and ODIM inputs already use ViT or BERT embeddings for image and text datasets, the

corresponding modification there is the EMA warm-up, and TabPFN replacement applies only to the selected tabular datasets. Among the methods reported in Tables B.8 and B.9, the modified ALTBI variant has the highest mean AUROC and AUPRC.

Table B.6: *Per-dataset AUROC of the variants of the IM-based methods* (Original, +EMA, Ours) on the 57 ADBench datasets. The best average AUROC is **bolded**.

Number	Dataset Name	ALTBI			ODIM		
		Original	+EMA	Ours	Original	+EMA	Ours
1	ALOI	0.521	0.526	0.601	0.525	0.530	0.606
2	annthyroid	0.633	0.642	0.691	0.632	0.625	0.685
3	backdoor	0.876	0.868	0.868	0.909	0.897	0.897
4	breastw	0.981	0.982	0.982	0.980	0.978	0.978
5	campaign	0.726	0.726	0.726	0.729	0.735	0.735
6	cardio	0.845	0.840	0.840	0.869	0.876	0.876
7	Cardiotocography	0.540	0.530	0.530	0.542	0.534	0.534
8	celeba	0.804	0.797	0.797	0.782	0.760	0.760
9	census	0.670	0.672	0.672	0.682	0.685	0.685
10	cover	0.897	0.910	0.910	0.920	0.914	0.914
11	donors	0.609	0.650	0.650	0.606	0.624	0.624
12	fault	0.661	0.658	0.658	0.678	0.671	0.671
13	fraud	0.920	0.925	0.925	0.913	0.920	0.920
14	glass	0.748	0.771	0.771	0.757	0.769	0.769
15	Hepatitis	0.811	0.818	0.818	0.703	0.719	0.719
16	http	0.996	0.996	0.996	0.996	0.996	0.996
17	InternetAds	0.733	0.745	0.745	0.731	0.734	0.734
18	Ionosphere	0.850	0.883	0.883	0.862	0.860	0.860
19	landsat	0.527	0.539	0.569	0.499	0.545	0.600
20	letter	0.722	0.733	0.733	0.730	0.736	0.736
21	Lymphography	0.988	0.988	0.988	0.980	0.994	0.994
22	magic.gamma	0.746	0.745	0.694	0.724	0.727	0.672
23	mammography	0.808	0.814	0.763	0.813	0.816	0.774
24	mnist	0.796	0.794	0.794	0.791	0.779	0.779
25	musk	1.000	1.000	1.000	0.971	0.945	0.945
26	optdigits	0.602	0.684	0.684	0.552	0.697	0.697
27	PageBlocks	0.868	0.882	0.807	0.864	0.861	0.815
28	pendigits	0.939	0.957	0.957	0.911	0.918	0.918
29	Pima	0.716	0.713	0.713	0.712	0.718	0.718
30	satellite	0.734	0.760	0.760	0.702	0.740	0.740
31	satimage-2	0.997	0.998	0.998	0.997	0.997	0.997
32	shuttle	0.983	0.984	0.985	0.982	0.981	0.709
33	skin	0.869	0.865	0.865	0.873	0.877	0.877
34	smtp	0.823	0.845	0.845	0.804	0.840	0.840
35	SpamBase	0.491	0.512	0.670	0.512	0.513	0.667
36	speech	0.471	0.468	0.468	0.471	0.469	0.469
37	Stamps	0.878	0.883	0.883	0.898	0.896	0.896
38	thyroid	0.938	0.939	0.939	0.938	0.942	0.942
39	vertebral	0.385	0.373	0.373	0.374	0.360	0.360
40	vowels	0.878	0.882	0.882	0.862	0.874	0.874
41	Waveform	0.749	0.713	0.713	0.763	0.780	0.780
42	WBC	0.994	0.993	0.993	0.992	0.993	0.993
43	WDBC	0.985	0.981	0.981	0.993	0.986	0.986
44	Wilt	0.349	0.375	0.409	0.364	0.382	0.401
45	wine	0.948	0.940	0.940	0.875	0.863	0.863
46	WPBC	0.506	0.503	0.503	0.502	0.526	0.526
47	yeast	0.411	0.411	0.469	0.398	0.409	0.471
48	CIFAR10	0.889	0.888	0.888	0.877	0.876	0.876
49	FashionMNIST	0.921	0.920	0.920	0.902	0.902	0.902
50	MNIST-C	0.739	0.741	0.741	0.726	0.727	0.727
51	MVTec-AD	0.867	0.860	0.860	0.849	0.845	0.845
52	SVHN	0.581	0.581	0.581	0.579	0.579	0.579
53	Agnews	0.868	0.864	0.864	0.862	0.862	0.862
54	Amazon	0.546	0.548	0.548	0.544	0.539	0.539
55	Imdb	0.523	0.527	0.527	0.524	0.526	0.526
56	Yelp	0.555	0.557	0.557	0.558	0.554	0.554
57	20news	0.725	0.725	0.725	0.725	0.722	0.722
	Average	0.757	0.762	**0.766**	0.751	0.756	**0.757**

Table B.7: *Per-dataset AUPRC of the variants of the IM-based methods* (Original, +EMA, Ours) on the 57 ADBench datasets. The best average AUPRC is **bolded**.

Number	Dataset Name	ALTBI			ODIM		
		Original	+EMA	Ours	Original	+EMA	Ours
1	ALOI	0.035	0.035	0.057	0.037	0.037	0.064
2	annthyroid	0.156	0.155	0.181	0.154	0.154	0.192
3	backdoor	0.138	0.122	0.122	0.406	0.396	0.396
4	breastw	0.943	0.952	0.952	0.930	0.915	0.915
5	campaign	0.266	0.247	0.247	0.246	0.243	0.243
6	cardio	0.539	0.519	0.519	0.526	0.513	0.513
7	Cardiotocography	0.424	0.409	0.409	0.390	0.402	0.402
8	celeba	0.098	0.090	0.090	0.064	0.054	0.054
9	census	0.090	0.089	0.089	0.096	0.097	0.097
10	cover	0.047	0.051	0.051	0.068	0.060	0.060
11	donors	0.071	0.079	0.079	0.072	0.076	0.076
12	fault	0.477	0.475	0.475	0.495	0.487	0.487
13	fraud	0.299	0.323	0.323	0.329	0.283	0.283
14	glass	0.083	0.094	0.094	0.104	0.133	0.133
15	Hepatitis	0.454	0.459	0.459	0.242	0.250	0.250
16	http	0.296	0.309	0.309	0.285	0.289	0.289
17	InternetAds	0.365	0.405	0.405	0.336	0.343	0.343
18	Ionosphere	0.801	0.839	0.839	0.818	0.812	0.812
19	landsat	0.202	0.208	0.219	0.192	0.207	0.238
20	letter	0.120	0.130	0.130	0.135	0.132	0.132
21	Lymphography	0.782	0.705	0.705	0.563	0.846	0.846
22	magic.gamma	0.689	0.686	0.643	0.664	0.672	0.577
23	mammography	0.067	0.069	0.070	0.083	0.083	0.080
24	mnist	0.396	0.391	0.391	0.334	0.330	0.330
25	musk	1.000	1.000	1.000	0.537	0.316	0.316
26	optdigits	0.038	0.046	0.046	0.030	0.045	0.045
27	PageBlocks	0.553	0.584	0.584	0.543	0.543	0.457
28	pendigits	0.253	0.335	0.335	0.173	0.187	0.187
29	Pima	0.498	0.494	0.494	0.501	0.510	0.510
30	satellite	0.681	0.698	0.698	0.634	0.670	0.670
31	satimage-2	0.907	0.920	0.920	0.948	0.942	0.942
32	shuttle	0.959	0.960	0.972	0.944	0.949	0.302
33	skin	0.449	0.441	0.441	0.457	0.462	0.462
34	smtp	0.079	0.033	0.033	0.507	0.409	0.409
35	SpamBase	0.380	0.389	0.528	0.394	0.392	0.512
36	speech	0.017	0.017	0.017	0.018	0.019	0.019
37	Stamps	0.309	0.312	0.312	0.338	0.331	0.331
38	thyroid	0.228	0.220	0.220	0.244	0.253	0.253
39	vertebral	0.097	0.095	0.095	0.097	0.094	0.094
40	vowels	0.308	0.297	0.297	0.242	0.290	0.290
41	Waveform	0.129	0.113	0.113	0.180	0.217	0.217
42	WBC	0.862	0.869	0.869	0.859	0.876	0.876
43	WDBC	0.625	0.530	0.530	0.783	0.568	0.568
44	Wilt	0.037	0.039	0.042	0.038	0.039	0.042
45	wine	0.502	0.416	0.416	0.259	0.233	0.233
46	WPBC	0.230	0.233	0.233	0.230	0.242	0.242
47	yeast	0.300	0.299	0.331	0.291	0.298	0.337
48	CIFAR10	0.526	0.525	0.525	0.378	0.376	0.376
49	FashionMNIST	0.695	0.697	0.697	0.473	0.473	0.473
50	MNIST-C	0.304	0.313	0.313	0.213	0.215	0.215
51	MVTec-AD	0.535	0.534	0.534	0.510	0.503	0.503
52	SVHN	0.065	0.065	0.065	0.065	0.065	0.065
53	Agnews	0.281	0.264	0.264	0.286	0.285	0.285
54	Amazon	0.054	0.055	0.055	0.054	0.054	0.054
55	Imdb	0.055	0.056	0.056	0.055	0.055	0.055
56	Yelp	0.057	0.057	0.057	0.057	0.057	0.057
57	20news	0.117	0.119	0.119	0.129	0.129	0.129
	Average	0.350	0.349	**0.352**	**0.334**	0.332	0.322

Table B.8: *Per-dataset AUROC comparison on the 57* `ADBench` *datasets.* Methods marked with '*' are implemented and evaluated in our framework, while the remaining baseline results are taken from (Livernoche et al., 2023). Best average AUROC for each method is **bolded**.

	CBLOF	COPOD	ECOD	FeatureBagging	HBOS	IForest	kNN	LODA	LOF	MCD	OCSVM	PCA	DAGMM	DeepSVDD	DROCC	GOAD	ICL	PlanarFlow	DDPM	DTE-NP	DTE-IG	DTE-C	ODIM*	ALTBI*	Ours* (w/ ALTBI)
aloi	0.556	0.515	0.531	0.792	0.531	0.542	0.613	0.495	0.767	0.520	0.549	0.549	0.517	0.514	0.500	0.497	0.548	0.520	0.532	0.645	0.541	0.525	0.525	0.521	0.601
annthyroid	0.676	0.777	0.789	0.788	0.608	0.816	0.761	0.453	0.710	0.918	0.682	0.676	0.548	0.739	0.631	0.453	0.599	0.966	0.814	0.781	0.923	0.964	0.632	0.633	0.691
backdoor	0.897	0.500	0.500	0.790	0.740	0.725	0.826	0.515	0.764	0.848	0.889	0.888	0.752	0.735	0.500	0.587	0.936	0.787	0.892	0.806	0.753	0.875	0.909	0.876	0.868
breastw	0.961	0.994	0.990	0.408	0.984	0.983	0.980	0.970	0.446	0.985	0.935	0.946	0.811	0.625	0.847	0.845	0.807	0.965	0.766	0.976	0.905	0.891	0.980	0.981	0.982
campaign	0.738	0.783	0.769	0.594	0.768	0.704	0.750	0.493	0.614	0.775	0.737	0.734	0.581	0.508	0.500	0.443	0.766	0.566	0.724	0.746	0.660	0.789	0.729	0.726	0.726
cardio	0.832	0.921	0.935	0.579	0.839	0.922	0.830	0.856	0.551	0.815	0.934	0.949	0.625	0.498	0.655	0.908	0.461	0.796	0.723	0.777	0.631	0.721	0.869	0.845	0.840
cardiotocography	0.561	0.664	0.784	0.538	0.595	0.681	0.503	0.708	0.527	0.500	0.691	0.747	0.546	0.488	0.449	0.624	0.372	0.643	0.579	0.493	0.506	0.510	0.542	0.540	0.530
celeba	0.753	0.757	0.763	0.514	0.754	0.707	0.736	0.600	0.432	0.803	0.781	0.792	0.627	0.491	0.726	0.432	0.684	0.703	0.796	0.699	0.700	0.812	0.782	0.804	0.797
census	0.664	0.500	0.500	0.538	0.611	0.607	0.671	0.454	0.562	0.731	0.655	0.662	0.491	0.527	0.443	0.488	0.668	0.604	0.659	0.672	0.629	0.646	0.682	0.670	0.672
cover	0.922	0.882	0.919	0.571	0.707	0.873	0.866	0.922	0.568	0.696	0.952	0.934	0.742	0.580	0.747	0.124	0.681	0.417	0.808	0.838	0.635	0.697	0.920	0.897	0.910
donors	0.808	0.815	0.888	0.691	0.743	0.771	0.829	0.566	0.629	0.765	0.770	0.825	0.558	0.511	0.747	0.225	0.739	0.899	0.806	0.832	0.796	0.785	0.606	0.609	0.650
fault	0.665	0.455	0.468	0.591	0.506	0.544	0.715	0.478	0.579	0.505	0.537	0.480	0.495	0.522	0.668	0.546	0.661	0.469	0.562	0.726	0.577	0.590	0.678	0.661	0.658
fraud	0.954	0.943	0.949	0.616	0.945	0.950	0.955	0.856	0.548	0.911	0.954	0.952	0.857	0.769	0.500	0.724	0.931	0.895	0.924	0.956	0.942	0.938	0.913	0.920	0.925
glass	0.855	0.760	0.710	0.659	0.820	0.790	0.870	0.624	0.618	0.795	0.661	0.715	0.630	0.517	0.743	0.545	0.729	0.766	0.560	0.881	0.681	0.864	0.757	0.748	0.771
hepatitis	0.635	0.807	0.737	0.469	0.768	0.683	0.669	0.557	0.468	0.721	0.704	0.748	0.600	0.361	0.582	0.637	0.616	0.654	0.461	0.631	0.451	0.577	0.703	0.811	0.818
http	0.996	0.991	0.980	0.288	0.991	1.000	0.051	0.060	0.338	1.000	0.994	0.997	0.838	0.249	0.500	0.996	0.921	0.994	0.998	0.051	0.973	0.995	0.996	0.996	0.996
internetads	0.616	0.676	0.677	0.494	0.696	0.686	0.616	0.541	0.587	0.660	0.615	0.609	0.515	0.583	0.500	0.614	0.592	0.608	0.614	0.634	0.635	0.656	0.731	0.733	0.745
ionosphere	0.892	0.783	0.717	0.876	0.544	0.833	0.922	0.788	0.864	0.951	0.838	0.777	0.641	0.514	0.766	0.829	0.629	0.884	0.758	0.924	0.697	0.911	0.862	0.850	0.883
landsat	0.548	0.422	0.368	0.540	0.575	0.474	0.614	0.382	0.549	0.607	0.423	0.366	0.533	0.631	0.626	0.506	0.649	0.464	0.496	0.602	0.473	0.544	0.499	0.527	0.569
letter	0.763	0.560	0.573	0.886	0.589	0.616	0.812	0.537	0.878	0.804	0.598	0.524	0.503	0.517	0.780	0.598	0.737	0.689	0.847	0.850	0.676	0.781	0.730	0.722	0.733
lymphography	0.994	0.996	0.995	0.523	0.995	0.999	0.995	0.900	0.636	0.989	0.996	0.997	0.840	0.681	0.878	0.995	0.884	0.940	0.958	0.989	0.852	0.834	0.980	0.988	0.988
magic.gamma	0.725	0.681	0.638	0.700	0.709	0.721	0.795	0.655	0.678	0.699	0.673	0.667	0.584	0.604	0.728	0.442	0.676	0.742	0.763	0.801	0.782	0.765	0.724	0.746	0.694
mammography	0.795	0.905	0.906	0.726	0.838	0.860	0.852	0.867	0.702	0.690	0.871	0.888	0.719	0.451	0.779	0.414	0.658	0.782	0.749	0.849	0.799	0.810	0.813	0.808	0.763
musk	1.000	0.948	0.953	0.575	1.000	0.998	0.964	0.993	0.581	1.000	1.000	1.000	0.912	0.538	0.575	1.000	0.790	0.748	1.000	0.882	0.785	0.965	0.971	1.000	1.000
optdigits	0.785	0.500	0.500	0.539	0.868	0.696	0.395	0.493	0.538	0.413	0.507	0.518	0.408	0.519	0.565	0.657	0.533	0.492	0.402	0.386	0.513	0.508	0.552	0.602	0.684
pageblocks	0.893	0.875	0.914	0.758	0.779	0.897	0.919	0.712	0.703	0.923	0.914	0.907	0.753	0.592	0.914	0.609	0.768	0.908	0.820	0.906	0.850	0.924	0.864	0.868	0.807
pendigits	0.864	0.906	0.927	0.518	0.925	0.947	0.828	0.895	0.534	0.834	0.929	0.936	0.548	0.383	0.520	0.592	0.650	0.780	0.700	0.786	0.624	0.713	0.911	0.939	0.957
pima	0.655	0.662	0.604	0.573	0.704	0.674	0.723	0.595	0.563	0.686	0.631	0.651	0.522	0.510	0.542	0.606	0.524	0.615	0.537	0.707	0.599	0.624	0.712	0.716	0.713
satellite	0.742	0.633	0.583	0.545	0.762	0.695	0.721	0.614	0.550	0.804	0.662	0.601	0.675	0.562	0.608	0.702	0.627	0.671	0.715	0.702	0.582	0.711	0.702	0.734	0.760
satimage-2	0.999	0.975	0.965	0.526	0.976	0.993	0.992	0.981	0.539	0.995	0.997	0.977	0.911	0.551	0.579	0.996	0.898	0.970	0.996	0.980	0.858	0.946	0.997	0.997	0.998
shuttle	0.621	0.995	0.993	0.493	0.986	0.997	0.732	0.389	0.526	0.990	0.992	0.990	0.898	0.576	0.500	0.208	0.642	0.852	0.975	0.698	0.669	0.976	0.982	0.983	0.985
skin	0.675	0.471	0.490	0.534	0.588	0.670	0.720	0.442	0.550	0.892	0.547	0.447	0.554	0.548	0.708	0.579	0.265	0.773	0.461	0.718	0.741	0.741	0.873	0.869	0.865
smtp	0.863	0.912	0.882	0.794	0.809	0.905	0.933	0.819	0.899	0.948	0.845	0.856	0.868	0.895	0.500	0.915	0.656	0.784	0.956	0.930	0.769	0.951	0.804	0.823	0.845
spambase	0.541	0.688	0.656	0.424	0.664	0.637	0.566	0.480	0.453	0.446	0.534	0.548	0.488	0.584	0.490	0.496	0.459	0.528	0.510	0.545	0.509	0.515	0.512	0.491	0.670
speech	0.471	0.489	0.470	0.509	0.473	0.476	0.480	0.466	0.512	0.494	0.466	0.469	0.522	0.512	0.483	0.458	0.512	0.496	0.466	0.487	0.488	0.495	0.471	0.471	0.468
stamps	0.660	0.929	0.877	0.502	0.904	0.907	0.870	0.831	0.512	0.838	0.882	0.909	0.719	0.465	0.760	0.774	0.505	0.838	0.556	0.820	0.692	0.753	0.898	0.878	0.883
thyroid	0.909	0.939	0.977	0.707	0.948	0.979	0.965	0.819	0.657	0.986	0.958	0.955	0.719	0.505	0.889	0.574	0.693	0.992	0.871	0.964	0.828	0.990	0.938	0.938	0.939
vertebral	0.463	0.263	0.417	0.473	0.317	0.362	0.379	0.294	0.487	0.389	0.426	0.378	0.470	0.394	0.425	0.468	0.449	0.409	0.563	0.400	0.451	0.458	0.374	0.385	0.373
vowels	0.884	0.496	0.593	0.933	0.679	0.763	0.951	0.705	0.932	0.732	0.779	0.604	0.464	0.514	0.738	0.791	0.784	0.888	0.903	0.964	0.705	0.914	0.862	0.878	0.882
waveform	0.701	0.739	0.603	0.715	0.694	0.707	0.750	0.594	0.693	0.572	0.669	0.635	0.523	0.609	0.674	0.592	0.661	0.640	0.617	0.729	0.523	0.602	0.763	0.749	0.713
wbc	0.977	0.994	0.994	0.388	0.987	0.996	0.982	0.992	0.607	0.988	0.987	0.993	0.821	0.503	0.821	0.949	0.853	0.934	0.948	0.979	0.894	0.871	0.992	0.994	0.993
wdbc	0.990	0.993	0.971	0.867	0.989	0.988	0.980	0.980	0.849	0.969	0.984	0.988	0.715	0.602	0.347	0.983	0.738	0.985	0.965	0.975	0.566	0.835	0.993	0.985	0.981
wilt	0.396	0.345	0.394	0.666	0.348	0.451	0.511	0.313	0.678	0.859	0.317	0.239	0.432	0.465	0.400	0.555	0.649	0.794	0.659	0.552	0.834	0.851	0.364	0.349	0.409
wine	0.453	0.865	0.738	0.323	0.907	0.786	0.470	0.822	0.330	0.975	0.671	0.819	0.513	0.507	0.621	0.734	0.455	0.390	0.374	0.425	0.310	0.557	0.875	0.948	0.940
wpbc	0.487	0.519	0.489	0.436	0.548	0.516	0.512	0.501	0.447	0.534	0.485	0.486	0.449	0.493	0.483	0.466	0.488	0.483	0.493	0.502	0.489	0.468	0.502	0.506	0.503
yeast	0.461	0.380	0.443	0.465	0.402	0.394	0.396	0.461	0.453	0.406	0.420	0.418	0.503	0.520	0.396	0.503	0.466	0.442	0.463	0.400	0.446	0.420	0.398	0.411	0.469
CIFAR10	0.663	0.548	0.567	0.687	0.572	0.629	0.659	0.591	0.686	0.639	0.663	0.659	0.530	0.555	0.503	0.659	0.557	0.621	0.663	0.660	0.595	0.629	0.877	0.889	0.888
MNIST-C	0.757	0.500	0.500	0.702	0.689	0.733	0.786	0.591	0.699	0.739	0.751	0.741	0.581	0.552	0.594	0.752	0.670	0.705	0.751	0.788	0.703	0.746	0.726	0.739	0.741
MVTec-AD	0.754	0.500	0.500	0.745	0.732	0.747	0.763	0.644	0.742	0.618	0.735	0.724	0.596	0.603	0.544	0.730	0.683	0.637	0.732	0.761	0.655	0.730	0.849	0.867	0.860
SVHN	0.601	0.500	0.500	0.629	0.542	0.580	0.604	0.534	0.628	0.583	0.604	0.599	0.528	0.521	0.521	0.597	0.571	0.580	0.605	0.607	0.567	0.600	0.579	0.581	0.581
mnist	0.843	0.500	0.500	0.664	0.574	0.811	0.867	0.564	0.658	0.856	0.849	0.848	0.631	0.605	0.615	0.698	0.691	0.645	0.816	0.853	0.756	0.819	0.791	0.796	0.794
FashionMNIST	0.871	0.500	0.500	0.748	0.748	0.831	0.875	0.672	0.738	0.840	0.860	0.853	0.664	0.647	0.564	0.860	0.758	0.819	0.861	0.873	0.767	0.841	0.902	0.921	0.920
20news	0.564	0.533	0.544	0.610	0.537	0.550	0.567	0.539	0.610	0.583	0.559	0.545	0.518	0.515	0.496	0.553	0.547	0.513	0.547	0.570	0.527	0.579	0.725	0.725	0.725
agnews	0.619	0.551	0.552	0.715	0.554	0.584	0.647	0.568	0.714	0.665	0.601	0.566	0.508	0.494	0.500	0.592	0.591	0.497	0.571	0.652	0.545	0.627	0.862	0.868	0.864
amazon	0.579	0.571	0.541	0.572	0.563	0.558	0.603	0.526	0.571	0.597	0.565	0.550	0.501	0.464	0.500	0.560	0.528	0.495	0.551	0.603	0.535	0.556	0.544	0.546	0.548
imdb	0.496	0.512	0.471	0.499	0.499	0.489	0.494	0.466	0.500	0.504	0.484	0.478	0.487	0.526	0.500	0.486	0.521	0.493	0.478	0.495	0.486	0.484	0.524	0.523	0.527
yelp	0.635	0.605	0.578	0.661	0.600	0.602	0.670	0.581	0.661	0.655	0.621	0.592	0.498	0.524	0.504	0.590	0.545	0.527	0.594	0.671	0.514	0.602	0.558	0.555	0.557
Average	0.731	0.692	0.689	0.608	0.714	0.737	0.728	0.627	0.611	0.748	0.725	0.718	0.614	0.543	0.600	0.627	0.643	0.696	0.700	0.722	0.660	0.728	0.751	0.757	**0.766**

Table B.9: *Per-dataset AUPRC comparison on the 57* `ADBench` *datasets.* Methods marked with '*' are implemented and evaluated in our framework, while the remaining baseline results are taken from (Livernoche et al., 2023). Best average AUPRC for each method is **bolded**.

	CBLOF	COPOD	ECOD	FeatureBagging	HBOS	IForest	kNN	LODA	LOF	MCD	OCSVM	PCA	DAGMM	DeepSVDD	DROCC	GOAD	ICL	PlanarFlow	DDPM	DTE-NP	DTE-IG	DTE-C	ODIM*	ALTBI*	Ours* (w/ ALTBI)
aloi	0.037	0.031	0.033	0.104	0.034	0.034	0.048	0.033	0.097	0.032	0.039	0.037	0.033	0.034	0.030	0.033	0.046	0.032	0.036	0.056	0.040	0.033	0.037	0.035	0.057
annthyroid	0.169	0.174	0.272	0.206	0.228	0.312	0.224	0.098	0.163	0.503	0.188	0.196	0.109	0.192	0.186	0.131	0.123	0.654	0.297	0.228	0.380	0.670	0.154	0.156	0.181
backdoor	0.547	0.025	0.025	0.217	0.052	0.045	0.479	0.101	0.358	0.122	0.534	0.531	0.250	0.372	0.025	0.347	0.717	0.336	0.520	0.473	0.438	0.481	0.406	0.138	0.122
breastw	0.890	0.989	0.982	0.284	0.954	0.956	0.932	0.955	0.297	0.962	0.897	0.946	0.660	0.482	0.776	0.826	0.635	0.908	0.537	0.921	0.770	0.715	0.930	0.943	0.952
campaign	0.287	0.368	0.354	0.145	0.352	0.279	0.289	0.131	0.158	0.325	0.283	0.284	0.163	0.149	0.113	0.105	0.267	0.191	0.299	0.281	0.237	0.321	0.246	0.266	0.247
cardio	0.482	0.576	0.567	0.161	0.458	0.559	0.402	0.428	0.159	0.364	0.536	0.609	0.193	0.177	0.272	0.540	0.108	0.471	0.278	0.376	0.184	0.268	0.526	0.539	0.519
cardiotocography	0.335	0.403	0.502	0.276	0.361	0.436	0.324	0.463	0.272	0.311	0.408	0.462	0.271	0.252	0.258	0.403	0.188	0.348	0.338	0.312	0.250	0.276	0.390	0.424	0.409
celeba	0.069	0.093	0.095	0.024	0.090	0.063	0.061	0.047	0.018	0.092	0.103	0.112	0.044	0.031	0.047	0.021	0.045	0.066	0.093	0.052	0.058	0.077	0.064	0.098	0.090
census	0.088	0.062	0.062	0.061	0.073	0.073	0.088	0.065	0.069	0.153	0.085	0.087	0.062	0.075	0.058	0.072	0.095	0.074	0.086	0.090	0.083	0.081	0.096	0.090	0.089
cover	0.070	0.068	0.113	0.019	0.026	0.052	0.054	0.090	0.019	0.016	0.099	0.075	0.044	0.048	0.056	0.005	0.022	0.010	0.046	0.048	0.025	0.021	0.068	0.047	0.051
donors	0.148	0.209	0.265	0.120	0.135	0.124	0.182	0.255	0.109	0.141	0.139	0.166	0.086	0.112	0.123	0.040	0.119	0.241	0.143	0.188	0.164	0.140	0.072	0.071	0.079
fault	0.473	0.313	0.325	0.396	0.360	0.395	0.522	0.337	0.388	0.334	0.401	0.332	0.361	0.375	0.496	0.381	0.473	0.329	0.392	0.532	0.417	0.422	0.495	0.477	0.475
fraud	0.145	0.252	0.215	0.003	0.209	0.145	0.169	0.146	0.003	0.488	0.110	0.149	0.084	0.250	0.002	0.257	0.127	0.447	0.146	0.137	0.188	0.648	0.329	0.299	0.323
glass	0.144	0.111	0.183	0.151	0.161	0.144	0.167	0.090	0.144	0.113	0.130	0.112	0.111	0.090	0.159	0.076	0.122	0.113	0.073	0.206	0.135	0.168	0.104	0.083	0.094
hepatitis	0.304	0.389	0.295	0.225	0.328	0.243	0.252	0.275	0.214	0.363	0.277	0.339	0.253	0.170	0.221	0.291	0.231	0.317	0.165	0.238	0.215	0.257	0.242	0.454	0.459
http	0.464	0.280	0.145	0.047	0.302	0.886	0.010	0.004	0.050	0.865	0.356	0.500	0.368	0.093	0.004	0.441	0.091	0.363	0.642	0.024	0.295	0.440	0.285	0.296	0.309
internetads	0.297	0.505	0.505	0.182	0.523	0.486	0.296	0.242	0.232	0.344	0.291	0.276	0.207	0.252	0.197	0.288	0.237	0.262	0.295	0.290	0.275	0.302	0.336	0.365	0.405
ionosphere	0.881	0.663	0.633	0.821	0.353	0.779	0.911	0.741	0.807	0.947	0.829	0.721	0.473	0.392	0.728	0.781	0.472	0.824	0.633	0.920	0.610	0.880	0.818	0.801	0.839
landsat	0.212	0.176	0.164	0.246	0.231	0.194	0.258	0.183	0.250	0.253	0.175	0.163	0.230	0.362	0.272	0.198	0.329	0.187	0.200	0.255	0.203	0.223	0.192	0.202	0.219
letter	0.166	0.068	0.077	0.445	0.078	0.086	0.203	0.083	0.433	0.174	0.113	0.076	0.083	0.099	0.252	0.099	0.208	0.153	0.367	0.255	0.181	0.257	0.135	0.120	0.130
lymphography	0.915	0.907	0.894	0.090	0.919	0.972	0.894	0.491	0.135	0.767	0.885	0.935	0.454	0.254	0.463	0.897	0.264	0.417	0.731	0.805	0.388	0.381	0.563	0.782	0.705
magic.gamma	0.666	0.588	0.533	0.539	0.617	0.638	0.724	0.579	0.520	0.632	0.625	0.589	0.450	0.499	0.627	0.326	0.548	0.692	0.651	0.730	0.657	0.664	0.664	0.689	0.643
mammography	0.140	0.430	0.435	0.070	0.132	0.218	0.181	0.218	0.085	0.036	0.187	0.204	0.111	0.025	0.114	0.046	0.046	0.074	0.099	0.175	0.082	0.170	0.083	0.067	0.070
musk	1.000	0.369	0.475	0.140	0.999	0.945	0.708	0.842	0.118	0.992	1.000	1.000	0.500	0.107	0.196	1.000	0.128	0.391	0.984	0.434	0.137	0.553	0.537	1.000	1.000
optdigits	0.059	0.029	0.029	0.036	0.192	0.046	0.022	0.029	0.035	0.022	0.027	0.027	0.026	0.039	0.032	0.039	0.030	0.027	0.022	0.021	0.028	0.028	0.030	0.038	0.046
pageblocks	0.547	0.370	0.520	0.341	0.319	0.464	0.556	0.410	0.292	0.617	0.531	0.525	0.255	0.288	0.632	0.373	0.285	0.538	0.493	0.530	0.507	0.555	0.543	0.553	0.584
pendigits	0.192	0.177	0.270	0.048	0.247	0.260	0.100	0.186	0.040	0.069	0.226	0.219	0.056	0.022	0.027	0.075	0.045	0.060	0.056	0.089	0.044	0.044	0.173	0.253	0.335
pima	0.484	0.536	0.484	0.412	0.577	0.510	0.530	0.404	0.406	0.498	0.477	0.492	0.372	0.366	0.413	0.476	0.385	0.476	0.400	0.528	0.437	0.447	0.501	0.498	0.494
satellite	0.656	0.570	0.526	0.378	0.688	0.649	0.582	0.613	0.381	0.768	0.654	0.606	0.527	0.406	0.465	0.658	0.451	0.596	0.662	0.563	0.380	0.529	0.634	0.681	0.698
satimage-2	0.972	0.797	0.666	0.042	0.760	0.918	0.690	0.857	0.041	0.682	0.965	0.872	0.289	0.052	0.076	0.949	0.102	0.484	0.783	0.507	0.095	0.138	0.948	0.907	0.920
shuttle	0.184	0.962	0.905	0.081	0.965	0.976	0.193	0.168	0.109	0.841	0.907	0.913	0.438	0.149	0.072	0.136	0.135	0.346	0.779	0.187	0.247	0.626	0.944	0.959	0.972
skin	0.289	0.179	0.183	0.207	0.232	0.254	0.290	0.180	0.221	0.490	0.220	0.172	0.226	0.221	0.285	0.232	0.173	0.335	0.175	0.290	0.316	0.302	0.457	0.449	0.441
smtp	0.403	0.005	0.589	0.001	0.005	0.005	0.415	0.312	0.022	0.006	0.383	0.382	0.179	0.240	0.000	0.358	0.004	0.006	0.502	0.411	0.012	0.422	0.507	0.079	0.033
spambase	0.402	0.544	0.518	0.344	0.518	0.488	0.415	0.387	0.360	0.349	0.402	0.409	0.389	0.456	0.383	0.387	0.370	0.433	0.384	0.407	0.399	0.400	0.394	0.380	0.528
speech	0.019	0.019	0.020	0.022	0.023	0.021	0.019	0.016	0.022	0.019	0.019	0.018	0.022	0.018	0.020	0.019	0.020	0.018	0.020	0.019	0.019	0.020	0.018	0.017	0.017
stamps	0.211	0.398	0.324	0.143	0.332	0.347	0.317	0.280	0.153	0.257	0.318	0.364	0.198	0.099	0.241	0.285	0.117	0.284	0.143	0.273	0.235	0.226	0.338	0.309	0.312
thyroid	0.272	0.179	0.472	0.069	0.501	0.562	0.392	0.189	0.077	0.702	0.329	0.356	0.126	0.024	0.338	0.318	0.066	0.734	0.325	0.360	0.118	0.705	0.244	0.228	0.220
vertebral	0.123	0.085	0.110	0.124	0.091	0.097	0.095	0.089	0.130	0.101	0.107	0.099	0.134	0.107	0.118	0.124	0.115	0.111	0.150	0.098	0.133	0.119	0.097	0.097	0.095
vowels	0.166	0.034	0.083	0.314	0.078	0.162	0.443	0.127	0.326	0.085	0.196	0.069	0.041	0.037	0.178	0.154	0.219	0.295	0.311	0.504	0.166	0.417	0.242	0.308	0.297
waveform	0.122	0.057	0.040	0.078	0.048	0.056	0.133	0.040	0.071	0.040	0.052	0.044	0.032	0.061	0.150	0.042	0.063	0.150	0.050	0.109	0.037	0.043	0.180	0.129	0.113
wbc	0.691	0.883	0.882	0.037	0.728	0.948	0.743	0.898	0.077	0.839	0.813	0.913	0.327	0.069	0.358	0.736	0.211	0.431	0.758	0.722	0.348	0.194	0.859	0.862	0.869
wdbc	0.688	0.760	0.493	0.155	0.761	0.702	0.521	0.527	0.128	0.395	0.539	0.613	0.152	0.063	0.039	0.589	0.065	0.568	0.483	0.465	0.074	0.157	0.783	0.625	0.530
wilt	0.040	0.037	0.042	0.081	0.039	0.044	0.049	0.036	0.083	0.153	0.035	0.032	0.047	0.046	0.041	0.065	0.109	0.115	0.076	0.054	0.211	0.163	0.038	0.037	0.042
wine	0.170	0.364	0.195	0.061	0.412	0.207	0.081	0.250	0.064	0.737	0.135	0.264	0.120	0.116	0.126	0.229	0.087	0.086	0.075	0.074	0.064	0.103	0.259	0.502	0.416
wpbc	0.227	0.234	0.217	0.206	0.241	0.237	0.234	0.227	0.210	0.257	0.222	0.229	0.214	0.240	0.234	0.214	0.234	0.236	0.238	0.227	0.238	0.231	0.230	0.230	0.233
yeast	0.314	0.308	0.332	0.326	0.328	0.304	0.294	0.330	0.315	0.298	0.303	0.302	0.353	0.350	0.284	0.332	0.318	0.309	0.320	0.295	0.306	0.306	0.291	0.300	0.331
CIFAR10	0.103	0.065	0.067	0.115	0.075	0.089	0.102	0.086	0.115	0.084	0.102	0.101	0.062	0.073	0.060	0.102	0.070	0.085	0.102	0.104	0.078	0.092	0.378	0.526	0.525
MNIST-C	0.173	0.050	0.050	0.128	0.126	0.178	0.191	0.101	0.127	0.166	0.179	0.170	0.092	0.097	0.096	0.177	0.098	0.154	0.178	0.192	0.141	0.157	0.213	0.304	0.313
MVTec-AD	0.570	0.236	0.236	0.536	0.546	0.570	0.580	0.464	0.532	0.451	0.555	0.540	0.362	0.387	0.317	0.546	0.404	0.454	0.546	0.578	0.439	0.517	0.510	0.535	0.534
SVHN	0.079	0.050	0.050	0.084	0.064	0.073	0.079	0.064	0.083	0.068	0.078	0.078	0.059	0.063	0.060	0.078	0.068	0.074	0.078	0.080	0.069	0.077	0.065	0.065	0.065
mnist	0.386	0.092	0.092	0.241	0.109	0.290	0.409	0.170	0.233	0.308	0.385	0.381	0.215	0.253	0.237	0.297	0.232	0.259	0.374	0.400	0.276	0.368	0.334	0.396	0.391
FashionMNIST	0.329	0.050	0.050	0.194	0.269	0.320	0.346	0.180	0.188	0.245	0.329	0.319	0.138	0.181	0.106	0.328	0.158	0.297	0.325	0.339	0.213	0.267	0.473	0.695	0.697
20news	0.067	0.061	0.062	0.087	0.061	0.062	0.069	0.062	0.088	0.072	0.064	0.062	0.054	0.058	0.055	0.063	0.063	0.056	0.063	0.072	0.060	0.068	0.129	0.117	0.119
agnews	0.072	0.059	0.058	0.125	0.059	0.064	0.082	0.064	0.125	0.077	0.068	0.061	0.053	0.053	0.051	0.066	0.069	0.050	0.062	0.085	0.063	0.076	0.286	0.281	0.264
amazon	0.061	0.060	0.055	0.058	0.059	0.058	0.062	0.054	0.058	0.062	0.059	0.057	0.049	0.046	0.050	0.058	0.052	0.050	0.057	0.062	0.055	0.057	0.054	0.054	0.055
imdb	0.047	0.050	0.045	0.049	0.047	0.047	0.047	0.046	0.049	0.049	0.047	0.046	0.049	0.053	0.050	0.047	0.054	0.051	0.046	0.047	0.047	0.047	0.055	0.055	0.056
yelp	0.073	0.072	0.065	0.085	0.070	0.070	0.083	0.067	0.085	0.075	0.073	0.069	0.049	0.058	0.051	0.068	0.054	0.056	0.069	0.085	0.054	0.066	0.057	0.057	0.057
Average	0.318	0.288	0.296	0.179	0.308	0.336	0.308	0.260	0.181	0.337	0.324	0.328	0.198	0.170	0.199	0.285	0.185	0.283	0.301	0.295	0.216	0.288	0.334	0.350	**0.352**