RETHINKING STRUCTURAL ANOMALY DETECTION: FROM DECISION BOUNDARIES TO PROJECTION OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many anomaly detection methods either rely on estimating a probability density or learning an enclosing decision boundary, implicitly assuming that normal data occupies a region of non-zero volume in the ambient space. In contrast, structural anomaly detection deals with perceptual data such as images, where normal samples concentrate near a low-dimensional manifold and anomalies correspond to deviations outside its support. This view reveals a mismatch between the volumetric bias of existing methods and the geometric structure of the data, often resulting in degraded performance in practice. To address this mismatch, we formalize the task as learning a projection operator onto the manifold of normal samples and define a sample as anomalous if it is altered by this projection. This formulation naturally integrates the inductive bias of manifold-supported data and reframes anomaly detection in terms of a projection residual, thereby resolving issues arising from modeling degenerate distributions. Notably, it provides a unifying interpretation of reconstruction-based methods by explaining their success and failure in terms of projection quality. In particular, it explains the strong generalization ability of projection-aligned models to unseen anomalies as a consequence of their contraction behavior that suppresses off-manifold variation. Empirically, we demonstrate that such models substantially outperform a range of boundary-based methods and achieve state-of-the-art performance while producing segmentation masks with fine, accurate boundaries.

1 INTRODUCTION

Conceptually, anomaly detection (AD) aims to identify observations that deviate from a notion of normality. In many practical scenarios, such as industrial inspection or medical imaging, anomalies arise as structural deviations in input images, including surface defects, shape deformations, or other irregularities disrupting normal visual patterns.

The vast majority of existing AD methods (Schölkopf et al., 2001; Tax & Duin, 2004; Parzen, 1962; Kim et al., 2023; Ruff et al., 2018; Tack et al., 2020; Ruff et al., 2021) define normality either through a probability density or, more generally, a decision boundary enclosing normal samples. Crucially, this relies on an implicit assumption that normal data occupies a region of non-zero volume in the ambient space. However, this assumption becomes brittle for high-dimensional perceptual data such as images, where normal samples concentrate near a low-dimensional manifold. Since an embedded manifold has no interior with respect to the ambient space, any enclosing decision boundary necessarily includes unsupported regions. In practice, this mismatch can lead to unstable learning behavior, sensitivity to noise in the scoring function, and increased vulnerability to the curse of dimensionality. Fig. 1 illustrates, using the example of a One-Class Support Vector Machine (OC-SVM) Schölkopf et al. (2001), how the misalignment between the volumetric assumption of boundary-based methods and the low-dimensional nature of manifold-supported data manifests geometrically. When combined with isotropic radial kernels (e.g., a Gaussian kernel), similarity is propagated equally in both tangent and normal directions of the data manifold. Consequently, the kernel width induces an unavoidable trade-off: large values admit larger off-manifold regions, leading to false negatives, while small values require dense sampling to ensure manifold coverage and otherwise lead to false positives. While this example clearly illustrates the geometric mismatch,

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.

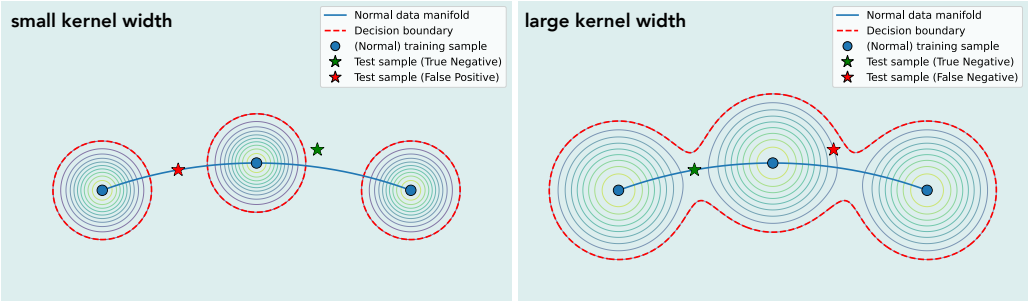

Figure 1: Illustration of the failure mode of an OC-SVM with an RBF kernel on manifold-supported data. Although normal samples concentrate near a low-dimensional manifold (blue curve), the isotropic similarity induced by the kernel in ambient space produces a volumetric inlier region. The effective radius of this region is controlled by the kernel width: decreasing it reduces coverage of the manifold and increases false positives (left), whereas increasing it expands the inlier region and increases false negatives (right). Note the change in classification of the star-shaped samples.

the underlying issue is not restricted to classical kernel methods. Although pretrained representations such as CLIP (Radford et al., 2021) or DINOv2 (Oquab et al., 2024) have been shown to substantially improve detection performance, they do not fully resolve the volumetric mismatch in the input space. The same limitation applies, to varying degrees, to patch-wise memory-bank methods (Roth et al., 2022) and deep one-class approaches (Ruff et al., 2018). See Proposition 1 in Appendix A.1.

Given the geometric misalignment between enclosing decision boundaries and manifold-supported data, a natural alternative is to measure the abnormality of a sample via its distance to the manifold. However, this idea immediately encounters fundamental challenges: the manifold itself is unknown and must be inferred from finitely many samples, and computing distances requires solving a non-trivial optimization problem to identify the closest point on the manifold. In high-dimensional settings, both aspects are computationally and statistically intractable. Instead, we propose a different approach. The key observation is that, for natural images, the global shape of the data manifold is induced by strong spatial correlations between pixels, which drastically reduce the effective degrees of freedom. Specifically, these dependencies give rise to a form of *index-induced regularity*, rendering the value of a pixel largely predictable from its local neighborhood. This enables implicit modeling of the global geometry from individual samples by learning a mapping that corrects deviations from the manifold. When applied to anomalous data, this mapping acts as an approximate projection operator, mapping inputs toward the manifold without requiring explicit geometric representation. In this way, global geometry becomes tied to local predictive structure shared across training samples, supporting the learning and generalization of a projection mapping. The learned projection, in turn, induces an implicit characterization of the data manifold as the zero-set of a reconstruction residual. This provides a global description of normality without requiring explicit geometric estimation or costly optimization to identify nearest points on the manifold.

To this end, we reinterpret and extend the family of correction-based methods (Zavrtanik et al., 2021a; Li et al., 2021; Zavrtanik et al., 2021b; Pirnay & Chai, 2022; Bauer et al., 2024) for structural anomaly detection (SAD) by viewing them as approximations to a learnable projection operator onto the manifold of normal data. Given an input, the model produces a corrected version projected toward the manifold, while abnormality is quantified through the resulting projection residual. At the pixel level, the discrepancy between the input and its projection enables localization of anomalous regions. Importantly, this perspective provides a unified framework for understanding and improving existing reconstruction-based approaches. First, it aligns the learning objective with the intrinsic geometry of the data, introducing an inductive bias consistent with manifold-supported distributions. Second, it offers a principled explanation for the limitations of decision-boundary methods and the success of reconstruction-based approaches in terms of projection quality. Third, by decoupling anomaly detection from probabilistic modeling, it can reduce the tendency to misclassify rare but normal samples, a widely recognized limitation of probabilistic approaches. Fourth, it provides a geometric interpretation of generalization through contractive behavior that collapses off-manifold variation. Collectively, these insights provide a roadmap for improved algorithm design. In particular, they

motivate geometry-aware regularization through Jacobian constraints, iterative refinement through fixed point iteration, and improved corruption strategies that encourage stable projection behavior and stronger generalization.

The remainder of the paper is organized as follows. Section 2 describes the training framework used to approximate the projection operator, together with the corresponding inference procedure. Section 3 develops the theoretical foundation of this framework in four steps. It first derives the fixed point condition and minimal correction support as the inductive biases required for SAD and identifies the conservative projection as an optimal solution. It then formulates an optimization problem whose Bayes-optimal solution satisfies these biases and explains how the practical training objective from Section 2 approximates this problem. Next, it shows how the contraction behavior toward the normal-data manifold induced by the corruption process enables progressive refinement of corrupted regions through fixed point iteration. Finally, it provides a geometric interpretation of model generalization as a consequence of collapsing off-manifold variation. Section 4 presents the experimental evaluation on two established industrial anomaly detection benchmarks and compares the proposed framework with state-of-the-art approaches. Section 5 concludes the paper.

2 METHODOLOGY

Formally, we approximate the nonlinear projection operator using an autoencoder

$$f_{\boldsymbol{\theta}} : [0,1]^{h\times w\times 3} \to [0,1]^{h\times w\times 3},$$

where $\boldsymbol{\theta}$ denotes the learnable parameters. Both input and output correspond to RGB images of resolution $h \times w$. In particular, we make no architectural assumptions. Specifically, the model is not required to contain a bottleneck. The term autoencoder is used in a general sense, referring simply to mappings with identical input and output dimensionality.

2.1 TRAINING

The model is trained in a self-supervised fashion using input-output pairs of normal samples and their corrupted versions. Let $\boldsymbol{x} \in [0,1]^{h\times w\times 3}$ denote a normal (anomaly-free) image. A corrupted version $\hat{\boldsymbol{x}}$ is obtained by altering randomly selected regions, specified by a real-valued mask $\mathbf{M} \in [0,1]^{h\times w\times 3}$. Its complement is defined as $\bar{\mathbf{M}} := \mathbf{1} - \mathbf{M}$, where $\mathbf{1}$ denotes the tensor of ones.

We adopt the loss formulation introduced in (Bauer et al., 2024) and derive its projection-based motivation in the next section:

$$\mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{x}, \mathbf{M}; \boldsymbol{\theta}) := \frac{1-\lambda}{\|\bar{\mathbf{M}}\|_1}\|\bar{\mathbf{M}} \odot (f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}) - \boldsymbol{x})\|_2^2 + \frac{\lambda}{\|\mathbf{M}\|_1}\|\mathbf{M} \odot (f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}) - \boldsymbol{x})\|_2^2, \tag{1}$$

where $\odot$ denotes elementwise tensor multiplication, $\|\cdot\|_p$ denotes the ℓ^p-norm, and $\lambda \in (0,1)$ balances the contribution of corrupted versus uncorrupted regions.

There are multiple ways of choosing a corruption pattern for generating the training data. An efficient and powerful augmentation technique is to use an auxiliary texture dataset such as DTD (Cimpoi et al., 2014), which provides background images $\mathcal{B}$ containing structural patterns absent from the normal data. Given a normal image $\boldsymbol{x} \in \mathcal{M}$, a corruption pattern $\boldsymbol{y} \in \mathcal{B}$, and a randomly selected smooth mask $\mathbf{M}$, a corrupted version is created according to $\hat{\boldsymbol{x}} = \mathbf{M} \odot \boldsymbol{y} + \bar{\mathbf{M}} \odot \boldsymbol{x}$.

2.2 TEST-TIME DETECTION

After training, anomaly localization is performed by comparing an input $\hat{\boldsymbol{x}}$ with its reconstruction $f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}})$. To this end, we define a pixel-wise discrepancy function

$$\Delta : [0,1]^{h\times w\times 3} \times [0,1]^{h\times w\times 3} \to [0,1]^{h\times w}$$

Different choices of Δ have been proposed, including Mean Squared Error (MSE), Structural Similarity Index Measure (SSIM) (Bergmann et al., 2019), and Gradient Magnitude Similarity (GMS) (Xue et al., 2014), each capturing different aspects of reconstruction quality. A binary segmentation mask can be obtained by thresholding the pixel-wise anomaly map. Regardless of the choice of Δ, spatial smoothing before thresholding improves stability.

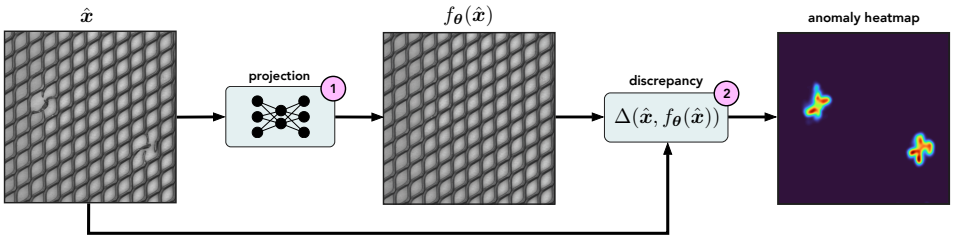

Figure 2: Illustration of our anomaly detection process. Given an input $\hat{\boldsymbol{x}}$, a trained model first produces an output $f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}})$, which preserves normal regions and replaces irregularities with locally consistent patterns. Second, we compute a pixel-wise discrepancy map $\Delta(\hat{\boldsymbol{x}}, f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}))$ between the input and the output to localize anomalous regions.

For image-level detection, a global anomaly score is obtained by aggregating the values of the anomaly map, typically via summation. Alternative strategies, such as taking the maximum or restricting the aggregation to the largest responses, can reduce sensitivity to the spatial extent of anomalies. The overall inference pipeline is illustrated in Fig. 2.

3 Formal Analysis

This section develops the theoretical foundation of the projection-based framework for SAD. We first derive the fixed point condition and minimal correction support as the inductive biases required for detection and localization, and identify the conservative projection as an optimal solution. We then formulate an idealized optimization problem whose Bayes-optimal solution satisfies these biases and explain how the practical training objective from Section 2 approximates this problem. We further discuss iterative refinement through fixed point iteration as a means of improving the approximation to the conservative projection at test time. Finally, we provide a geometric interpretation of model generalization as a consequence of collapsing off-manifold variation.

3.1 Geometric Inductive Biases

We now formalize the inductive biases required for detecting and localizing structural deviations.

Inductive Bias 1 (Fixed Point Condition)**.** *Let $\mathcal{M} \subset \mathbb{R}^d$ denote the set of normal data points. A reconstruction mapping $f : \mathbb{R}^d \to \mathbb{R}^d$ must satisfy the fixed point condition with respect to $\mathcal{M}$:*

$$f(\boldsymbol{x}) = \boldsymbol{x} \quad \Longleftrightarrow \quad \boldsymbol{x} \in \mathcal{M}. \tag{2}$$

This property ensures identity reconstruction exclusively on the data manifold while excluding trivial identity mappings in the ambient space, a behavior often violated by regularized autoencoders. It therefore constitutes a minimal requirement for anomaly detection: any mapping satisfying this condition is sufficient for the detection task, as off-manifold inputs necessarily violate the fixed point condition. In particular, any projection operator f with $\mathrm{Im}(f) = \mathcal{M}$ satisfies this property. While vanilla and regularized autoencoders are often interpreted as learning the underlying data manifold (Zhang et al., 2017; Connor et al., 2021), they primarily optimize reconstruction of normal samples, leaving their behavior in the ambient space largely unconstrained. Common regularizers, including bottleneck compression and Jacobian penalties (Alain & Bengio, 2014), are likewise insufficient to confine near-identity reconstruction to the data manifold, as frequently observed in visual settings (Bhattad et al., 2018; Aguila et al., 2025). Consequently, these models often provide only weak separation between normal and anomalous samples.

In contrast, assuming the model accurately realizes the behavior prescribed by the inductive bias in (2), the residual mapping $\Phi(\boldsymbol{x}) := f(\boldsymbol{x}) - \boldsymbol{x}$ fully captures the normal data manifold as

$$\mathcal{M} = \Phi^{-1}(\mathbf{0}). \tag{3}$$

If Φ is smooth and its Jacobian $D\Phi(\boldsymbol{x})$ has constant rank $d - k$ in a neighborhood of $\mathcal{M}$, then by the constant-rank theorem, the level set $\Phi^{-1}(\mathbf{0})$ locally defines a k-dimensional embedded submanifold of $\mathbb{R}^d$. In this sense, the reconstruction mapping f admits a natural interpretation as a projection

operator that implicitly captures the geometry of the normal manifold as its fixed point set, thereby separating it from the ambient space and yielding a principled criterion for distinguishing between normal and anomalous samples. Notably, this perspective directly aligns with the core objective of anomaly detection, namely learning a notion of normality. In contrast to common approaches, the proposed formulation explicitly defines normality via Eq. (3), while anomalous samples are implicitly characterized by their deviation from it.

Beyond separating normal and anomalous samples, domains such as images (or time series) additionally require localization of anomalies at the pixel level. For this purpose, we leverage the residual mapping Φ, which provides a spatial discrepancy map enabling localization of anomalous regions, as introduced in Section 2. Given an anomalous sample identified with a vector $\hat{\boldsymbol{x}} \in \mathbb{R}^d$, we define the anomalous region as a minimal set of pixels $S \subset \{1, ..., d\}$ that must be corrected to produce a normal sample $\boldsymbol{x} \in \mathcal{M}$. That is, S is uniquely defined and $\hat{\boldsymbol{x}}_{\bar{S}} = \boldsymbol{x}_{\bar{S}}$, where $\bar{S} := \{1, ..., d\} \setminus S$ denotes the complement of S, and $\hat{\boldsymbol{x}}_S$, $\boldsymbol{x}_S$ denote vector slices restricted to indices in S. It is worth noting that the uniqueness assumption does not generally hold, as discussed in (Bauer et al., 2024). However, the resulting ambiguity, which arises from the finite resolution of digitized images, is typically limited in extent and can often be neglected in practice.

Having established the formal definition of a *structural anomaly* as an off-manifold sample $\hat{\boldsymbol{x}} \notin \mathcal{M}$, with uniquely localized anomalous part $\hat{\boldsymbol{x}}_S$, we now can derive an optimal projection operator for the task of SAD, given as

$$\Pi_{\text{con}}(\boldsymbol{x}) := \arg\min_{\boldsymbol{y} \in \mathcal{M}} \|\boldsymbol{x} - \boldsymbol{y}\|_0, \tag{4}$$

which we refer to as a conservative projection. Here, $\|\cdot\|_0$ denotes the so-called ℓ_0 pseudo-norm, defined as the number of nonzero components of a vector $\|\boldsymbol{x}\|_0 = \#\{i : x_i \neq 0\}$. In this context, minimizing $\|\boldsymbol{x}-\boldsymbol{y}\|_0$ enforces a conservative correction by minimizing the support of the modification, i.e., the number of altered components required to project $\boldsymbol{x}$ onto $\mathcal{M}$.

Unlike orthogonal projection, which minimizes geometric distance, the conservative projection minimizes correction support. Since the objective $\|\boldsymbol{x} - \boldsymbol{y}\|_0$ depends on support cardinality, the resulting operator is inherently non-smooth. Under uniqueness assumptions on the correction support, it may nevertheless exhibit piecewise-continuous or piecewise-differentiable behavior. Specifically, within regions where the optimal support remains unchanged, the mapping reduces to a constrained correction problem on the manifold, while non-smoothness arises at transitions between distinct supports. Consequently, such behavior may be locally approximated by a differentiable autoencoder that learns a smooth relaxation of conservative correction.

Under uniqueness of the correction support, the conservative projection Π_{con} yields an optimal solution for SAD. First, it naturally satisfies Inductive Bias 1. Anomalous regions are then identified from the individual components of the discrepancy map Φ, such that a pixel $i \in \{1, \ldots, d\}$ is anomalous if and only if $|\Phi_i(\hat{\boldsymbol{x}})| > 0$. In practice, however, deviations from the idealized assumptions, together with reconstruction inaccuracies and numerical noise, motivate the use of thresholds $\tau, \nu > 0$: specifically, $\|\Phi(\hat{\boldsymbol{x}})\|_2 > \tau$ for anomaly detection and $|\Phi_i(\hat{\boldsymbol{x}})| > \nu$ for pixel-wise localization.

The conservative projection further reveals an additional inductive bias for projection operators required for accurate localization: minimal correction support. While the fixed point condition ensures that anomalous samples are altered by the reconstruction mapping f, thereby enabling detection, accurate localization additionally requires full preservation of normal regions.

Inductive Bias 2 (Minimal Correction Support)**.** *A reconstruction mapping $f : \mathbb{R}^d \to \mathbb{R}^d$ must preserve all normal components while modifying only the anomalous regions. Formally, for any input $\hat{\boldsymbol{x}}$ with anomalous support $S \subseteq \{1, \ldots, d\}$,*

$$f_i(\hat{\boldsymbol{x}}) \neq \hat{x}_i \quad \iff \quad i \in S. \tag{5}$$

It is worth noting that the requirement in (5) does not strictly imply $f(\hat{\boldsymbol{x}}) \in \mathcal{M}$. While preservation of normal regions is necessary, alteration of the corrupted region is both necessary and sufficient for accurate localization. In particular, Inductive Bias 2 implies Inductive Bias 1. In our framework, these inductive biases are approximated by minimizing reconstruction loss on partially occluded images with varying opacity. Specifically, the objective in (1) directly promotes preservation of normal regions while replacing corrupted regions with manifold-consistent patterns.

3.2 Approximation of the Conservative Projection

The following analysis clarifies how our training framework encodes the inductive biases discussed in the previous section. To make this connection precise, the theorem below characterizes the population-optimal mapping induced by the corresponding optimization problem.

For this purpose, we define the set-valued mappings $N, A\colon \mathbb{R}^d \to 2^{\{1,\dots,d\}}$, which return the indices of the normal and anomalous regions of an input, respectively, with $N(\hat{\boldsymbol{x}}) = \{1,\dots,d\} \setminus A(\hat{\boldsymbol{x}})$. As previously, we assume that this decomposition is uniquely defined. We further impose the standard measurability, integrability, support, and regularity conditions required for the conditional distributions and expectations below to be well defined and for the Bayes-optimal mapping to be representable by the considered hypothesis class.

Theorem 1. *Let $p(\boldsymbol{x})$ denote the distribution of normal data and let $q(\hat{\boldsymbol{x}} \mid \boldsymbol{x})$ denote the conditional distribution of corrupted samples induced by a corruption process for which $N(\hat{\boldsymbol{x}})$ and $A(\hat{\boldsymbol{x}})$ are well defined. Consider the following optimization problem:*

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^k}{\text{minimize}}\ \mathbb{E}_{p(\boldsymbol{x})}\left[\mathbb{E}_{q(\hat{\boldsymbol{x}}|\boldsymbol{x})}\left[\|f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}) - \boldsymbol{x}\|^2\right]\right]. \tag{6}$$

Provided sufficient capacity of $\{f_{\boldsymbol{\theta}}\}$, every minimizer $f^* := f_{\boldsymbol{\theta}^*}$ satisfies, for q-a.e. $\hat{\boldsymbol{x}}$,

$$f^*_{N(\hat{\boldsymbol{x}})}(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{x}}_{N(\hat{\boldsymbol{x}})} \qquad \textit{and} \qquad f^*_{A(\hat{\boldsymbol{x}})}(\hat{\boldsymbol{x}}) = \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})}\left[\boldsymbol{x}_{A(\hat{\boldsymbol{x}})}\right], \tag{7}$$

where the posterior distribution is given by $p(\boldsymbol{x} \mid \hat{\boldsymbol{x}}) \propto p(\boldsymbol{x})\, q(\hat{\boldsymbol{x}} \mid \boldsymbol{x})$.

In particular, the Bayes-optimal solution in (7) satisfies the inductive bias of minimal correction support and, consequently, the fixed point condition required for accurate detection and localization. The practical objective in (1) provides an empirical approximation of the population objective in (6), while normalizing the reconstruction errors on normal and corrupted regions separately and weighting their relative contributions by $\lambda \in (0, 1)$. In this sense, the resulting model approximates the conservative projection defined in (4). A formal proof of Theorem 1 is provided in Appendix A.2.

Theorem 1 highlights an important side effect of reconstruction-based methods. When correction of an anomalous region is ambiguous, the reconstruction loss drives the model toward the posterior mean over feasible restorations under the training distribution. Provided sufficient training data and model capacity, the approximation quality to the conservative projection is therefore primarily limited by reconstruction ambiguity within corrupted regions. In practice, we observed that this averaging effect can vary substantially depending on the structure and sparsity of the training data. Importantly, perfect reconstruction is not strictly necessary for accurate SAD. Any conservative alteration of the anomalous pattern is sufficient, as discussed in Section 3.1.

Nevertheless, higher reconstruction fidelity in corrupted regions may be beneficial in practice, as it can sharpen the localization signal and improve visual interpretability. An important consequence of our corruption process is that it promotes contractive behavior toward the data manifold, allowing the reconstruction to be refined progressively through repeated model application. Specifically, we iteratively apply the autoencoder to its previous output, eventually converging to a fixed point on the manifold. We summarize this result in the following theorem and provide a formal proof together with the underlying assumptions in Appendix A.3.

Theorem 2 (Fixed Point Iteration)**.** *Let $\mathcal{M} \subset \mathbb{R}^d$ be a closed manifold and $f : \mathbb{R}^d \to \mathbb{R}^d$ a nonlinear operator satisfying the conservativity, slice-contraction, and vanishing-correction assumptions stated in Appendix A.3. Then the following statements hold:*

(i) For all $\boldsymbol{x} \in \mathbb{R}^d$,

$$f(\boldsymbol{x}) = \boldsymbol{x} \quad \Longleftrightarrow \quad \boldsymbol{x} \in \mathcal{M}. \tag{8}$$

(ii) For all $\hat{\boldsymbol{x}} \in \mathbb{R}^d \setminus \mathcal{M}$ with anomaly index set S, there exists $\boldsymbol{x}^* \in \mathcal{M}$ such that

$$\lim_{n\to\infty} f^{(n)}(\hat{\boldsymbol{x}}) = \boldsymbol{x}^* \quad \textit{and} \quad \boldsymbol{x}^*_{\bar{S}} = \hat{\boldsymbol{x}}_{\bar{S}}. \tag{9}$$

As explained in detail in Appendix A.3, under the proposed corruption model $\hat{\boldsymbol{x}} \sim \mathcal{C}(\boldsymbol{x})$ based on partial occlusions with varying opacity levels and the stated modeling assumptions, the Bayes-optimal reconstruction progressively reduces the ambiguity among compatible normal sources $\boldsymbol{x} \in \mathcal{M}$

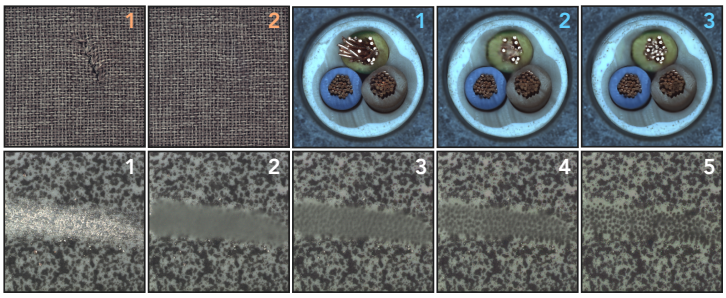

Figure 3: Illustration of improved reconstruction fidelity and varying convergence speeds to a fixed point under iterative application of our model, separately trained on different categories of the MVTec AD dataset. Each example shows (from left to right) the input together with reconstruction results obtained through repeated model application, converging toward a fixed point on the manifold.

satisfying $\boldsymbol{x}_{N(\hat{\boldsymbol{x}})} = \hat{\boldsymbol{x}}_{N(\hat{\boldsymbol{x}})}$. This induces the slice-contraction and vanishing-correction properties required by Theorem 2, thereby providing a theoretical justification for our training framework as a means of approximating a conservative projection onto the normal manifold.

Fig. 3 illustrates several examples of convergence under iterative model application, showing progressively improved reconstruction fidelity for defective samples from the MVTec AD dataset.

3.3 Generalization through Contraction toward the Data Manifold

The strong generalization of correction-based autoencoders to unseen anomalies, relative to supervised segmentation as observed in (Zavrtanik et al., 2021b), can be explained by the specific contraction behavior of the learned projection operator toward the normal-data manifold. Theorem 1 shows that, on corrupted regions, the Bayes-optimal output is an average over normal samples compatible with the corrupted input under the training corruption process. Consequently, accurate reconstruction does not require detailed modeling of the corruption patterns themselves. Moreover, when anomalous patterns exhibit substantially greater variation than normal structure, minimizing reconstruction error favors learning compatibility with normal data rather than modeling anomalous variation. This substantially reduces the dependence of the reconstruction on the particular appearance of a corruption pattern. Provided sufficient variation during training, Theorem 2 extends this mechanism across the input space as contraction toward the normal-data manifold. Repeated application moves corrupted inputs toward the manifold while preserving normal regions, converging to a fixed point. In particular, these results explain why artificial and real anomalies, despite potentially different appearances, are effectively handled by the same correction mechanism.

Specifically, we argue that the model realizes this contraction through a filtering strategy. Rather than learning separate responses to individual corruption patterns, the network suppresses irregular input structure, infers missing structure from normal context, and integrates the reconstruction into the remainder of the image. Inspection of the learned representations supports this interpretation. As illustrated in Fig. 4, selected feature maps of a trained U-Net reflect the successive suppression, reconstruction, and merging of anomalous regions. Our corruption process promotes this behavior by exposing the model to diverse off-manifold deviations with predominantly normal corruption directions and at arbitrarily small distances from the manifold.

4 Experiments

In this section, we evaluate the training framework presented in Section 2 and compare it with state-of-the-art approaches. Based on our geometric motivation, we refer to the resulting model as a projecting autoencoder (PAE).

Datasets. We evaluate the performance on two established benchmarks for industrial anomaly detection: MVTec AD (Bergmann et al., 2021) and VisA (Zou et al., 2022). Together, these datasets cover a broad range of industrial inspection scenarios, comprising 16,175 images across 27 categories.

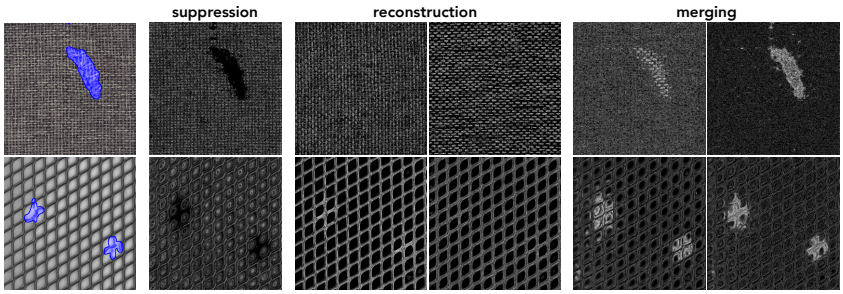

Figure 4: Activation patterns in selected feature maps from models (U-Net) trained on the carpet (top) and grid (bottom) categories of the MVTec AD dataset. The first column shows the anomalous inputs, with anomalous regions marked in blue. The presented activations support the interpretation that, at different stages of the forward pass, the model suppresses anomalous patterns, reconstructs missing content from the normal context, and merges the reconstructed part with the normal regions.

Training setup. We apply category-specific augmentations such as random rotations and flips, while reserving 5% of the training data for validation. A separate model is trained per category using a modified U-Net (Ronneberger et al., 2015) with dilated convolutions and optimized with Adam (Kingma & Ba, 2015) at a learning rate of 10^{-4}. We set $\lambda = 0.5$ in (1) and evaluate using the SSIM metric. Texture categories are resized to 512×512, whereas object categories are downscaled to 256×256. Accordingly, object categories use a larger dilation kernel (5 instead of 3). Several VisA categories contain substantial background clutter that is not annotated as anomalous. Where necessary, we apply foreground segmentation as a preprocessing step using a U-Net trained independently of the anomaly detection task. The resulting masks suppress irrelevant background regions prior to training, mitigating annotation noise that disproportionately affects projection-based methods.

Results and Interpretation. We report the Area Under the Receiver Operating Characteristic curve (AUROC) at the image level (I-AUROC) for anomaly detection and at the pixel level (P-AUROC) for anomaly segmentation in Table 1. In addition, we report pixel-level Average Precision (P-AP) and the area under the Per-Region Overlap curve (P-PRO). The best and second-best results are highlighted in red and blue, respectively. We compare our approach with a range of established baselines and recent state-of-the-art methods and summarize the main observations below.

Table 1 supports three observations consistent with our geometric interpretation. First, boundary-based approaches (Reiss et al., 2021) built on pretrained ResNet backbones, including OC-SVM (Schölkopf et al., 2001) and DeepSVDD (DSVDD) (Ruff et al., 2018), as well as OC-SVM on frozen CLIP embeddings (OC_{CLIP}), are substantially outperformed by PAE. Second, among reconstruction-based methods, AE (Bergmann et al., 2019), RIAD (Zavrtanik et al., 2021a), and DRAEM (Zavrtanik et al., 2021b) approximate projection behavior with increasing fidelity in this order, which correlates with the corresponding increase in their performance scores. However, these methods either rely on simplistic perturbations or deviate from a reconstruction objective aligned with the projection goal, suggesting a weaker approximation of the desired inductive biases. Third, PAE remains competitive with engineered hybrid approaches, including CutPaste (Li et al., 2021), SimpleNet (Liu et al., 2023), PatchCore (Roth et al., 2022), RealNet (Zhang et al., 2024), PBAS (Chen et al., 2025), PMSR (Li et al., 2026), and GLASS (Chen et al., 2024). While PAE performs on par with the strongest methods in terms of I/P-AUROC, it achieves higher P-AP scores, consistent with the visual sharpness of the resulting segmentation masks. The P-PRO metric, in contrast, places greater emphasis on recall and is comparatively tolerant of false-positive predictions, which can explain the higher PRO values of competing methods despite their comparatively lower AUROC and AP scores. On VisA, P-PRO decreases more noticeably, which we attribute to systematically incomplete annotation masks for displaced object parts: only the new location is annotated, while the anomalous absence at the original location is ignored. Positive predictions in these regions are therefore counted as false positives, increasing the false-positive rate and further reducing the PRO score. Fig. 7 in the appendix illustrates the segmentation quality achieved by PAE and includes an example of the aforementioned annotation error, while Fig. 8 highlights the qualitative difference of PAE's segmentation masks compared with GLASS. Additional sensitivity analyses and ablations are provided in the appendix A.4.

Table 1: Experimental results for image-level anomaly detection and pixel-level segmentation.

Dataset	Metric	AE	OC-SVM	DSVDD	OC_{CLIP}	RIAD	CutPaste	DRAEM	SimpleNet	PatchCore	RealNet	PBAS	PMSR	GLASS	**PAE**
MVTec AD	I-AUROC	63.0	70.8	77.9	87.8	91.7	95.2	98.0	99.6	99.2	99.7	99.8	99.9	99.9	99.9
	I-AP	61.1	73.1	79.1	94.7	92.3	95.7	98.5	99.8	99.8	99.5	99.8	99.8	99.9	99.9
	P-AUROC	87.0	–	–	–	94.2	96.0	97.3	98.1	98.4	99.0	98.6	98.8	99.3	99.4
	P-AP	23.0	–	–	–	53.2	56.0	68.4	58.1	57.1	69.4	74.1	73.9	74.2	78.6
	P-PRO	69.4	–	–	–	89.9	90.1	91.3	91.1	92.4	93.5	97.3	96.6	96.7	96.8
VisA	I-AUROC	54.9	59.1	62.2	77.1	81.1	85.1	88.7	97.1	94.7	97.8	97.7	98.5	98.2	98.5
	I-AP	56.1	61.4	65.0	79.8	85.1	87.0	90.5	97.9	92.7	97.1	98.1	98.5	96.9	98.7
	P-AUROC	77.4	–	–	–	88.7	90.6	93.5	98.2	98.5	98.8	98.6	98.0	98.6	98.9
	P-AP	17.4	–	–	–	24.8	24.2	26.6	37.3	38.5	38.1	47.6	43.1	45.6	54.0
	P-PRO	61.1	–	–	–	69.0	70.3	72.4	90.7	91.8	92.8	93.3	92.3	90.8	84.8
Mean	I-AUROC	59.0	65.0	70.1	82.5	86.4	90.2	93.4	98.4	97.0	98.8	98.8	**99.2**	99.1	**99.2**
	I-AP	58.6	67.3	72.1	87.3	88.7	91.4	94.5	98.9	96.3	98.3	98.9	99.2	98.4	**99.3**
	P-AUROC	82.2	–	–	–	91.5	93.3	95.4	98.2	98.5	98.9	98.7	98.4	98.9	**99.2**
	P-AP	20.2	–	–	–	39.0	40.1	47.5	47.7	47.8	53.8	60.8	58.5	59.9	**66.3**
	P-PRO	65.3	–	–	–	79.5	80.2	81.9	90.9	92.1	93.2	**95.3**	94.5	93.8	90.8

5 Conclusion

We developed a formal framework for SAD by casting the task as learning a projection onto the manifold of normal data, with several important consequences. First, unlike common approaches, this yields a precise definition of normality as the zero set of the reconstruction residual. Second, we derived the fixed point condition and minimal correction support as two inductive biases required for accurate detection and localization, and formulated a learning problem whose Bayes-optimal solution satisfies these properties. Third, we showed how fixed point iteration can refine the correction of corrupted regions at test time, thereby improving approximation to the conservative projection operator. More broadly, our framework places correction- and reconstruction-based approaches under the common objective of learning a projection onto normal data, allowing their empirical successes and failures to be explained in terms of alignment with or violation of these two inductive biases.

The proposed framework also provides an explanation for the strong generalization of correction-based autoencoders to unseen anomalies as a consequence of the specific contractive behavior of the learned projection. Under the conditions of our contraction result, repeated model application progressively reduces the distance to the normal-data manifold while preserving normal regions. Locally, this corresponds to suppressing variation along off-manifold directions, indicating that the correction mechanism is driven more by the geometry of the normal-data manifold and less by a specific corruption pattern. We argued that the model realizes this contraction through a filtering strategy that suppresses irregular structure, reconstructs compatible normal content, and integrates it with the remainder of the image. This shared correction mechanism, in particular, explains how models trained on artificial corruptions can effectively transfer to previously unseen real anomalies.

It also sharpens the boundary between structural and logical anomalies. Its effectiveness in SAD relies on exploiting the local predictive structure of images, consistent with our definition of structural anomalies as off-manifold deviations. Logical anomalies, in contrast, may preserve local appearance statistics while violating semantic or relational structure and can therefore remain on the data manifold. Such anomalies are not reliably detectable from reconstruction residuals alone, making explicit a limitation that is shared by the majority of existing anomaly detection approaches.

Empirical evaluation shows that PAE performs on par with state-of-the-art methods in I/P-AUROC while achieving higher P-AP and producing segmentation masks with finer boundaries. In contrast, competing methods often achieve similarly high AUROC while producing more diffuse segmentations with reduced boundary sharpness, as reflected by the corresponding variation in P-AP scores and supported by the direct visual comparison. This behavior is consistent with our theoretical characterization of the optimal projection operator, in particular the role of minimal correction support in promoting accurate and spatially precise corrections.

AI use statement

In this work, we used generative AI tools solely for language editing and improving the clarity, grammar, and presentation of author-written text. Generative AI tools were not used to generate research ideas, formulate technical methods, derive proofs or theoretical results, conduct experiments, analyze results, or generate figures or images. All scientific content, technical claims, mathematical

arguments, and experimental design were developed by the authors. All AI-assisted edits were reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

The theoretical results are stated together with their assumptions in the main paper, with complete derivations and proofs provided in the appendix. The empirical evaluation is described in sufficient detail to reproduce the reported comparisons, including datasets, preprocessing, training settings, baselines, and evaluation protocols. We will additionally release the training and evaluation code.

References

Ana Lawry Aguila, Peirong Liu, Oula Puonti, and Juan Eugenio Iglesias. Conditional diffusion models for guided anomaly detection in brain images using fluid-driven anomaly randomization. *CoRR*, abs/2506.10233, 2025.

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *J. Mach. Learn. Res.*, 15(110):3563–3593, 2014.

Alexander Bauer, Shinichi Nakajima, and Klaus-Robert Müller. Efficient exact inference with loss augmented objective in structured learning. *IEEE Trans. Neural Netw. Learn. Syst.*, 28(11): 2566–2579, 2017.

Alexander Bauer, Shinichi Nakajima, Nico Görnitz, and Klaus-Robert Müller. Partial optimality of dual decomposition for MAP inference in pairwise MRFs. In *Proc. Int. Conf. Artif. Intell. Stat.*, volume 89 of *Proc. Mach. Learn. Res.*, pp. 1696–1703, 2019a.

Alexander Bauer, Shinichi Nakajima, Nico Görnitz, and Klaus-Robert Müller. Optimizing for measure of performance in max-margin parsing. *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–5, 2019b.

Alexander Bauer, Shinichi Nakajima, and Klaus-Robert Müller. Polynomial-time constrained message passing for exact MAP inference on discrete models with global dependencies. *Mathematics*, 11 (12):2628, 2023.

Alexander Bauer, Shinichi Nakajima, and Klaus-Robert Müller. Self-supervised autoencoders for visual anomaly detection. *Mathematics*, 12(24):3988, 2024.

Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *Proc. Int. Conf. Comput. Vis. Theory Appl.*, pp. 372–380, 2019.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *Int. J. Comput. Vis.*, 129(4):1038–1059, 2021.

Anand Bhattad, Jason Rock, and David A. Forsyth. Detecting anomalous faces with "no peeking" autoencoders. *CoRR*, abs/1802.05798, 2018.

Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In *Proc. Eur. Conf. Comput. Vis.*, volume 15125 of *Lect. Notes Comput. Sci.*, pp. 37–54, 2024.

Qiyu Chen, Huiyuan Luo, Han Gao, Chengkan Lv, and Zhengtao Zhang. Progressive boundary guided anomaly synthesis for industrial anomaly detection. *IEEE Trans. Circuits Syst. Video Technol.*, 35(2):1193–1208, 2025.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014.

Marissa Connor, Gregory Canal, and Christopher Rozell. Variational autoencoder with learned latent structure. In *Proc. Int. Conf. Artif. Intell. Stat.*, volume 130 of *Proc. Mach. Learn. Res.*, pp. 2359–2367, 2021.

Minkyung Kim, Junsik Kim, Jongmin Yu, and Jun Kyun Choi. Active anomaly detection based on deep one-class classification. *Pattern Recognit. Lett.*, 167:18–24, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. Learn. Represent.*, 2015.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 9664–9674, 2021.

Lanxiao Li, Chunjuan Yan, Da Song, Binghui Wang, and Chuanxu Wang. A prototype correction multi-scale feature reconstruction network for industrial anomaly detection. *Pattern Recognit.*, 177:113331, 2026.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 20402–20411, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 33(3): 1065–1076, 1962.

Jonathan Pirnay and Keng Chai. Inpainting transformer for anomaly detection. In *Proc. Int. Conf. Image Anal. Process.*, volume 13232 of *Lect. Notes Comput. Sci.*, pp. 394–406, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. PANDA: Adapting pretrained features for anomaly detection and segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2806–2814, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. Med. Image Comput. Comput.-Assist. Interv.*, volume 9351 of *Lect. Notes Comput. Sci.*, pp. 234–241, 2015.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter V. Gehler. Towards total recall in industrial anomaly detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 14298–14308, 2022.

Lukas Ruff, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert A. Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In Jennifer G. Dy and Andreas Krause (eds.), *Proc. 35th Int. Conf. Mach. Learn.*, volume 80 of *Proc. Mach. Learn. Res.*, pp. 4390–4399, 2018.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proc. IEEE*, 109(5):756–795, 2021.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13 (7):1443–1471, 2001.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: Novelty detection via contrastive learning on distributionally shifted instances. In *Adv. Neural Inf. Process. Syst.*, 2020.

David M. J. Tax and Robert P. W. Duin. Support vector data description. *Mach. Learn.*, 54(1):45–66, 2004.

Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Trans. Image Process.*, 23(2):684–695, 2014.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognit.*, 112:107706, 2021a.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Dræm – A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 8310–8319, 2021b.

Dejiao Zhang, Yifan Sun, Brian Eriksson, and Laura Balzano. Deep unsupervised clustering using mixture of autoencoders. *CoRR*, abs/1712.07788, 2017.

Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 16699–16708, 2024.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *Proc. Eur. Conf. Comput. Vis.*, volume 13690 of *Lect. Notes Comput. Sci.*, pp. 392–408, 2022.

A Appendix

This supplementary material provides additional theoretical analysis and proofs supporting the main paper. Section A.1 provides a formal explanation of the volumetric bias induced by decision-boundary approaches. Section A.2 provides the formal proof of Theorem 1. Section A.3 discusses theoretical guarantees for convergence to a fixed point on the manifold under iterative model application, completing Theorem 2 with its accompanying proof, and provides further details on the role of the corruption process and the limitations with respect to logical anomalies.

A.1 On the Volumetric Bias of Enclosing Decision Boundaries

In the main paper, we used an OC-SVM with an RBF kernel in Fig. 1 to illustrate how the volumetric assumption of boundary-based methods conflicts with the low-dimensional structure of manifold-supported data. The Gaussian kernel provides a particularly simple geometric interpretation: reducing off-manifold acceptance requires decreasing the kernel width, which in turn demands increasingly dense sampling of the manifold. Replacing the kernel or using pretrained representations such as CLIP changes the shape of the induced decision boundary, but does not eliminate this fundamental mismatch. More generally, any volumetric decision boundary defined after a continuous feature mapping induces a volumetric decision boundary in the input space and therefore necessarily accepts unsupported off-manifold regions, as formalized below.

Proposition 1. *Let $\Phi\colon \mathbb{R}^n \to \mathbb{R}^m$ be a continuous feature map. Any nonempty open acceptance region $\mathcal{U} \subseteq \mathbb{R}^m$ in the latent space satisfying $\mathcal{U} \cap \mathrm{Im}(\Phi) \neq \varnothing$ induces a volumetric acceptance region $\Phi^{-1}(\mathcal{U})$ in the input space.*

The proof is straightforward. Since $\mathcal{U}$ is open and Φ is continuous, its preimage $\Phi^{-1}(\mathcal{U})$ is also open in $\mathbb{R}^n$. Moreover, $\mathcal{U} \cap \mathrm{Im}(\Phi) \neq \varnothing$ implies that $\Phi^{-1}(\mathcal{U})$ is nonempty. Every nonempty open subset of $\mathbb{R}^n$ has positive Lebesgue measure and is therefore volumetric.

A.2 Proof of Theorem 1

We first prove a simpler claim.

Lemma 1. *Let $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ be fixed and let $p(\boldsymbol{x} \mid \hat{\boldsymbol{x}})$ denote the posterior distribution over normal samples. Consider the unconstrained optimization problem*

$$\underset{\boldsymbol{\mu} \in \mathbb{R}^d}{\text{minimize}} \quad \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\|\boldsymbol{\mu} - \boldsymbol{x}\|^2 \right]. \tag{10}$$

Then the unique minimizer is

$$\boldsymbol{\mu}^* = \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\boldsymbol{x} \right]. \tag{11}$$

Proof. Consider the following derivations:

$$\begin{aligned}
\mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\|\boldsymbol{\mu} - \boldsymbol{x}\|^2 \right] &= \int \|\boldsymbol{\mu} - \boldsymbol{x}\|^2 p(\boldsymbol{x} \mid \hat{\boldsymbol{x}}) \, \mathrm{d}\boldsymbol{x} \\
&= \int \left(\|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\mu}^\top \boldsymbol{x} + \|\boldsymbol{x}\|^2 \right) p(\boldsymbol{x} \mid \hat{\boldsymbol{x}}) \, \mathrm{d}\boldsymbol{x} \\
&= \|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\mu}^\top \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\boldsymbol{x} \right] + \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\|\boldsymbol{x}\|^2 \right].
\end{aligned}$$

We now look for stationary points of the mapping

$$g \colon \boldsymbol{\mu} \mapsto \|\boldsymbol{\mu}\|^2 - 2\boldsymbol{\mu}^\top \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\boldsymbol{x} \right] + \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\|\boldsymbol{x}\|^2 \right]$$

by setting its gradient to zero. It follows that

$$\nabla_{\boldsymbol{\mu}} g = 2\boldsymbol{\mu} - 2\mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\boldsymbol{x} \right] = \mathbf{0}.$$

Hence,

$$\boldsymbol{\mu}^* = \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\boldsymbol{x} \right].$$

Since $g(\boldsymbol{\mu})$ is strictly convex, this is the unique optimal solution. □

We now prove Theorem 1.

Proof. First, we rewrite the objective in (6) as follows:

$$\begin{aligned}
\mathbb{E}_{p(\boldsymbol{x})} \left[\mathbb{E}_{q(\hat{\boldsymbol{x}}|\boldsymbol{x})} \left[\|f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}) - \boldsymbol{x}\|^2 \right] \right] &= \int \int \|f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}) - \boldsymbol{x}\|^2 p(\boldsymbol{x}) \, q(\hat{\boldsymbol{x}} \mid \boldsymbol{x}) \, \mathrm{d}\hat{\boldsymbol{x}} \, \mathrm{d}\boldsymbol{x} \\
&\overset{(*)}{=} \int \int \|f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}) - \boldsymbol{x}\|^2 q(\hat{\boldsymbol{x}}) \, p(\boldsymbol{x} \mid \hat{\boldsymbol{x}}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\hat{\boldsymbol{x}} \\
&= \mathbb{E}_{q(\hat{\boldsymbol{x}})} \left[\mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\|f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}) - \boldsymbol{x}\|^2 \right] \right].
\end{aligned}$$

In step $(*)$, we used Bayes' theorem,

$$p(\boldsymbol{x}) \, q(\hat{\boldsymbol{x}} \mid \boldsymbol{x}) = q(\hat{\boldsymbol{x}}) \, p(\boldsymbol{x} \mid \hat{\boldsymbol{x}}).$$

For each fixed $\hat{\boldsymbol{x}}$, Lemma 1 implies that the unique optimal output is

$$f^*(\hat{\boldsymbol{x}}) = \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\boldsymbol{x} \right]. \tag{12}$$

Provided sufficient capacity of $\{f_{\boldsymbol{\theta}}\}$, this mapping can be realized for q-almost every $\hat{\boldsymbol{x}}$ and therefore minimizes (6).

It follows directly from the definition of $N(\hat{\boldsymbol{x}})$ that $\boldsymbol{x}_{N(\hat{\boldsymbol{x}})} = \hat{\boldsymbol{x}}_{N(\hat{\boldsymbol{x}})}$ for normal samples compatible with $\hat{\boldsymbol{x}}$ under the corruption process. Consequently,

$$\begin{aligned}
f^*_{N(\hat{\boldsymbol{x}})}(\hat{\boldsymbol{x}}) &= \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\boldsymbol{x}_{N(\hat{\boldsymbol{x}})} \right] \\
&= \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\boldsymbol{x}_{N(\hat{\boldsymbol{x}})} \;\middle|\; \boldsymbol{x}_{N(\hat{\boldsymbol{x}})} = \hat{\boldsymbol{x}}_{N(\hat{\boldsymbol{x}})} \right] \\
&= \hat{\boldsymbol{x}}_{N(\hat{\boldsymbol{x}})}.
\end{aligned}$$

For the corrupted region, (12) directly gives

$$f^*_{A(\hat{\boldsymbol{x}})}(\hat{\boldsymbol{x}}) = \mathbb{E}_{p(\boldsymbol{x}|\hat{\boldsymbol{x}})} \left[\boldsymbol{x}_{A(\hat{\boldsymbol{x}})} \right].$$

Together, these establish (7). □

A.3 Fixed Point Convergence under Iterative Model Application

The following theorem provides the complete version of Theorem 2 presented in the main body of the paper. It establishes convergence to a fixed point on the manifold under repeated application of the model, assuming a contraction property toward the data manifold.

As a first simplifying modeling assumption, we assume that for any given anomalous sample $\hat{\boldsymbol{x}} \in \mathbb{R}^d \setminus \mathcal{M}$, the anomalous index set $S := A(\hat{\boldsymbol{x}})$ is uniquely identifiable, since ambiguous boundary cases occur with negligible probability in practice.

As an additional simplifying assumption, we assume that for a given anomalous input $\hat{\boldsymbol{x}}$, the anomaly index set S remains unchanged (up until convergence) under iterative application of the model.

Theorem 3 (Fixed Point Iteration)**.** *Let $\mathcal{M} \subset \mathbb{R}^d$ be a closed manifold and let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a non-linear operator satisfying:*

(i) (Conservativity) For every input $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ with anomaly index set S, the operator preserves the complementary coordinates corresponding to normal regions:

$$f_{\bar{S}}(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{x}}_{\bar{S}}.$$

(ii) (Slice Contraction) There exists $\rho \in (0, 1)$ such that for all $\boldsymbol{u} \in \mathbb{R}^d$,

$$\operatorname{dist}\big(f_S(\boldsymbol{u}), \mathcal{M}_S(\boldsymbol{u})\big) \le \rho \cdot \operatorname{dist}\big(\boldsymbol{u}_S, \mathcal{M}_S(\boldsymbol{u})\big),$$

where the slice

$$\mathcal{M}_S(\boldsymbol{u}) := \{\boldsymbol{x}_S \colon \boldsymbol{x} \in \mathcal{M},\ \boldsymbol{x}_{\bar{S}} = \boldsymbol{u}_{\bar{S}}\}$$

and

$$\operatorname{dist}(z, A) := \inf_{y \in A} \|z - y\|_2 \qquad \text{for } z \in \mathbb{R}^k,\ A \subset \mathbb{R}^k,\ k = |S|.$$

(iii) (Vanishing correction near the slice) There exists $C > 0$ such that for all $\boldsymbol{u} \in \mathbb{R}^d$,

$$\|f_S(\boldsymbol{u}) - \boldsymbol{u}_S\|_2 \le C \operatorname{dist}\big(\boldsymbol{u}_S, \mathcal{M}_S(\boldsymbol{u})\big).$$

Then the following statements hold:

(a) A point $\boldsymbol{x} \in \mathbb{R}^d$ belongs to $\mathcal{M}$ if and only if it is a fixed point of f.

(b) For every anomalous sample $\hat{\boldsymbol{x}}$ with anomaly index set S, there exists $\boldsymbol{x}^* \in \mathcal{M}$ such that $\boldsymbol{x}^*_{\bar{S}} = \hat{\boldsymbol{x}}_{\bar{S}}$ and

$$\lim_{n \to \infty} f^{(n)}(\hat{\boldsymbol{x}}) = \boldsymbol{x}^*.$$

Proof. We prove (a) and (b).

(a) ($\Rightarrow$) Let $\boldsymbol{x} \in \mathcal{M}$. By the convention, normal samples have empty anomaly index set, i.e. $S = \emptyset$ and hence $\bar{S} = \{1, \dots, d\}$. Applying (i) yields

$$f(\boldsymbol{x}) = f_{\bar{S}}(\boldsymbol{x}) = \boldsymbol{x}_{\bar{S}} = \boldsymbol{x},$$

so $\boldsymbol{x}$ is a fixed point of f.

(a) ($\Leftarrow$) Let $\boldsymbol{x} \in \mathbb{R}^d$ be a fixed point, i.e. $f(\boldsymbol{x}) = \boldsymbol{x}$, and let S be its anomaly index set. If $S = \emptyset$, then $\boldsymbol{x}$ is normal by definition and hence $\boldsymbol{x} \in \mathcal{M}$.

Assume $S \neq \emptyset$. Using (ii) with $\boldsymbol{u} = \boldsymbol{x}$ and $f_S(\boldsymbol{x}) = \boldsymbol{x}_S$ gives

$$\operatorname{dist}\big(\boldsymbol{x}_S, \mathcal{M}_S(\boldsymbol{x})\big) = \operatorname{dist}\big(f_S(\boldsymbol{x}), \mathcal{M}_S(\boldsymbol{x})\big) \le \rho \operatorname{dist}\big(\boldsymbol{x}_S, \mathcal{M}_S(\boldsymbol{x})\big).$$

Since $\rho \in (0, 1)$, we obtain

$$\operatorname{dist}(\boldsymbol{x}_S, \mathcal{M}_S(\boldsymbol{x})) = 0.$$

By definition of $\mathcal{M}_S(\boldsymbol{x})$, there exists a sequence $(\boldsymbol{y}^{(n)}) \subset \mathcal{M}$ satisfying

$$\boldsymbol{y}^{(n)}_{\bar{S}} = \boldsymbol{x}_{\bar{S}} \qquad \text{and} \qquad \boldsymbol{y}^{(n)}_S \to \boldsymbol{x}_S.$$

Hence $\boldsymbol{y}^{(n)} \to \boldsymbol{x}$ in $\mathbb{R}^d$. Moreover, the set

$$\{\boldsymbol{y} \in \mathcal{M} : \ \boldsymbol{y}_{\bar{S}} = \boldsymbol{x}_{\bar{S}}\}$$

is closed as the intersection of the closed set $\mathcal{M}$ with an affine subspace. Since each $\boldsymbol{y}^{(n)}$ belongs to a closed set and $\boldsymbol{y}^{(n)} \to \boldsymbol{x}$, we conclude that the limit $\boldsymbol{x}$ also belongs to this set. Therefore $\boldsymbol{x} \in \mathcal{M}$.

(b) Fix an anomalous sample $\hat{\boldsymbol{x}}$ with anomaly index set S and define the iterates

$$\boldsymbol{x}^{(0)} := \hat{\boldsymbol{x}}, \qquad \boldsymbol{x}^{(n+1)} := f(\boldsymbol{x}^{(n)}).$$

By (i), for all $n \geq 0$,

$$\boldsymbol{x}_{\bar{S}}^{(n)} = \hat{\boldsymbol{x}}_{\bar{S}}.$$

Consequently, the slice is constant along the iterates:

$$\mathcal{M}_S(\boldsymbol{x}^{(n)}) = \{\boldsymbol{x}_S : \boldsymbol{x} \in \mathcal{M},\ \boldsymbol{x}_{\bar{S}} = \boldsymbol{x}_{\bar{S}}^{(n)}\} = \{\boldsymbol{x}_S : \boldsymbol{x} \in \mathcal{M},\ \boldsymbol{x}_{\bar{S}} = \hat{\boldsymbol{x}}_{\bar{S}}\} =: \mathcal{A}.$$

Define

$$\delta_n := \operatorname{dist}(\boldsymbol{x}_S^{(n)}, \mathcal{A}).$$

Applying (ii) with $\boldsymbol{u} = \boldsymbol{x}^{(n)}$ yields

$$\delta_{n+1} = \operatorname{dist}\big(\boldsymbol{x}_S^{(n+1)}, \mathcal{A}\big) = \operatorname{dist}\big(f_S(\boldsymbol{x}^{(n)}), \mathcal{M}_S(\boldsymbol{x}^{(n)})\big) \leq \rho \operatorname{dist}\big(\boldsymbol{x}_S^{(n)}, \mathcal{M}_S(\boldsymbol{x}^{(n)})\big) = \rho\, \delta_n,$$

and hence

$$\delta_n \leq \rho^n \delta_0 \qquad \text{for all } n.$$

We now show that $(\boldsymbol{x}_S^{(n)})$ is a Cauchy sequence. By (iii),

$$\|\boldsymbol{x}_S^{(n+1)} - \boldsymbol{x}_S^{(n)}\|_2 = \|f_S(\boldsymbol{x}^{(n)}) - \boldsymbol{x}_S^{(n)}\|_2 \leq C \operatorname{dist}\big(\boldsymbol{x}_S^{(n)}, \mathcal{M}_S(\boldsymbol{x}^{(n)})\big) = C\, \delta_n \leq C\, \rho^n \delta_0.$$

Therefore, for $m > n$,

$$\begin{aligned}\|\boldsymbol{x}_S^{(m)} - \boldsymbol{x}_S^{(n)}\|_2 &\leq \sum_{t=n}^{m-1} \|\boldsymbol{x}_S^{(t+1)} - \boldsymbol{x}_S^{(t)}\|_2 \leq C\delta_0 \sum_{t=n}^{m-1} \rho^t = C\delta_0 \left(\sum_{t=0}^{m-1} \rho^t - \sum_{t=0}^{n-1} \rho^t\right) \\ &= C\delta_0 \left(\frac{1-\rho^m}{1-\rho} - \frac{1-\rho^n}{1-\rho}\right) \leq \frac{C\delta_0}{1-\rho}\, \rho^n \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

Thus $(\boldsymbol{x}_S^{(n)})$ converges to some limit $\boldsymbol{x}_S^* \in \mathbb{R}^{|S|}$.

Since $\operatorname{dist}(\boldsymbol{x}_S^{(n)}, \mathcal{A}) = \delta_n \leq \rho^n \delta_0 \to 0$ and $\boldsymbol{x}_S^{(n)} \to \boldsymbol{x}_S^*$, it follows that $\operatorname{dist}(\boldsymbol{x}_S^*, \mathcal{A}) = 0$. Because $\mathcal{A}$ is closed, we conclude that $\boldsymbol{x}_S^* \in \mathcal{A}$.

Moreover, $\boldsymbol{x}_{\bar{S}}^{(n)} = \hat{\boldsymbol{x}}_{\bar{S}}$ for all n, hence $\boldsymbol{x}^{(n)} \to \boldsymbol{x}^*$ where $\boldsymbol{x}_{\bar{S}}^* = \hat{\boldsymbol{x}}_{\bar{S}}$ and $\boldsymbol{x}_S^* \in \mathcal{A}$. By definition of $\mathcal{A}$, this implies $\boldsymbol{x}^* \in \mathcal{M}$, which concludes the proof. □

The remaining question is under which conditions the model satisfies assumptions (ii) and (iii) of the theorem. In the following, we provide an intuition together with a practical heuristic used in our experiments, while leaving a rigorous formal investigation to future work. Specifically, we rely on out-of-distribution images as a source of anomalous patterns. The main intuition is that the resulting perturbations are predominantly orthogonal to the manifold of normal data.

Corruption model. Let $p(\boldsymbol{x})$ denote the distribution of normal data supported on a differentiable manifold $\mathcal{M} \subset \mathbb{R}^d$. A partially corrupted observation is generated as

$$\hat{\boldsymbol{x}} = \boldsymbol{x} + \alpha\,\boldsymbol{\delta}, \qquad \boldsymbol{x} \sim p, \quad \alpha \sim \pi, \qquad \boldsymbol{\delta}_{N(\hat{\boldsymbol{x}})} = \mathbf{0}, \tag{13}$$

where π is a transparency distribution on $(0, 1]$.

We assume that the corruption vector is predominantly normal to the manifold. That is, there exists $\eta \in [0, 1)$ such that, for typical pairs $(\boldsymbol{x}, \boldsymbol{\delta})$,

$$\big\|\Pi_{T_{\boldsymbol{x}}\mathcal{M}}(\boldsymbol{\delta})\big\| \le \eta\,\big\|\Pi_{N_{\boldsymbol{x}}\mathcal{M}}(\boldsymbol{\delta})\big\|, \tag{14}$$

where $\Pi_{T_{\boldsymbol{x}}\mathcal{M}}$ and $\Pi_{N_{\boldsymbol{x}}\mathcal{M}}$ denote the orthogonal projections onto the tangent and normal subspaces of the manifold at $\boldsymbol{x}$, respectively. In practice, we approximate this condition by setting $\boldsymbol{\delta}_{A(\hat{\boldsymbol{x}})} = \boldsymbol{y}_{A(\hat{\boldsymbol{x}})} - \boldsymbol{x}_{A(\hat{\boldsymbol{x}})}$ for some out-of-distribution image $\boldsymbol{y} \notin \mathcal{M}$ that is sufficiently distant from the manifold. Informally, this corresponds to using images that look noticeably different from normal examples. Furthermore, we assume that π assigns nonzero probability to arbitrarily small values of α.

Compatible normal sources. For a partially corrupted input $\hat{\boldsymbol{x}}$ we define the set of normal samples compatible with $\hat{\boldsymbol{x}}$ as

$$\mathrm{CNS}(\hat{\boldsymbol{x}}) = \Big\{\boldsymbol{x} \in \mathcal{M} \;:\; \boldsymbol{x}_{N(\hat{\boldsymbol{x}})} = \hat{\boldsymbol{x}}_{N(\hat{\boldsymbol{x}})} \text{ and } \exists\,\alpha \in (0, 1],\ \boldsymbol{\delta} \text{ s.t. } \hat{\boldsymbol{x}} = \boldsymbol{x} + \alpha\boldsymbol{\delta}\Big\}. \tag{15}$$

We measure ambiguity on the corrupted region by the diameter

$$\mathrm{diam}\big(\mathrm{CNS}(\hat{\boldsymbol{x}})\big) := \sup_{\boldsymbol{x},\boldsymbol{x}' \in \mathrm{CNS}(\hat{\boldsymbol{x}})} \|\boldsymbol{x}_{A(\hat{\boldsymbol{x}})} - \boldsymbol{x}'_{A(\hat{\boldsymbol{x}})}\|. \tag{16}$$

We now show how our corruption process induces the contraction toward the manifold that ensures properties (ii) and (iii) in Theorem 3.

Proposition 2. *Consider the following unconstrained optimization problem:*

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^k}{\textit{minimize}} \quad \mathbb{E}_{p(\boldsymbol{x})}\left[\mathbb{E}_{q(\hat{\boldsymbol{x}}|\boldsymbol{x})}\left[\|f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}) - \boldsymbol{x}\|^2\right]\right], \tag{17}$$

under the corruption model equation 13–equation 14. Let f^* denote a Bayes-optimal solution of the above optimization problem. Assume that the posterior distribution over compatible normal sources is locally balanced, in the sense that its mean remains uniformly separated from the boundary of the local compatible-source cone, and that the curvature of $\mathcal{M}$ is uniformly bounded in a neighborhood of interest. Then there exists a sufficiently small neighborhood $\mathcal{U}$ of $\mathcal{M}$ and a constant $C \in (0, 1)$ such that, for every admissible corrupted sample $\hat{\boldsymbol{x}} \in \mathcal{U}$,

$$\mathrm{diam}\big(\mathrm{CNS}(f^*(\hat{\boldsymbol{x}}))\big) \le C\,\mathrm{diam}\big(\mathrm{CNS}(\hat{\boldsymbol{x}})\big). \tag{18}$$

Proof. The assumption about the perturbations with normal-dominant components implies that locally the set $\mathrm{CNS}(\hat{\boldsymbol{x}})$ is confined within a cone around the normal direction of $\mathcal{M}$ in the corrupted coordinates of $\boldsymbol{\delta}$. By Theorem 1, the Bayes-optimal mapping produces the posterior mean of the compatible normal sources induced by the corruption process. Since the posterior distribution over the compatible normal sources is locally balanced, this mean remains close to the central axis of the corresponding compatible-source cone.

For a sufficiently small slice of a smooth manifold with uniformly bounded curvature κ, the deviation of the posterior mean from the corresponding manifold slice is of order

$$\mathrm{dist}(f^*(\hat{\boldsymbol{x}}), \mathcal{M}) = \mathcal{O}\big(\kappa D(\hat{\boldsymbol{x}})^2\big)\,,$$

where

$$D(\hat{\boldsymbol{x}}) := \mathrm{diam}\big(\mathrm{CNS}(\hat{\boldsymbol{x}})\big).$$

Moreover, the uniform normal-dominance condition in equation 14 bounds the aperture of the local cone of compatible normal sources. Because the posterior is locally balanced, its mean remains uniformly separated from the boundary of this cone and hence close to its central axis. Therefore, for a sufficiently small normal displacement

$$h := \mathrm{dist}(f^*(\hat{\boldsymbol{x}}), \mathcal{M})\,,$$

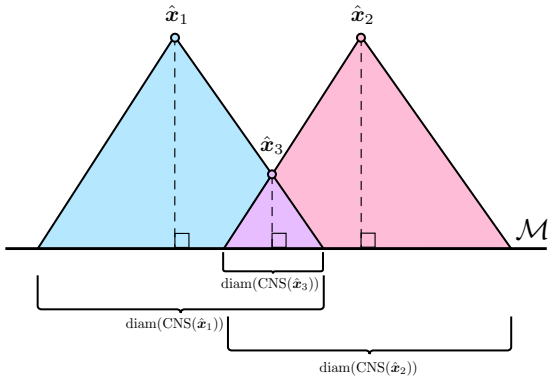

Figure 5: Illustration of the progressive reduction in the diameter of compatible normal sources under iterative application of the model. The manifold of normal data $\mathcal{M}$ is represented schematically by the horizontal black line. The leftmost triangle (blue) indicates the cone of compatible normal sources for the corrupted sample $\hat{\boldsymbol{x}}_1$, while the rightmost triangle (red) indicates the corresponding cone for $\hat{\boldsymbol{x}}_2$. The small triangle (lilac) illustrates that, as the corrected sample moves closer to $\mathcal{M}$, the set of compatible normal sources becomes more restricted, resulting in a smaller diameter.

the bounded cone aperture implies that all compatible normal sources of $f^*(\hat{\boldsymbol{x}})$ lie within a tangential extent proportional to h. Thus, there exists a constant $K_1 > 0$, independent of $\hat{\boldsymbol{x}}$ within the neighborhood considered, such that

$$D(f^*(\hat{\boldsymbol{x}})) \leq K_1 \operatorname{dist}(f^*(\hat{\boldsymbol{x}}), \mathcal{M})\,.$$

Combining this bound with

$$\operatorname{dist}(f^*(\hat{\boldsymbol{x}}), \mathcal{M}) = \mathcal{O}\big(\kappa D(\hat{\boldsymbol{x}})^2\big)$$

yields a constant $K > 0$ such that

$$D(f^*(\hat{\boldsymbol{x}})) \leq K\kappa D(\hat{\boldsymbol{x}})^2.$$

Restricting the neighborhood further such that $D(\hat{\boldsymbol{x}}) \leq D_{\max}$ and $K\kappa D_{\max} < 1$, and defining

$$C := K\kappa D_{\max},$$

we obtain

$$D(f^*(\hat{\boldsymbol{x}})) \leq CD(\hat{\boldsymbol{x}}), \qquad C \in (0, 1),$$

which proves equation 18. This idea is illustrated in Fig. 5. □

Together, Theorem 1 and Proposition 2 provide the geometric mechanism underlying assumptions (ii) and (iii) of Theorem 3. Theorem 1 states that a Bayes-optimal solution f^* maps each corrupted input to the posterior mean of its compatible normal sources, while Proposition 2 shows that, locally, the diameter of the corresponding compatible-source set decreases by a uniform multiplicative factor under repeated application of f^*.

Since the compatible normal sources are defined by fixing the normal coordinates and varying only the corrupted coordinates, their diameter measures the remaining ambiguity along the corresponding manifold slice. Geometrically, this diameter corresponds to the local chord length between compatible normal completions on the same slice, as illustrated in Fig. 6.

By Proposition 2,

$$D(f^*(\hat{\boldsymbol{x}})) \leq CD(\hat{\boldsymbol{x}}), \qquad 0 < C < 1,$$

within the neighborhood considered in the proposition. Consequently, the compatible-source ambiguity decreases geometrically under repeated application of f^*. Together with the quadratic relation between compatible-source diameter and curvature-induced deviation from the corresponding manifold slice, this provides the local slice-contraction behavior required in assumption (ii) of Theorem 3.

Moreover, the bounded cone aperture implies that the magnitude of the Bayes-optimal correction is controlled locally by the distance to the corresponding manifold slice. As this distance decreases

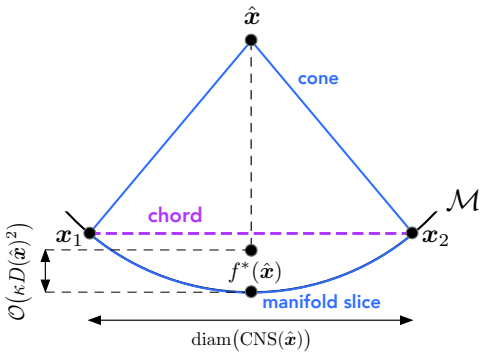

Figure 6: Geometric illustration of the compatible-source diameter for a corrupted sample $\hat{\boldsymbol{x}}$. The compatible normal sources lie on the same local manifold slice (blue) and span a chord (purple), whose length corresponds to $\mathrm{diam}(\mathrm{CNS}(\hat{\boldsymbol{x}}))$. The Bayes-optimal output $f^*(\hat{\boldsymbol{x}})$ is the posterior mean of these compatible sources and therefore generally lies off the curved slice. For a smooth slice with locally bounded curvature, this deviation decreases quadratically with the compatible-source diameter.

under repeated application, the correction therefore decreases and vanishes in the limit, providing the vanishing-correction behavior required in assumption (iii). Thus, iterative application of f^* progressively reduces both the ambiguity of the compatible normal sources and the magnitude of the remaining correction.

Role of transparency and corruption geometry. Training with transparent corruptions, i.e., a continuum of perturbation magnitudes, exposes the model to inputs arbitrarily close to the normal-data manifold and therefore determines the local behavior of the learned mapping in a neighborhood of $\mathcal{M}$. However, transparency alone is not sufficient to guarantee contractive behavior. A contraction bias is obtained only if the corruption directions that appear close to the manifold are predominantly normal to $\mathcal{M}$. In this case, small-amplitude corruptions correspond to displacements away from the manifold, and minimizing the reconstruction objective enforces corrections that reduce these displacements. Conversely, if small-amplitude corruptions are dominated by tangent components, they correspond to valid variations along the manifold and should not be removed.

Thus, transparency restricts the perturbation directions that influence the local Jacobian of the learned mapping: near $\mathcal{M}$, the model is exposed only to corruption patterns whose normal component dominates. Under this condition, the learned operator is biased toward contractive behavior in the corrupted coordinates.

On limitations for logical anomaly detection. While the proposed formulation provides a principled framework for SAD, where anomalies manifest as deviations from locally predictable appearance structure, it cannot fully address logical anomalies. Projection-based models implicitly approximate a manifold of normal appearance statistics, $\mathcal{M}_{\mathrm{app}}$, embedded in the broader manifold of natural images, $\mathcal{M}_{\mathrm{nat}}$. Logical anomalies, however, may violate higher-level constraints such as object counts, spatial relations, or semantic consistency while still preserving locally plausible appearance statistics. Such constraints can therefore be understood as defining a further restriction $\mathcal{M}_{\mathrm{logic}} \subset \mathcal{M}_{\mathrm{app}}$: an image may lie outside $\mathcal{M}_{\mathrm{logic}}$ while remaining within $\mathcal{M}_{\mathrm{app}}$. Consequently, logical anomalies need not induce an off-manifold deviation with respect to the learned appearance manifold and may produce little or no projection residual. Future work will investigate the integration of projection-based appearance modeling with structured modeling and graph-based approaches (Bauer et al., 2023; 2017; 2019a;b).

A.4 Sensitiivity Analysis

There are multiple hyperparameters in the presented framework that influence performance, including the balancing value $\lambda \in (0, 1)$ in the training objective (1) and the choice of evaluation metric

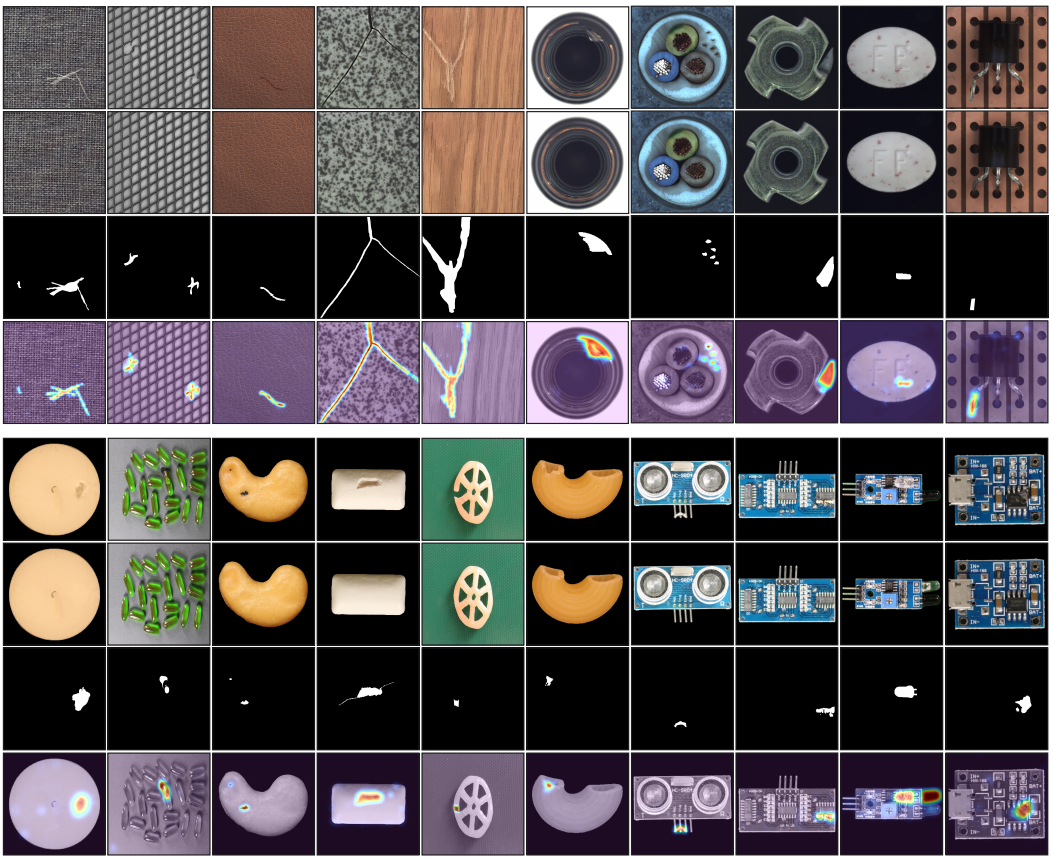

Figure 7: Representative examples of anomaly heatmaps on the MVTec AD and VisA datasets produced by our model. Each column shows (top to bottom) the input image, its reconstruction, the corresponding human-annotated anomaly mask for reference, and the predicted anomaly heatmap. In the VisA examples, the second column from the right illustrates an annotation error for a displaced object: only its new location is marked as anomalous, while the corresponding absence at the original location is left unannotated, producing an incomplete mask.

$\Delta \in \{\text{MSE}, \text{SSIM}\}$. We present the corresponding ablation studies on the texture categories of MVTec AD dataset in Table 2.

Overall, the results indicate that the proposed framework is relatively insensitive to the precise choice of λ, with performance remaining consistently high over the considered range. At the same time, choosing a larger value of λ can have a beneficial side effect when the training data is contaminated. In this case, the model may otherwise overfit to anomalous patterns present in the nominal training data and consequently fail to detect similar patterns during testing. Increasing λ places greater emphasis on correction quality in corrupted regions relative to preserving the uncorrupted image content, thereby reducing the tendency of the model to reproduce incidental artifacts and outliers occurring in the training data. Fig. 9 illustrates the corresponding effect on two examples from the carpet and tile categories of the MVTec AD dataset, where models trained with larger values of λ remain sensitive to anomalous patterns occurring in the training data. Furthermore, the choice of the discrepancy measure Δ exhibits a complementary trend across the pixel-level metrics. MSE tends to favor P-AP, suggesting that it produces more confident predictions of anomalous pixels, whereas SSIM tends to favor P-AUROC and P-PRO, reflecting stronger overall segmentation quality.

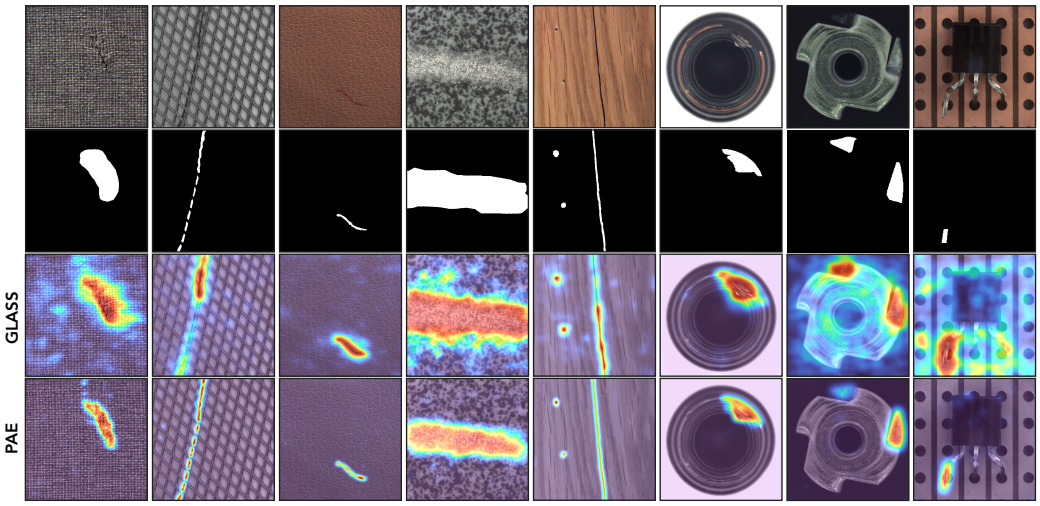

Figure 8: Qualitative comparison with GLASS. PAE produces sharper and more localized anomaly maps that are better aligned with the annotations.

Table 2: Sensitivity analysis for the parameter λ and evaluation metric Δ.

		MSE			*SSIM*		
Category	Metric	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.9$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.9$
carpet	I-AUROC	99.45 ± 0.15	99.46 ± 0.14	$\mathbf{99.74} \pm 0.27$	99.43 ± 0.16	99.60 ± 0.39	99.62 ± 0.38
	I-AP	99.85 ± 0.12	99.85 ± 0.12	$\mathbf{99.92} \pm 0.07$	99.68 ± 0.08	99.68 ± 0.08	99.86 ± 0.13
	P-AUROC	99.40 ± 0.03	99.60 ± 0.03	99.56 ± 0.03	99.51 ± 0.06	99.46 ± 0.05	$\mathbf{99.67} \pm 0.09$
	P-AP	83.20 ± 0.12	83.89 ± 0.11	84.77 ± 0.23	86.91 ± 0.35	87.01 ± 0.35	$\mathbf{87.06} \pm 0.36$
	P-PRO	95.99 ± 0.14	96.21 ± 0.25	97.56 ± 0.15	98.22 ± 0.45	98.30 ± 0.40	$\mathbf{98.49} \pm 0.39$
grid	I-AUROC	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$
	I-AP	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$
	P-AUROC	99.69 ± 0.01	99.71 ± 0.00	99.69 ± 0.01	99.71 ± 0.01	$\mathbf{99.78} \pm 0.01$	99.74 ± 0.00
	P-AP	70.59 ± 0.04	$\mathbf{70.85} \pm 0.01$	70.35 ± 0.20	64.51 ± 1.50	69.29 ± 0.16	67.47 ± 0.17
	P-PRO	99.00 ± 0.02	98.94 ± 0.06	99.00 ± 0.03	99.20 ± 0.01	$\mathbf{99.32} \pm 0.01$	99.20 ± 0.01
leather	I-AUROC	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	98.66 ± 1.32	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	99.07 ± 0.90
	I-AP	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	99.53 ± 0.46	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	99.68 ± 0.32
	P-AUROC	98.50 ± 0.32	99.09 ± 0.51	99.29 ± 0.49	99.60 ± 0.29	$\mathbf{99.67} \pm 0.03$	99.55 ± 0.39
	P-AP	54.36 ± 0.06	57.25 ± 0.04	49.92 ± 0.72	63.84 ± 0.16	$\mathbf{65.18} \pm 1.91$	58.51 ± 0.80
	P-PRO	96.58 ± 0.03	97.34 ± 0.05	96.53 ± 0.10	98.85 ± 0.08	$\mathbf{99.09} \pm 0.10$	98.66 ± 0.15
tile	I-AUROC	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$
	I-AP	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$
	P-AUROC	98.53 ± 0.03	98.63 ± 0.06	98.56 ± 0.05	98.42 ± 0.03	$\mathbf{98.86} \pm 0.34$	98.59 ± 0.61
	P-AP	80.08 ± 0.44	$\mathbf{81.72} \pm 1.13$	80.31 ± 0.37	79.02 ± 0.35	79.82 ± 0.31	79.16 ± 1.53
	P-PRO	90.08 ± 0.81	88.91 ± 1.02	93.91 ± 0.16	95.66 ± 0.91	$\mathbf{96.27} \pm 0.01$	95.98 ± 0.59
wood	I-AUROC	99.65 ± 0.10	99.91 ± 0.01	98.82 ± 0.10	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	99.91 ± 0.01
	I-AP	99.89 ± 0.06	99.97 ± 0.01	99.94 ± 0.03	$\mathbf{100} \pm 0.00$	$\mathbf{100} \pm 0.00$	99.97 ± 0.00
	P-AUROC	$\mathbf{98.39} \pm 0.74$	96.99 ± 0.79	98.08 ± 0.62	98.04 ± 0.59	97.19 ± 0.82	98.11 ± 0.73
	P-AP	76.21 ± 0.49	$\mathbf{79.93} \pm 0.64$	76.65 ± 0.55	74.59 ± 0.59	76.70 ± 0.67	76.37 ± 0.65
	P-PRO	89.83 ± 0.41	85.99 ± 0.17	93.25 ± 0.61	$\mathbf{96.68} \pm 0.37$	95.18 ± 0.02	95.93 ± 0.58

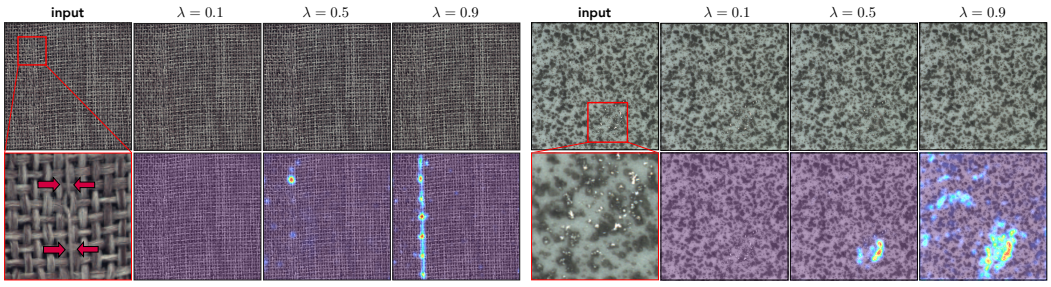

Figure 9: Effect of the parameter λ in (1) on reconstruction behavior for representative training examples (declared as normal but contaminated), from the carpet (left) and tile (right) categories of MVTec AD. For each example, columns show the input image with a zoom-in window, followed by the reconstruction and corresponding anomaly heatmap for different values of $\lambda \in \{0.1, 0.5, 0.9\}$. Larger values of λ preserve sensitivity to anomalous patterns present in the training data.