WHEN DOES UNSUPERVISED LEARNING SUCCEED OR FAIL? A POS PERSPECTIVE ON RECONSTRUCTION-BASED ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstruction-based unsupervised learning can fail in two opposing ways: a model may reconstruct anomalies too accurately or discard valid nominal variation. Using the Pursuit of Subspaces hypothesis, we characterize these failures through the meet, union, and join geometries induced by the nominal components. Excess learned range produces join blindness, while insufficient capacity produces meet preference and loss of nominal fidelity. We show that the compact nominal union is optimal among nominal faithful ranges and generally requires a nonlinear reconstruction map. Based on this geometry, we introduce Dynamic Push and Pull, which learns from controlled perturbations without anomaly labels, and nested manifold carving, which applies the same principle recursively in latent space. Experiments confirm the predicted changes in latent geometry across every tested Push and Pull configuration. The proposed methods improve reconstruction-based anomaly detection across standard benchmarks and unseen image degradations, while also improving pretrained ECG representations for downstream classification. These results connect reconstruction failures to identifiable geometric conditions and provide practical mechanisms for learning compact representations.

1 INTRODUCTION

The recently proposed Pursuit of Subspaces (PoS) hypothesis provides a postulate-based geometric foundation for representation learning (Yamac et al., 2026). It formalizes recurring properties of successful networks through four idealized and falsifiable postulates. **P1: Geometric compaction** states that ideal learning refines an m-dimensional candidate data manifold $\mathcal{M}_D \subset \mathbb{R}^n$ into a structured target $\mathcal{M} := \bigcup_{i=1}^{M} \mathcal{M}_{D_i} \subseteq \mathcal{M}_D$, where $\dim(\mathcal{M}_{D_i}) = k_i < m$. **P2: Nonlinear orthogonal projection** specifies the ideal learned operation: within an encoder-decoder encapsulation, the network acts as $P_{\mathcal{M}} := f_D \circ f_E : \mathbb{R}^n \to \mathcal{M}$, producing $\hat{s} = P_{\mathcal{M}}(s) \in \mathcal{M}$ and satisfying $P_{\mathcal{M}}^2 = P_{\mathcal{M}}$; hence $P_{\mathcal{M}}(\hat{s}) = \hat{s}$. **P3: Normal-space removal** realizes this projection as $\hat{s} = P_{\mathcal{M}}(s) = s - P_{\mathcal{M}}^{\perp}(s)$, where $P_{\mathcal{M}}^{\perp}(s) \in N_{\hat{s}}\mathcal{M}_{D_i}$ and $P_{\mathcal{M}} = I - P_{\mathcal{M}}^{\perp}$. Thus, the identity path preserves s, while the learned branch estimates the normal-space component to be removed; *normal space* denotes the geometric object and *residual branch* its architectural realization. **P4: Nested projections through depth** reapplies P1-P3 at every level: the representation at level $\ell - 1$ becomes the input to $P^{(\ell)} := f_D^{(\ell)} \circ f_E^{(\ell)} : \mathbb{R}^{n_{\ell-1}} \to \mathcal{M}^{(\ell)}$, where $\mathcal{M}^{(\ell)} := \bigcup_{i=1}^{M_\ell} \mathcal{M}_{D_i}^{(\ell)}$. Depth therefore realizes a hierarchy of nested, progressively refined nonlinear projections.

Remarks 1–3 of PoS Yamac et al. (2026) show that, once the network has learned $P_{\mathcal{M}}$ as a nonlinear orthogonal projector onto a union of submanifolds, it can be extended while preserving this property. Provided that the added transformation Φ is learned as an isometry, $\Phi \circ P_{\mathcal{M}} \circ \Phi^{-1}$ remains a nonlinear orthogonal projector onto $\Phi(\mathcal{M})$. These results explain how an established PoS projection can be preserved during network growth, but not how the network can be encouraged to learn $P_{\mathcal{M}}$ as such a projection in the first place. This paper addresses precisely that open problem:

> *Which learning mechanisms encourage neural modules to approximate nonlinear orthogonal projection onto compact unions of submanifolds?*

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.

The PoS postulates enable a falsifiable predict-then-measure approach: they specify both the target representation geometry and how representations should evolve as compactness increases. We derive two failure predictions and two corresponding remedies, then test the predicted dynamics at the level of individual samples and across configurations. We distinguish properties enforced directly by the proposed objectives from geometric conclusions that additionally assume the ideal projection in P2. Let $\mathcal{M}_D$ contain the component submanifolds $\{\mathcal{M}_{D_i}\}_{i=1}^{M}$, and locally represent them in common coordinates by tangent subspaces S_i. Define their *meet* $\mathcal{W} := \bigcap_i S_i$, local nominal *union* $\mathcal{U} := \bigcup_i S_i$, and *join* $\mathcal{V} := \sum_i S_i = \mathrm{span}(\bigcup_i S_i)$, the smallest linear subspace containing every S_i. Since $\mathcal{W} \subseteq \mathcal{U} \subseteq \mathcal{V}$, set inclusion reverses distance:

$$\mathrm{dist}(s, \mathcal{V}) \leq \mathrm{dist}(s, \mathcal{U}) \leq \mathrm{dist}(s, \mathcal{W}). \tag{1}$$

This relation predicts two limiting failures: a learned range approaching the join can reduce anomaly scores and cause false negatives, whereas one approaching the meet assigns positive scores to nominal points in $\mathcal{U} \setminus \mathcal{W}$, causing false positives. These distances coincide with reconstruction errors only under the nearest-point projection prescribed by P2; we do not assume that an arbitrary learned range lies between $\mathcal{W}$ and $\mathcal{V}$. Formal definitions and proofs are given in Appendix A.

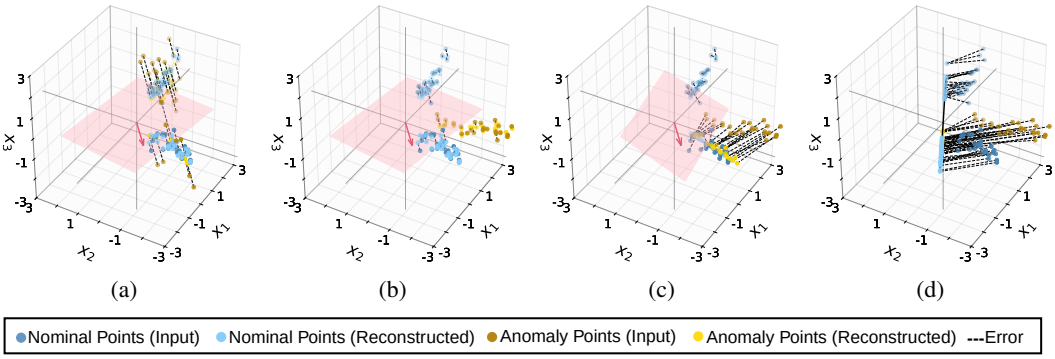

Figure 1: **Toy geometry of successful and failed anomaly detection.** The shaded plane denotes the join $\mathcal{V}$ of the two nominal subspaces; dark blue and ochre denote nominal and anomalous inputs, light blue and yellow their reconstructions, and dashed segments reconstruction errors. **(a)** A linear network with range $\mathcal{V}$ detects anomalies outside the join through their nonzero reconstruction errors. **(b)** The same range is blind to anomalies in $\mathcal{V} \setminus \mathcal{U}$, which can be reconstructed as accurately as nominal data. **(c)** A nonlinear network learns the compact union $\mathcal{U}$, preserving the nominal components while producing nonzero errors for anomalies both inside and outside $\mathcal{V}$. **(d)** Rank deficient learning produces a low dimensional compromise range $T_{\mathrm{under}} \subset \mathcal{V}$ that poorly represents both nominal and anomalous data, inflating both scores and reducing their fidelity to nominal union membership.

PoS Prediction I (join blindness): A reconstruction objective constrains the model on nominal samples but does not bound its learned range outside them. The range may therefore expand toward the join $\mathcal{V}$, violating the compaction required by P1. As Figure 1(a) illustrates, anomalies outside $\mathcal{V}$ remain detectable through reconstruction error. Under P2, however, an anomaly $s_a \in \mathcal{V} \setminus \mathcal{U}$ receives zero reconstruction error (Figure 1(b)), while even an approximate join can substantially reduce its score. Join blindness is therefore a loss of *selectivity*: the model retains directions that should have been excluded.

PoS Prediction II (meet preference): Let $\mathcal{W} := \bigcap_i S_i$ denote the directions shared by all nominal component tangent spaces. Limited capacity may preferentially retain $\mathcal{W}$ while discarding component specific directions; we call the limiting case $T = \mathcal{W}$ *meet collapse*. Proposition 10 establishes sufficient conditions under which a rank constrained MSE optimal linear autoencoder exhibits this preference. Through local tangent approximations, the same mechanism applies to nonlinear geometry. Under P2, meet collapse assigns positive errors to nominal points in $\mathcal{U} \setminus \mathcal{W}$, causing a loss of *fidelity*. Figure 1(d) visualizes this consequence through rank deficient underlearning; its meet is trivial, so it illustrates the fidelity failure rather than exact meet collapse.

Why the compact union requires nonlinearity: P1 identifies the desired learned range as the compact nominal union $\mathcal{U}$. Lemma 7 shows that $\mathcal{U}$ is optimal among nominal-faithful ranges, while Proposition 4 shows that a linear reconstruction map preserving $\mathcal{U}$ must also contain its join $\mathcal{V}$.

Reconstruction alone does not constrain the learned range toward $\mathcal{U}$: excess capacity can produce join blindness, whereas restricted capacity can discard component-specific directions and produce meet preference. Learning the compact union therefore requires both nonlinearity and an explicit separation signal. Dynamic Push and Pull supplies this missing signal. Figure 1(c) illustrates the desired case: a nonlinear network learns the compact union, preserving nominal reconstruction while detecting anomalies both inside and outside the join. Moreover, any geometrically nominal faithful range $T \supseteq \mathcal{U}$ satisfies $\mathrm{dist}(s, T) \leq \mathrm{dist}(s, \mathcal{U})$, where $\mathcal{U}$ maximizes ideal pointwise separation and, under nominal support on $\mathcal{U}$, expected geometric score gap among such ranges. Formal statements and proofs are given in Appendices A.5, A.2.1, and A.3.

Remedy I (dynamic Push and Pull learning): P1 to P3 motivate a training objective based on controlled perturbations of nominal samples, without requiring anomaly examples or class labels. Reconstructions are pulled toward the corresponding clean samples and pushed away from their perturbed versions, with a margin adapted to perturbation severity. This discourages the excess range responsible for join blindness, as detailed in (Section 3.1).

Remedy II (nested manifold carving): P4 motivates applying this compaction recursively in latent space. Each nested stage is trained to preserve nominal information while removing perturbation directions, while sufficient latent capacity reduces premature contraction toward shared directions (Section 3.2).

Although anomaly detection exposes these failure geometries particularly clearly, the remedies target unsupervised representation learning more generally. They learn data structure without semantic labels. Reconstruction error uses this structure to detect departures from the learned data, while downstream classifiers can reuse the same representation for label efficient prediction. Anomaly detection is therefore one application of the learned geometry, not the scope of the method.

We evaluate this progression in three stages. First, we compare our strategies with state of the art unsupervised anomaly detection methods on standard benchmarks. Second, we introduce a real world nominal only task: learning from high quality images and detecting unseen degradations without low quality training examples. Third, we move beyond anomaly detection by pretraining on unlabelled ECG signals, freezing the encoder, and training a small classifier with limited labelled data. We additionally test two PoS predictions for representation dynamics: relative disentanglement of nominal components and projection of samples outside the union toward the nominal union.

Our contributions are threefold. First, we derive two opposing failure predictions, join blindness and meet preference, and establish their consequences for anomaly scores, the need for nonlinearity, and the optimality of the compact nominal union. Second, we introduce Dynamic Push and Pull and nested manifold carving as complementary PoS-derived strategies for learning compact representations without anomaly examples or class labels. Third, we validate the framework through benchmark anomaly detection, real-world detection of unseen image degradations from high-quality nominal data, and unsupervised ECG pretraining for downstream classification, while testing two predicted dynamics of the learned geometry.

2 Related Work

Autoencoders are widely used for unsupervised anomaly detection, where reconstruction error serves as an anomaly score under the assumption that models trained on nominal data reconstruct anomalous inputs less accurately (Sakurada & Yairi, 2014; Zhai et al., 2016; Zhou & Paffenroth, 2017; Zhao et al., 2017). However, this assumption is not generally reliable, since autoencoders may also reconstruct unseen anomalies with small error (Zong et al., 2018; Gong et al., 2019; Zaheer et al., 2020). Existing approaches regularize the learned representation through latent compactness (Cao et al., 2019), sparse autoencoders (Nanaumi et al., 2026), denoising objectives (Marchi et al., 2015), or masked reconstruction (Huang et al., 2023; Georgescu, 2023; Elrashid et al., 2026), thereby encouraging the model to capture stable structure in nominal data rather than learning an unrestricted identity mapping. Architectural approaches instead constrain the reconstruction process using memory banks of nominal prototypes (Gong et al., 2019; Gao et al., 2023), latent density models (Zong et al., 2018), or adversarial and one class components (Perera et al., 2019).

Despite this broad effort, the mechanisms that cause autoencoder based anomaly detection to fail, and the conditions under which it can be reliable, remain insufficiently understood. Bouman & Heskes

(2025) show that autoencoders can accurately reconstruct anomalies through interpolation between nominal groups and extrapolation beyond the observed nominal data. We interpret both behaviors as manifestations of an excessively broad learned range and characterize the limiting case as join blindness; we further identify the opposite failure of meet preference, in which nominal directions specific to individual components are lost. Self supervised methods such as CutPaste, DRAEM, NSA, and CSI learn discriminative anomaly criteria from synthetic anomalies or distribution shifted views (Li et al., 2021; Zavrtanik et al., 2021; Schlüter et al., 2022; Tack et al., 2020). Reconstruction based perturbation based methods train autoencoders using noisy, synthetic, or adversarially perturbed nominal samples (Marchi et al., 2015; Astrid et al., 2021; 2024; Salehi et al., 2021). Methods that reconstruct the clean sample from its perturbation correspond to the Pull component of our formulation, whereas Dynamic Push and Pull additionally imposes an explicit separation margin from the perturbed input to target join blindness. Nested Manifold Carving complements this objective by recursively targeting off union directions retained in latent representations.

3 PoS-Guided Learning Strategies

3.1 Dynamic push-pull learning

Let d be a distance function chosen for the signal modality. For a nominal observation s_n, we generate a perturbed observation by adding a controlled perturbation e. Using $[z]_+ := \max\{z, 0\}$, define

$$\begin{aligned} \widetilde{s}_n := s_n + e, \qquad \widehat{s}_n := P_\theta(\widetilde{s}_n), \qquad \delta_e := d(\widetilde{s}_n, s_n), \\ \mathcal{L}_{\mathrm{PP}}(\theta; s_n, e) := d(\widehat{s}_n, s_n) + [\alpha\delta_e - d(\widehat{s}_n, \widetilde{s}_n)]_+ , \qquad \alpha > 0. \end{aligned} \tag{2}$$

The first term pulls the reconstruction toward the nominal observation, while the hinge term pushes it away from the perturbed input by the adaptive margin $\alpha\delta_e$. In the general noisy setting, s_n need only be concentrated near the nominal union. Exact membership is imposed only in the ideal result below. The experiments do not optimize the push–pull loss in isolation. They retain the ordinary reconstruction objective on unperturbed nominal observations, defined by

$$\mathcal{L}_{\mathrm{AE}}(\theta) := \mathbb{E}_{s_n}\left[d(P_\theta(s_n), s_n)\right]. \tag{3}$$

and optimize the complete objective

$$\begin{aligned} \mathcal{L}_{\mathrm{train}}(\theta) := \mathcal{L}_{\mathrm{AE}}(\theta) + \lambda_{\mathrm{PP}}\, \mathbb{E}_{s_n, e}\left[\mathcal{L}_{\mathrm{PP}}(\theta; s_n, e)\right] \\ + \lambda_{\mathrm{sp}}\, \Omega_{\mathrm{sp}}(\theta), \qquad \lambda_{\mathrm{PP}} > 0, \qquad \lambda_{\mathrm{sp}} \geq 0, \end{aligned} \tag{4}$$

where, Ω_{sp} is an optional sparsity penalty, with $\lambda_{\mathrm{sp}} = 0$ when sparsity is not used. Thus, $\mathcal{L}_{\mathrm{AE}}$ anchors the reconstruction of unperturbed nominal observations, while the push–pull term governs the response to their perturbed counterparts. The pairwise results below characterize the push-pull component; the operating point of the shared network is also shaped by the nominal reconstruction and optional sparsity terms in Eq. (4).

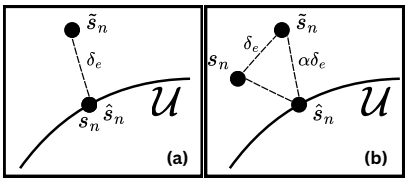

Figure 2: Ideal and realizable regimes of Lemma 1. (a) If $s_n \in \mathcal{U}$ and e is normal to $\mathcal{U}$, exact Pull gives $\widehat{s}_n = s_n$. For $\alpha \leq 1$, Push is inactive at this optimum but supplies a gradient whenever its margin is violated. (b) If $\mathrm{dist}_d(s_n, \mathcal{U}) = \eta(s_n) > 0$, margins $\alpha > 1$ become admissible and impose a constraint unavailable to Pull alone. The triangle inequality gives $\alpha\delta_e \leq \delta_e + \eta(s_n)$, recovering the zero-push-error case of Eq. (6).

Lemma 1 (Push and pull guarantees)**.** *Let d be a metric. Write Eq. (2) as $\mathcal{L}_{\mathrm{PP}} = \mathcal{L}_{\mathrm{pull}} + \mathcal{L}_{\mathrm{push}}$, and define $T := \mathrm{range}(P_\theta)$, $\rho_{\theta,d}(x) := d(x, P_\theta(x))$, and $\mathrm{dist}_d(x, T) := \inf_{t\in T} d(x, t)$. Suppose $\eta(s_n) := \mathrm{dist}_d(s_n, \mathcal{U}) \leq \varepsilon$, where $\eta(s_n) = 0$ when $s_n \in \mathcal{U}$. If $\mathcal{L}_{\mathrm{pull}} \leq \varepsilon_{\mathrm{pull}}$ and $\mathcal{L}_{\mathrm{push}} \leq \varepsilon_{\mathrm{push}}$, then $\mathrm{dist}_d(s_n, T) \leq \varepsilon_{\mathrm{pull}}$ and $\rho_{\theta,d}(\widetilde{s}_n) \geq \alpha\delta_e - \varepsilon_{\mathrm{push}}$. Under P2, these bounds jointly give*

$$\mathrm{dist}_d(\widetilde{s}_n, T) \geq \max\{\delta_e - \varepsilon_{\mathrm{pull}}, \alpha\delta_e - \varepsilon_{\mathrm{push}}, 0\}. \tag{5}$$

If, in addition, T provides exact nominal coverage $\mathcal{U} \subseteq T$, the triangle inequality gives $\mathrm{dist}_d(\widetilde{s}_n, T) \le \delta_e + \eta(s_n)$. *Consequently, for $\delta_e > 0$,*

$$\alpha \le 1 + \frac{\eta(s_n) + \varepsilon_{\mathrm{push}}}{\delta_e}. \tag{6}$$

At zero push error, this bound permits $\alpha > 1$ precisely when $\eta(s_n) > 0$.

The proof is given in Appendix A.6. The guarantee is local to the sampled perturbations. Suppressing join blindness throughout $\mathcal{V} \setminus \mathcal{U}$ therefore requires the perturbation sampler to probe this region sufficiently; otherwise, only the sampled neighborhoods are certifiably carved. Perturbations with components in $S_i^{\perp} \cap \mathcal{V}$ directly probe directions inside the join but outside the active nominal component.

3.2 Nested manifold carving

Following P4, we apply push–pull recursively in latent space. At level ℓ, the previously learned encoder–decoder pairs are frozen, their encoder stack maps s_n to $x_n^{(\ell-1)}$, and a new pair (E_ℓ, D_ℓ) is trained on this representation using the push–pull objective. The resulting reconstruction is then decoded through all preceding decoders. Each level therefore carves the representation inherited from the level above, while the pull term preserves nominal information and the push term prevents the newly added pair from simply reproducing perturbed latent inputs. The latent width is chosen sufficiently large to retain component-specific nominal directions and avoid premature meet preference. Let $\overline{T}^{(\ell)}$ denote the range of the complete reconstruction map through level ℓ, decoded back to the input space. Under the compatibility, capacity, and exact-optimization conditions stated in Appendix B, the construction yields

$$\overline{T}^{(1)} \supseteq \overline{T}^{(2)} \supseteq \cdots \supseteq \overline{T}^{(L)} \supseteq \mathcal{U}. \tag{7}$$

Thus, successive levels remove probed off-union regions without sacrificing nominal coverage. Eq. (7) describes the ideal construction target; its capacity, nonlinearity, and perturbation-coverage conditions are detailed in Appendix B, with the linear obstruction given in Appendix A.2.1.

3.3 PoS-predicted representation dynamics

PoS predicts how latent geometry changes as a network approaches the ideal behavior defined by P1 to P4. Let x and $\widehat{x}$ denote, respectively, less compact and more compact representations of the same samples, where compactness refers to agreement with the PoS postulates rather than merely smaller variance. From x to $\widehat{x}$, P1 and P4 predict increasing relative separation of nominal components, meaning that their radii contract more than the distances between their centers. P2 to P4 further predict that off-union anomalies move toward nominal components. Consequently, although nominal structure becomes more compact and separated, nominal and anomalous representations can become less distinguishable. A more compact pretrained network can therefore provide worse features for a downstream classifier that must distinguish anomalies from nominal samples. We test these predictions in Section 4.2. The exact metrics and their computation are provided in Appendix D.

4 Experiments

4.1 A controlled study of the predicted failure modes

We first use MNIST to examine the complete geometric picture in a controlled setting. Following Bouman & Heskes (2025), digits 4, 5, and 7 are treated as nominal, while the remaining digits are anomalies. This setting is particularly informative because the nominal distribution is multimodal and different anomalous digits interact differently with its learned geometry.

Baseline and controlled modifications: We reproduce the convolutional autoencoder of Bouman & Heskes (2025) for fair comparison, including its two-dimensional bottleneck and evaluation protocol. We then retain the same autoencoder backbone but increase the Stage-0 latent dimension from 2 to 9, reducing the capacity restriction that can merge component-specific nominal directions. Starting

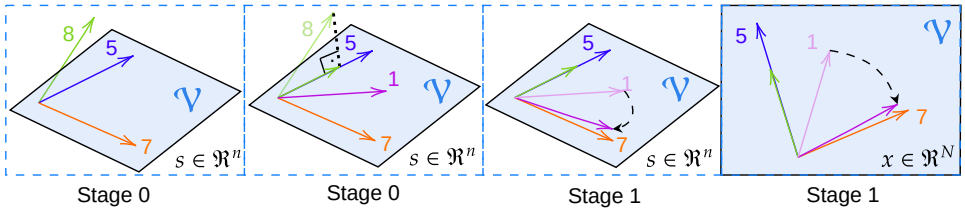

Figure 3: **Geometric interpretation of the two-stage MNIST experiment.** At Stage 0, the nominal digits 4, 5, and 7 define a compact union $\mathcal{U}$ within a coarser learned range $\mathcal{V}$. An anomaly such as 8, lying outside $\mathcal{V}$, produces a large Stage-0 reconstruction error. By contrast, digit 1 can lie inside $\mathcal{V} \setminus \mathcal{U}$ and therefore remain undetected at Stage 0. Stage 1 operates in the frozen Stage-0 latent space and further carves the retained range, projecting such off-union representations toward a nominal component. Anomalous information already removed by Stage 0 is unavailable to Stage 1.

from this higher-capacity model, we evaluate the individual and combined effects of nested carving, dynamic push-pull, masking, and sparsity. For the nested models, a second encoder–decoder pair is trained in the frozen Stage-0 latent space, whose dimension is expanded to 60 before carving. This controlled construction separates the effect of latent capacity from the effects of the proposed learning strategies.

Why different levels detect different anomalies: Figure 3 summarizes the mechanism investigated in this experiment. At Stage 0, anomalies lying outside the learned coarse range produce large reconstruction errors and are directly detectable. However, an anomalous digit such as 1 may lie close to the range spanning the nominal digits 4, 5, and 7. It can therefore be reconstructed with a small Stage-0 error despite not belonging to any nominal component. This is the predicted join-blindness mechanism. Nested carving addresses this ambiguity in the latent space. After the Stage-0 encoder and decoder are frozen, the second stage is trained to preserve the latent representations of nominal digits while rejecting their controlled perturbations. It can therefore remove off-union latent regions that Stage 0 retained and assign a larger error to anomalies such as 1. The converse situation is equally important. Some anomalous inputs are already mapped by Stage 0 onto, or very close to, one of the nominal components. Their anomalous information is then removed before entering the next latent stage, so Stage 1 cannot be expected to recover it. These anomalies remain detectable through the Stage-0 reconstruction error. Stage 0 and Stage 1 are therefore complementary: the former detects departures removed by the first projection, whereas the latter detects off-union structure that survives within the Stage-0 representation.

DAE versus AE-PP: Setting $\alpha = 0$ eliminates the Push hinge and reduces Push and Pull to a denoising autoencoder, while $\alpha > 0$ imposes the additional separation constraint characterized by Lemma 1 and Figure 2. DAE provides only a marginal improvement over AE and gains nothing from nesting, whereas AE-PP improves further at Stage 1, indicating that its gain cannot be explained by denoising alone; complete results are reported in Section C and Table 12.

Overall ablation results: Table 1 summarizes the central configurations, with complete results in Tables 10 and 11. Increasing Stage 0 capacity improves the reproduced baseline, but adding a reconstruction only nested module provides no further gain because it admits an identity solution. Push and Pull supplies the separation signal required for latent carving, improving both the single stage and nested models, while masking and sparsity provide complementary gains in the strongest configuration. Thus, the improvement arises from explicitly carving the learned range rather than from additional capacity, depth, or perturbation alone.

Per digit geometric analysis: The image and latent reconstruction errors provide direct evidence for the geometry predicted in Figure 3. Digit 8 produces a large image error but a small Stage 1 latent error, supporting the prediction that its anomalous direction lies largely outside the coarse range $\mathcal{V}$, is removed at Stage 0, and leaves a latent representation close to nominal digit 5. Conversely, digit 1 produces a small image error but a large latent error, supporting the prediction that it lies within $\mathcal{V} \setminus \mathcal{U}$ and is therefore retained by Stage 0 but differentiated through nested carving. Its Stage 1 reconstructions also become predominantly 7-like, supporting the predicted representation dynamics that increasing compactness moves off-union anomalies toward a nearby nominal component. The

Table 1: **Controlled MNIST study with nominal digits** $4, 5, 7$**.** The baseline reproduces Bouman & Heskes (2025) with a two-dimensional bottleneck. The remaining models use the same backbone with Stage-0 latent dimension 9. Nested configurations report the complete Stage-1 reconstruction score. PP denotes push–pull. Complete results are provided in Tables 10 and 11.

Configuration	F1	ROC-AUC	PR-AUC
Bouman-Heskes baseline	0.693	0.866	0.936
Higher-capacity AE	0.780	0.893	0.962
AE + mask	0.766	0.881	0.957
AE + PP	0.792	0.897	0.963
AE + nested + PP	0.798	0.909	0.967
AE + nested + PP + sparsity	0.808	0.928	0.972
AE + nested + PP + mask + sparsity	**0.815**	**0.929**	**0.973**

complete analysis, flag decomposition, and reconstructed images are provided in Section C.1, Table 13, and Figure 5.

4.2 Testing the predicted representation dynamics

We compare the less compact representation x with the more compact representation $\widehat{x}$ across four normal digit sets, two masking choices, and two sparsity choices. This gives 16 configurations for each training objective and 32 configurations in total. As summarized in Table 2, all 16 configurations trained with Push and Pull satisfy both PoS predictions. Across all 16 Push and Pull configurations, nominal relative separation increases and all four anomaly proximity indicators decrease. Ordinary AE training increases relative separation in only $7/16$ configurations and satisfies all four proximity predictions in $14/16$. Thus, Push and Pull produces the predicted PoS dynamics consistently across every tested normal set, masking choice, and sparsity choice. Complete results and metric definitions are provided in Appendix D.

Table 2: Observed PoS representation dynamics. Each objective includes 16 configurations obtained from four normal digit sets, two masking choices, and two sparsity choices.

Predicted change	**AE**	**AE + PP**
Increased nominal relative separation, $\Delta S > 0$	$7/16$	$\mathbf{16/16}$
All four anomaly proximity indicators decrease	$14/16$	$\mathbf{16/16}$

Table 3: One-class anomaly detection on Arrhythmia (Guvenir et al., 1997), measured by F1 score (%). AE-PP and AE-PoS results are reported as mean ± standard deviation over 10 seeds. Bold indicates the highest mean.

	Baselines												Ours	
Dataset	IF	OC-SVM	DSEBM-e	AnoGAN	DAGMM	ALAD	RCGAN	DSVDD	GOAD	MEMAE	MEMGAN	TOLL	AE-PP	AE-PoS
Arrhythmia	53.03	45.18	46.01	42.42	49.83	51.52	54.14	34.79	52.00	51.13	55.72	60.00	67.69±5.75	**70.77±5.75**

4.3 Standard One-Class Anomaly Detection

We evaluate one-class anomaly detection on MNIST, Fashion-MNIST, CIFAR-10, CIFAR-100, and Arrhythmia. Training uses only nominal samples. Following the validation protocol of TOLL (Ghafourian et al., 2025), checkpoints are selected on a validation set containing both nominal and anomalous samples. For a controlled comparison, all methods share the Stage 0 autoencoder architecture and optimization settings of TOLL. Our AE baseline is the reconstruction-only model obtained by setting the TOLL compactness coefficient to $\beta = 0$; it therefore uses the same backbone but not the TOLL compactness objective. AE-PP augments this baseline with Dynamic Push and Pull, whereas AE-PoS additionally applies Nested Manifold Carving to the frozen Stage 0 representation.

Table 4: One-class anomaly detection on CIFAR-10 (Krizhevsky, 2009), measured by AUC-ROC (%). Values are averaged over the ten nominal-class tasks. Full class-wise results, including uncertainty over 10 seeds for AE-PP and AE-PoS, are reported in Table 23.

	Baselines											Ours	
	OCGAN	VAE	DSVDD	DSEBM	DAGMM	IF	AnoGAN	ALAD	MEMAE	MEMGAN	TOLL	AE-PP	AE-PoS
Average	65.6	58.3	64.8	60.4	54.4	55.5	61.8	59.3	58.5	65.3	69.6	71.30	**71.73**

Table 5: Detection of unseen degradations in high-quality images. Entries are unweighted averages within each degradation family, reported as AUC-ROC and PR-AUC. Parentheses give the number of evaluation settings; the overall row averages all 16 settings. Complete results are provided in Table 26. Bold indicates the best result in each metric.

	AE (0.83M)		VAE (83.7M)		AE-PP (0.83M)		AE-PoS (1.67M)	
Family	ROC	PR	ROC	PR	ROC	PR	ROC	PR
Additive noise (1)	**1.000**	**1.000**	0.998	0.998	**1.000**	**1.000**	**1.000**	**1.000**
Blur (7)	0.568	0.536	0.133	0.334	0.889	0.881	**0.903**	**0.896**
Rain (5)	0.440	0.462	0.791	0.687	0.946	0.931	**0.958**	**0.938**
Underwater (3)	0.295	0.378	0.449	0.439	0.967	0.962	**0.985**	**0.981**
Overall (16)	0.504	0.512	0.452	0.506	0.929	0.919	**0.941**	**0.932**

As shown in Tables 3 and 4, AE-PoS reaches 70.77% F1 on Arrhythmia, compared with 60.00% for the strongest baseline, and 71.73% AUC-ROC on CIFAR-10, compared with 69.6%. Nested Manifold Carving improves AE-PP from 67.69% to 70.77% on Arrhythmia and from 71.30% to 71.73% on CIFAR-10, with gains across all ten CIFAR-10 nominal classes. Together with the controlled multi-component MNIST experiment in Section 4.1, these results suggest that Stage 1 provides further gains when the Stage-0 representation retains residual multiple components, while largely preserving performance when Stage 0 is already compact, as in the controlled single-component MNIST setting and Fashion-MNIST. Complete results for MNIST, Fashion-MNIST, and CIFAR-100, together with experimental details, baseline definitions, dataset-specific training strategies, and the AE/AE-PP/AE-PoS ablation, are provided in Appendix E. The CIFAR-100 results further expose a limitation of the current reconstruction-error anomaly score, which remains behind transformation-based methods on several nominal superclasses. Figure 4 shows the predicted saturation: once the Push hinge remains active for most training pairs, its gradient no longer depends on α, and performance plateaus at $\alpha = 2$ or 3.

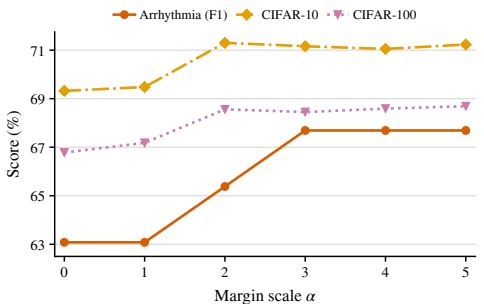

Figure 4: Effect of the margin scale α on Stage-0 AE-PP. Arrhythmia uses F1; CIFAR-10 and CIFAR-100 use AUC-ROC.

4.4 Detecting Unseen Image Degradations

Our objective is to learn a high-fidelity autoencoder representation of clean, high-quality images that preserves their fine-scale residual components while remaining sensitive to diverse unseen degradations. This contrasts with the VAE, whose probabilistic latent regularization reduces nominal-image fidelity and can preferentially represent low-frequency content, allowing strongly blurred images to be reconstructed more accurately than clean images and thereby reversing the desired anomaly-score ordering. AE-PP and AE-PoS are designed to preserve nominal detail while pushing degraded views away from the learned representation. Table 5 summarizes detection performance across additive noise, blur, rain, and underwater degradations, while Appendix E.7 provides the complete per-setting results, dataset construction, training strategy, clean-reference protocol, and reconstruction-fidelity analysis. As shown in Table 6, the reconstruction PSNR, computed between each input s and its reconstruction $\hat{s}$, increases for the VAE from 27.59 dB on clean images to 37.01 dB under the strongest blur. Thus, blurred anomalies receive lower reconstruction errors than

Table 6: Reconstruction PSNR (dB) under the strongest REDS blur. Higher PSNR indicates lower reconstruction error. VAE is the only model whose reconstruction quality improves substantially under blur, reversing the expected anomaly-score ordering.

	AE	VAE	AE-PP	AE-PoS
Clean	63.04	27.59	43.42	43.07
Blur 1.0	62.72	37.01	30.51	31.79
Δ	-0.32	$+\mathbf{9.42}$	-12.91	-11.28

clean inputs, reversing the expected score ordering and explaining the VAE's ROC-AUC below 0.5. Complete PSNR results are reported in Table 27. Training perturbations are limited to parametric blur kernels, sensor noise, resampling, and JPEG compression. REDS (Nah et al., 2019) and GoPro (Nah et al., 2017) therefore test transfer to realistic motion blur constructed through temporal averaging of high-frame-rate video; rain and underwater degradations are absent from training; and Kodak24 Gaussian noise provides an in-family control.

4.5 Transferable ECG Representations

Table 7: Frozen-encoder AUROC. Stage 0 and Stage 1 denote our depth-controlled models. Bold indicates the best result in each column.

	PTB-XL				ICBEB
Encoder	Super	Sub	Form	Rhythm	CPSC2018
ECG-CPC	0.8630	0.8860	0.8730	0.9380	0.9020
ECG-FM	0.8730	0.8660	0.7820	0.8650	0.9060
MERL	0.8870	0.8950	0.8640	0.8770	0.9140
ST-MEM	0.8954	0.8935	0.8548	0.9463	0.9559
Stage 0	0.8969	0.8917	0.8523	0.9449	0.9530
Stage 1	**0.9136**	**0.9277**	**0.8825**	**0.9482**	**0.9623**

Beyond representation-based anomaly detection, we evaluate nested carving as a general unsupervised representation-learning strategy. Clinical ECG labels require costly expert interpretation, whereas unlabeled recordings are abundant. We therefore learn structure from unlabeled ECGs and train only a linear diagnostic classifier on the frozen representation. Table 7 shows that nested carving improves performance across all five classification tasks; experimental and implementation details are provided in Appendix F.

5 Conclusion and Limitations

We characterized join blindness and meet preference as opposing failures of reconstruction-based unsupervised learning, and introduced Dynamic Push and Pull and Nested Manifold Carving to better approximate the compact nominal union prescribed by PoS, improving both anomaly detection and transferable representation learning. More importantly, the results demonstrate a predict-then-measure approach: Figure 3 predicts that anomalies outside the coarse range are removed at Stage 0, those in $\mathcal{V} \setminus \mathcal{U}$ survive and are exposed by latent carving, and carving moves them toward a nominal component; accordingly, digit 8 shows large reconstruction error but little latent change, whereas digit 1 is detected mainly by Stage 1 and becomes predominantly 7-like. At the population level, PoS predicts how geometry evolves as compactness increases: nominal components disentangle, with their relative distance (inter-centre distance normalized by component radius) increasing, while off-union samples move toward the nominal union. Both changes occur in all 16 Push and Pull configurations. PoS further predicts that increased compactness may weaken downstream discrimination between such samples and nominal data. These representation dynamics offer new design targets for unsupervised learning. Current limitations include metric dependence, manual depth selection, non-unique projections of structured anomalies, and validation of the dynamics only in a controlled multi-component setting, as discussed in Appendix G.

AI USE STATEMENT

In this work, we used large language model based generative AI tools to review and polish the manuscript, including improvements to grammar, clarity, organization, and consistency. All AI-assisted suggestions and revisions were critically reviewed and edited by the authors. We take full responsibility for the final content of this work, including its text, claims, and artifacts.

ETHICS STATEMENT

This work uses existing benchmark datasets and involves no new data collection or interaction with human participants. The clinical ECG datasets are used in de-identified form under their original licenses and access conditions. The ECG experiments evaluate representation learning and should not be interpreted as establishing clinical diagnostic validity; deployment in clinical settings would require independent validation and appropriate regulatory and ethical review.

REPRODUCIBILITY STATEMENT

The appendices provide the complete mathematical assumptions and proofs, model architectures, training objectives, optimization settings, preprocessing steps, perturbation schemes, and evaluation protocols. Upon acceptance, we will publicly release the complete source code, pretrained network checkpoints, all configuration files, and the exact test-set selections used for the unseen image-degradation experiments. Where the original dataset licenses prevent redistribution, we will provide the corresponding file lists and scripts needed to reconstruct the exact evaluation subsets from the source datasets. The release will also include scripts for reproducing the reported tables and figures.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1122–1131, 2017. doi: 10.1109/CVPRW.2017.150.

Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. GANomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision – ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pp. 622–637. Springer, 2019.

MA Al-Masud, Juan Miguel Lopez Alcaraz, and Nils Strodthoff. Benchmarking ecg fms: A reality check across clinical tasks. In *International Conference on Learning Representations*, volume 2026, pp. 146689–146715, 2026.

Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011. doi: 10.1109/TPAMI.2010.161.

Marcella Astrid, Muhammad Zaigham Zaheer, Jae-Yeong Lee, and Seung-Ik Lee. Learning not to reconstruct anomalies. *arXiv preprint arXiv:2110.09742*, 2021.

Marcella Astrid, Muhammad Zaigham Zaheer, Djamila Aouada, and Seung-Ik Lee. Exploiting autoencoder's weakness to generate pseudo anomalies. *Neural Computing and Applications*, 36 (23):14075–14091, 2024.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

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

Thomas Blumensath and Mike E Davies. Sampling theorems for signals from the union of finite-dimensional linear subspaces. *IEEE Transactions on Information Theory*, 55(4):1872–1882, 2009.

Roel Bouman and Tom Heskes. Autoencoders for anomaly detection are unreliable. *arXiv preprint arXiv:2501.13864*, 2025.

Van Loi Cao, Miguel Nicolau, and James McDermott. Learning neural representations for network anomaly detection. *IEEE Transactions on Cybernetics*, 49(8):3074–3087, 2019. doi: 10.1109/TCYB.2018.2838668.

Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. Image anomaly detection with generative adversarial networks. In *Joint european conference on machine learning and knowledge discovery in databases*, pp. 3–17. Springer, 2018.

Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Ozer Can Devecioglu, Serkan Kiranyaz, Mehmet Yamac, and Moncef Gabbouj. Expert operational gans: Towards real-color underwater image restoration. In *2025 IEEE International Conference on Image Processing Workshops (ICIPW)*, pp. 191–196, 2025. doi: 10.1109/ICIPW68931.2025.11386072.

Maab Elrashid, Anthony Deschênes, Cem Subakan, Mirco Ravanelli, Rémi Georges, and Michael Morin. Toward faithful explanations in acoustic anomaly detection. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5276–5280. IEEE, 2026.

Rich Franzen. Kodak lossless true color image suite. `http://r0k.us/graphics/kodak/`, 1999. Online accessed 24 Oct 2021.

Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Honghao Gao, Binyang Qiu, Ramón J. Durán Barroso, Walayat Hussain, Yueshen Xu, and Xinheng Wang. Tsmae: A novel anomaly detection approach for internet of things time series data using memory-augmented autoencoder. *IEEE Transactions on Network Science and Engineering*, 10(5): 2978–2990, 2023. doi: 10.1109/TNSE.2022.3163144.

Mariana-Iuliana Georgescu. Masked autoencoders for unsupervised anomaly detection in medical images. *Procedia Computer Science*, 225:969–978, 2023.

Amin Ghafourian, Huanyi Shui, Devesh Upadhyay, Rajesh Gupta, Dimitar Filev, and Iman Soltani. Targeted collapse regularized autoencoder for anomaly detection: Black hole at the center. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):10348–10358, 2025. doi: 10.1109/TNNLS.2024.3472456.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, 31, 2018.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714, 2019.

Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3512–3516. IEEE, 2019.

H. Altay Guvenir, Burak Acar, Haldun Muderrisoglu, and R. Quinlan. Arrhythmia. UCI Machine Learning Repository, 1997. DOI: https://doi.org/10.24432/C5BS32.

Chaoqin Huang, Qinwei Xu, Yanfeng Wang, Yu Wang, and Ya Zhang. Self-supervised masking for unsupervised anomaly detection and localization. *IEEE Transactions on Multimedia*, 25: 4426–4438, 2023. doi: 10.1109/TMM.2022.3175611.

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

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

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9659–9669, 2021. doi: 10.1109/CVPR46437.2021.00954.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1132–1140, 2017. doi: 10.1109/CVPRW.2017.151.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=ZvJ2lQQKjz`.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Yue M Lu and Minh N Do. A theory for sampling signals from a union of subspaces. *IEEE transactions on signal processing*, 56(6):2334–2345, 2008.

Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2017. doi: 10.1109/TIP.2016.2631888.

Alireza Makhzani and Brendan J Frey. Winner-take-all autoencoders. *Advances in neural information processing systems*, 28, 2015.

Erik Marchi, Fabio Vesperini, Florian Eyben, Stefano Squartini, and Björn Schuller. A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1996–2000, 2015. doi: 10.1109/ICASSP.2015.7178320.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: an open electrocardiogram foundation model. *JAMIA Open*, 8(5):ooaf122, 10 2025. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooaf122. URL https://doi.org/10.1093/jamiaopen/ooaf122.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International conference on learning representations*, volume 2024, pp. 15012–15035, 2024.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Tomoyasu Nanaumi, Yukino Tsuzuki, Junichi Okubo, Junichiro Fujii, and Takayoshi Yamashita. Spg: Sparse-projected guides with sparse autoencoders for zero-shot anomaly detection. *arXiv preprint arXiv:2604.02871*, 2026.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Lintao Peng, Chunli Zhu, and Liheng Bian. U-shape transformer for underwater image enhancement. *IEEE transactions on image processing*, 32:3066–3079, 2023.

Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2893–2901, 2019. URL https://api.semanticscholar.org/CorpusID:84186723.

Erick Andres Perez Alday, Annie Gu, Amit Shah, Chengyu Liu, Ashish Sharma, Salman Seyedi, Ali Bahrami Rad, Matthew Reyna, and Gari Clifford. Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020. *PhysioNet*, July 2022. doi: 10.13026/dvyd-kd57. Version 1.0.2.

Tom Pollard, Benjamin E Moody, Li-wei H Lehman, Brian J Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G Mark, and Thomas Heldt. Physionet as a global platform for biomedical research. *Nature Health*, pp. 1–4, 2026.

Antônio H. Ribeiro, Gabriela M. M. Paixão, Emilly M Lima, Manoel Horta Ribeiro, Marcelo M Pinto Filho, Paulo R. Gomes, Derick M. de Oliveira, Wagner Meira, Thomas Bo Schön, and Antônio Luiz P. Ribeiro. Code-15%: a large scale annotated dataset of 12-lead ecgs. 2021. URL https://api.semanticscholar.org/CorpusID:245000169.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning*, pp. 4393–4402. PMLR, 2018.

Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, pp. 4–11, 2014.

Mohammadreza Salehi, Atrin Arya, Barbod Pajoum, Mohammad Otoofi, Amirreza Shaeiri, Mohammad Hossein Rohban, and Hamid R Rabiee. Arae: Adversarially robust training of autoencoders improves novelty detection. *Neural Networks*, 144:726–736, 2021.

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 *Information Processing in Medical Imaging*, pp. 146–157, Cham, 2017. Springer International Publishing. ISBN 978-3-319-59050-9.

Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European conference on computer vision*, pp. 474–489. Springer, 2022.

Bernhard Schölkopf, Robert C. Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in Neural Information Processing Systems*, 12, 1999.

Stability AI. sd-vae-ft-mse: Improved autoencoders (fine-tuned kl-f8 decoder). `https://huggingface.co/stabilityai/sd-vae-ft-mse`, 2022. Hugging Face model repository.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Mehmet Yamac, Mert Duman, Ugur Akpinar, Felix Rojas Casadiego, Serkan Kiranyaz, Marcel van Gerven, and Moncef Gabbouj. Axiomatizing neural networks via pursuit of subspaces. *arXiv preprint arXiv:2605.20534*, 2026.

Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1685–1694. IEEE, 2017.

Ziyi Yang, Iman Soltani Bozchalooi, and Eric Darve. Regularized cycle consistent generative adversarial network for anomaly detection. In *European Conference on Artificial Intelligence (ECAI)*, volume 325, pp. 1618–1625, 2020. doi: 10.3233/FAIA200272. URL `https://ebooks.iospress.nl/DOI/10.3233/FAIA200272`.

Ziyi Yang, Teng Zhang, Iman Soltani Bozchalooi, and Eric Darve. Memory-augmented generative adversarial networks for anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2324–2334, 2021. doi: 10.1109/TNNLS.2021.3132928.

Fei Ye, Chaoqin Huang, Jinkun Cao, Maosen Li, Ya Zhang, and Cewu Lu. Attribute restoration framework for anomaly detection. *IEEE Transactions on Multimedia*, 24:116–127, 2022. doi: 10.1109/TMM.2020.3046884.

Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14171–14181. IEEE, 2020. doi: 10.1109/CVPR42600.2020.01419.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dræm–a discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8310–8319. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00822.

Houssam Zenati, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar. Adversarially learned anomaly detection. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 727–736. IEEE, 2018. URL https://ieeexplore.ieee.org/document/8594897.

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.

He Zhang and Vishal M. Patel. Density-aware single image de-raining using a multi-stream dense network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 695–704, 2018. doi: 10.1109/CVPR.2018.00079.

He Zhang, Vishwanath Sindagi, and Vishal M. Patel. Image de-raining using a conditional generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11): 3943–3956, 2020. doi: 10.1109/TCSVT.2019.2920407.

Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1933–1941, 2017. doi: 10.1145/3123266.3123451.

Jianwei Zheng, Huimin Chu, Daniele Carlo Struppa, Jianming Zhang, Sir Magdi Yacoub, Hesham Mohamed El-Askary, Anthony Chang, Louis Ehwerhemuepha, Islam Abudayyeh, Alexander Barrett, Guohua Fu, Hai Yao, Dongbo Li, Hangyuan Guo, and Cyril Rakovski. Optimal multi-stage arrhythmia classification approach. *Scientific Reports*, 10(1):2898, February 2020. doi: 10.1038/s41598-020-59821-7. URL https://doi.org/10.1038/s41598-020-59821-7.

Jianwei Zheng, Hangyuan Guo, and Huimin Chu. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0.0). *PhysioNet*, August 2022. doi: 10.13026/wgex-er52. URL https://doi.org/10.13026/wgex-er52. Version 1.0.0.

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*, pp. 665–674, 2017.

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.

APPENDIX

A MEET–UNION–JOIN GEOMETRY: DEFINITIONS AND PROOFS

This appendix formalizes the tangent-space model used in the main text. In each local chart, the component submanifolds are represented by tangent subspaces in common ambient coordinates, where the meet, union, and join relations hold pointwise. Applied throughout the tangent bundle, these fiberwise relations provide a global characterization of the nonlinear geometry without assuming that the manifolds or networks are globally linear. Whenever $\mathrm{dist}(s, T)$ is interpreted as a reconstruction score, we assume the ideal nearest-point projection prescribed by P2. For a generic autoencoder, its output need not be the nearest point in its learned range, so geometric distance and reconstruction error are kept distinct.

A.1 AXIOMATIC DEPENDENCE OF THE RESULTS

Table 8: Dependence of the main results and constructions on the PoS postulates. A checkmark denotes direct use, while an open circle denotes that the postulate is required only to transfer a geometric statement to its neural-network interpretation. Blank entries rely only on the stated hypotheses and standard geometry.

Result or construction	P1	P2	P3	P4
Score ordering, union optimality, and linear obstruction		○		
Join blindness and meet preference as PoS failures	✓	○		
Dynamic Push and Pull construction	✓	✓	✓	
Push and Pull geometric guarantees		✓		
Nested Manifold Carving and the nested-range relation	✓	✓	✓	✓
Increasing relative separation of nominal components	✓			✓
Movement of off-union samples toward nominal components		✓	✓	✓

Table 8 makes the axiomatic dependence explicit. A checkmark denotes a postulate used directly in the construction or result; an open circle denotes a postulate needed to transfer a geometric statement to the corresponding neural-network interpretation. Throughout the paper, every use of a PoS postulate is stated explicitly; an unmarked step uses only the displayed hypotheses and standard metric, set, or linear geometry. All definitions and results below apply to any input or latent representation space. We suppress the level index throughout; at level ℓ, the symbols S_i, $\mathcal{W}$, $\mathcal{U}$, $\mathcal{V}$, and T are understood as their level specific counterparts.

A.2 MEET, UNION, AND JOIN

Let $S_1, \ldots, S_M$ be linear subspaces of a common finite-dimensional Euclidean space $\mathbb{R}^n$. In the manifold interpretation, they are tangent spaces expressed in a common local coordinate system.

Definition 2 (Meet, union, and join)**.** *Define*

$$\mathcal{W} := \bigcap_{i=1}^{M} S_i, \qquad \mathcal{U} := \bigcup_{i=1}^{M} S_i, \qquad \mathcal{V} := \sum_{i=1}^{M} S_i. \tag{8}$$

Then $\mathcal{W} \subseteq \mathcal{U} \subseteq \mathcal{V}$. The union $\mathcal{U}$ is generally not a subspace, whereas $\mathcal{W}$ and $\mathcal{V}$ are subspaces. This chain concerns the three reference sets only and imposes no ordering on an arbitrary learned range T.

Definition 3 (Geometric and reconstruction scores)**.** *For a nonempty set $T \subseteq \mathbb{R}^n$, define the geometric score*

$$a_{\mathrm{geo}}(s; T) := \mathrm{dist}(s, T) = \inf_{t \in T} \|s - t\|_2.$$

For an autoencoder P_θ with learned range $T_\theta := \mathrm{range}(P_\theta)$, define its reconstruction score

$$a_{\mathrm{rec}}(s; \theta) := \|s - P_\theta(s)\|_2.$$

Since $P_\theta(s) \in T_\theta$,

$$a_{\mathrm{geo}}(s; T_\theta) \leq a_{\mathrm{rec}}(s; \theta).$$

Equality holds if and only if $P_\theta(s)$ is a nearest point to s in T_θ. Under P2, equality therefore holds wherever the learned nonlinear orthogonal projector selects a nearest point in its range. Such a nearest point need not be unique; conditions ensuring uniqueness are outside the scope of this paper.

A.2.1 Linear autoencoders are non-compact learners of subspace unions

The compact and non-compact terminology follows the original PoS formulation (Yamac et al., 2026). When $\mathcal{U} \neq \mathcal{V}$, a linear autoencoder whose range contains all nominal components cannot compact its learned range to their union $\mathcal{U}$, because linearity forces the range to contain their entire join $\mathcal{V}$. The following proposition formalizes the contrast between linear and ReLU networks observed in the PoS toy examples and complements the off-support reconstruction analysis of Bouman & Heskes (2025).

Proposition 4 (Linear non-compaction)**.** *Let $P = f_D \circ f_E : \mathbb{R}^n \to \mathbb{R}^n$ be a linear autoencoder map with learned range $T := \mathrm{range}(P)$. If $\mathcal{U} \subseteq T$, then*

$$\mathcal{V} = \mathrm{span}(\mathcal{U}) \subseteq T. \tag{9}$$

Consequently, if $T \subseteq \mathcal{V}$, then $T = \mathcal{V}$. Whenever $\mathcal{U} \neq \mathcal{V}$, no linear autoencoder whose range contains $\mathcal{U}$ can have learned range $T = \mathcal{U}$.

Proof. The range T of a linear map is a subspace. Since T contains $\mathcal{U}$, it must also contain the smallest linear subspace containing $\mathcal{U}$, namely $\mathrm{span}(\mathcal{U}) = \sum_i S_i = \mathcal{V}$. The remaining statements follow immediately. □

Thus, a linear autoencoder whose range covers $\mathcal{U}$ necessarily also contains the off-union directions $\mathcal{V} \setminus \mathcal{U}$. Under P2, these directions become zero-score blind regions. Exact carving of a genuine union $\mathcal{U} \neq \mathcal{V}$ therefore requires nonlinearity somewhere in the complete reconstruction map.

Remark 5 (Architectural consequence for nested carving)**.** *The obstruction concerns the complete reconstruction map $P = f_D \circ f_E$, not the encoder alone. At any level ℓ where $\mathcal{U}^{(\ell-1)} \neq \mathcal{V}^{(\ell-1)}$, a linear stage whose pull anchors enforce $\mathcal{U}^{(\ell-1)} \subseteq T^{(\ell)}$ must also satisfy $\mathcal{V}^{(\ell-1)} \subseteq T^{(\ell)}$. It may remove directions outside the join, but it cannot exclude the interpolating directions $\mathcal{V}^{(\ell-1)} \setminus \mathcal{U}^{(\ell-1)}$. Exact union carving therefore requires the complete stage map to be nonlinear. This requirement does not prescribe where activations or gating must be placed, and it does not apply when the nominal union is already a subspace.*

Proposition 6 (Geometric score sandwich)**.** *For every $s \in \mathbb{R}^n$,*

$$\mathrm{dist}(s, \mathcal{V}) \leq \mathrm{dist}(s, \mathcal{U}) = \min_{1 \leq i \leq M} \mathrm{dist}(s, S_i) \leq \mathrm{dist}(s, \mathcal{W}). \tag{10}$$

Proof. For sets $A \subseteq B$, distance reverses inclusion: $\mathrm{dist}(s, B) \leq \mathrm{dist}(s, A)$. Applying this to $\mathcal{W} \subseteq \mathcal{U} \subseteq \mathcal{V}$ gives the two inequalities. Since $\mathcal{U} = \bigcup_{i=1}^{M} S_i$ is a finite union, its distance is the minimum of the component distances. □

Under P2, this ordering transfers to reconstruction scores when the corresponding networks realize nearest-point projections. It does not hold automatically for generic autoencoders.

A.3 Optimality of the nominal union among geometrically faithful ranges

Assume that p_{nom} is supported on the nominal union $\mathcal{U}$. Since $\mathcal{U}$ is a finite union of closed subspaces, it is closed. Define the closed candidate ranges that geometrically preserve the complete nominal support by

$$\mathfrak{T}_+ := \{T \subseteq \mathbb{R}^n : \mathcal{U} \subseteq T,\ T \text{ is closed}\}. \tag{11}$$

For every $T \in \mathfrak{T}_+$, $\mathrm{dist}(s_n, T) = 0$ for all $s_n \in \mathcal{U}$. Under P2, this corresponds to perfect nominal reconstruction.

For an anomaly distribution p_a for which the following expectations exist, define the expected geometric score gap

$$\Delta(T) := \mathbb{E}_{s_a \sim p_a}[\mathrm{dist}(s_a, T)] - \mathbb{E}_{s_n \sim p_{\mathrm{nom}}}[\mathrm{dist}(s_n, T)]. \tag{12}$$

The second term vanishes for every $T \in \mathfrak{T}_+$, so $\Delta(T)$ equals the expected anomaly distance from T. The question is therefore which geometrically nominal-faithful range keeps anomalies farthest away. This analysis transfers to reconstruction error only when the network realizes the nearest-point projection prescribed by P2. Moreover, $\Delta(T)$ is an expected score gap and does not by itself determine AUC.

Lemma 7 (Nominal union optimality)**.** *Assume that the nominal distribution is supported on $\mathcal{U}$, that is, $p_{\mathrm{nom}}(\mathcal{U}) = 1$. For every $T \in \mathfrak{T}_+$ and every $s \in \mathbb{R}^n$,*

$$\mathrm{dist}(s, T) \leq \mathrm{dist}(s, \mathcal{U}). \tag{13}$$

Consequently,

$$\max_{T \in \mathfrak{T}_+} \Delta(T) = \Delta(\mathcal{U}), \tag{14}$$

so the nominal union $\mathcal{U}$ is an optimal geometrically faithful range.

Proof. Since $\mathcal{U} \subseteq T$, set inclusion gives $\mathrm{dist}(s, T) \leq \mathrm{dist}(s, \mathcal{U})$. Moreover, $p_{\mathrm{nom}}(\mathcal{U}) = 1$ implies that both nominal distance terms in $\Delta(T)$ and $\Delta(\mathcal{U})$ vanish. Taking expectations over p_a gives $\Delta(T) \leq \Delta(\mathcal{U})$. Since $\mathcal{U}$ is closed, $\mathcal{U} \in \mathfrak{T}_+$, so the upper bound is attained at $T = \mathcal{U}$. □

Remark 8. *The lemma identifies $\mathcal{U}$ as an optimal geometric target, independently of the particular model used to represent the learned range. Nominal reconstruction fidelity preserves $\mathcal{U}$, while the push and pull terms use controlled perturbations of nominal samples to discourage excess range. Together, these signals encourage the learned range T to approach the compact union $\mathcal{U}$ without requiring observed anomalies.*

The optimum need not be unique for a fixed anomaly distribution. A larger range $T \supsetneq \mathcal{U}$ may attain the same expected score if $\mathrm{dist}(s_a, T) = \mathrm{dist}(s_a, \mathcal{U})$ for p_a-almost every anomaly.

If the exact support assumption is relaxed and, for some $\varepsilon \geq 0$,

$$\mathbb{E}_{s_n \sim p_{\mathrm{nom}}} [\mathrm{dist}(s_n, \mathcal{U})] \leq \varepsilon, \tag{15}$$

then every $T \in \mathfrak{T}_+$ satisfies

$$\Delta(T) \leq \Delta(\mathcal{U}) + \varepsilon. \tag{16}$$

Thus, when nominal observations concentrate near $\mathcal{U}$, the nominal union remains optimal up to ε in expected geometric score gap.

A.4 Join blindness

Corollary 9 (Join blindness)**.** *Suppose join collapse occurs, so the learned operator has range $T_\theta := \mathrm{range}(P_\theta) = \mathcal{V}$. Under the ideal geometric score, every sample $s \in \mathcal{V}$ satisfies $\sigma_{\mathcal{V}}(s) = 0$. Hence nominal data supported on $\mathcal{U}$ and any anomaly distribution supported on $\mathcal{V}$ receive identical constant scores. Under the standard half-credit convention for ties, the corresponding AUC is $1/2$.*

Proof. The distance from any member of $\mathcal{V}$ to $\mathcal{V}$ is zero. Therefore every nominal and anomalous score is zero. □

A.5 Meet preference and fidelity loss

Recall that the meet

$$\mathcal{W} := \bigcap_{i=1}^{M} S_i$$

is the subspace shared by all component tangent spaces. Limited capacity need not recover exactly $\mathcal{W}$, but may preferentially retain its shared directions while discarding component specific directions. We call the limiting case $T = \mathcal{W}$ meet collapse. The following proposition gives a sufficient condition for this preference under rank constrained MSE learning.

Proposition 10 (Shared direction preference under balanced compression)**.** *Let nominal data be sampled uniformly from M components, with $\mathbb{E}[s \mid i] = 0$. Suppose $\dim \mathcal{W} > 0$, and write*

$$S_i = \mathcal{W} \oplus C_i, \qquad C_i := S_i \cap \mathcal{W}^{\perp}.$$

Assume that the C_i are mutually orthogonal and that

$$\mathrm{Cov}(s \mid i) = \sigma_{\mathcal{W}}^2 P_{\mathcal{W}} + \sigma_C^2 P_{C_i}. \tag{17}$$

Then the pooled covariance is

$$\Sigma = \sigma_{\mathcal{W}}^2 P_{\mathcal{W}} + \frac{\sigma_C^2}{M} \sum_{i=1}^{M} P_{C_i}. \tag{18}$$

If

$$\sigma_{\mathcal{W}}^2 > \frac{\sigma_C^2}{M}, \tag{19}$$

every direction in $\mathcal{W}$ has strictly larger covariance eigenvalue than every component specific direction. Consequently, a globally MSE optimal linear autoencoder constrained to rank r retains shared directions first. In particular, if $r = \dim \mathcal{W}$, its learned range is exactly $\mathcal{W}$.

Proof. Averaging the conditional covariances over the M components gives Eq. (18). Mutual orthogonality gives eigenvalue $\sigma_{\mathcal{W}}^2$ on $\mathcal{W}$ and σ_C^2/M on each C_i. Condition (19) therefore places all shared directions before all component specific directions in the covariance spectrum. By the linear autoencoder characterization of Baldi & Hornik (1989), the global MSE optimum projects onto the leading principal subspace. At rank $r = \dim \mathcal{W}$, condition (19) makes this subspace exactly $\mathcal{W}$, proving the result. □

Remark 11. *Exact meet collapse occurs at $r = \dim \mathcal{W}$. If $r < \dim \mathcal{W}$, the learned range is a proper subspace of $\mathcal{W}$. If $r > \dim \mathcal{W}$, it retains $\mathcal{W}$ together with selected component specific directions. Thus, the proposition establishes shared direction preference, not universal convergence to $\mathcal{W}$. Mutual orthogonality of the C_i is a simplifying sufficient assumption. For positive variances, condition (19) is equivalent to*

$$M > \frac{\sigma_C^2}{\sigma_{\mathcal{W}}^2},$$

so shared direction preference strengthens with the number of balanced components when $\sigma_{\mathcal{W}}^2$ and σ_C^2 are fixed.

Proposition 12 (Nominal distortion under meet collapse)**.** *Suppose meet collapse occurs, so that the learned range is $T = \mathcal{W}$. For every nominal sample $s \in S_i$,*

$$a_{\mathrm{geo}}(s; \mathcal{W}) = \mathrm{dist}(s, \mathcal{W}) = \left\| P_{S_i \cap \mathcal{W}^{\perp}} s \right\|_2. \tag{20}$$

Thus, every nominal sample with nonzero component specific tangent content outside $\mathcal{W}$ receives a positive geometric score. Under P2, this score equals its reconstruction error.

Proof. Since $\mathcal{W} \subseteq S_i$,

$$s = P_{\mathcal{W}} s + P_{S_i \cap \mathcal{W}^{\perp}} s.$$

The two terms are orthogonal, and $P_{\mathcal{W}} s$ is the nearest point to s in $\mathcal{W}$. Therefore,

$$\mathrm{dist}(s, \mathcal{W}) = \left\| s - P_{\mathcal{W}} s \right\|_2 = \left\| P_{S_i \cap \mathcal{W}^{\perp}} s \right\|_2,$$

proving Eq. (20). □

Remark 13. *Meet collapse guarantees loss of nominal fidelity, but it does not determine AUC without specifying the anomaly distribution. In particular, chance AUC requires additional distributional conditions, such as identical nominal and anomaly score distributions. The geometric prediction is therefore a loss of fidelity, not a universal claim of chance detection.*

Table 9: Toy verification of the learned geometries and their score consequences. The linear two dimensional model learns the join, the ReLU model approximates the compact union, and the one dimensional linear model exhibits rank deficient underlearning. The meet is trivial in this toy, so the final case illustrates nominal fidelity loss rather than exact meet collapse. Distances report the mean and standard deviation of reconstruction scores.

Learned geometry	$d_{\mathrm{nom,train}}$	$d_{\mathrm{nom,test}}$	d_{anom}	$\frac{d_{\mathrm{anom}}}{d_{\mathrm{nom,test}}}$	AUROC
Join (linear AE, 2D)	0.134 ± 0.11	0.143 ± 0.10	0.136 ± 0.12	0.95	0.456
Compact union (ReLU AE, 2D)	0.226 ± 0.12	0.244 ± 0.11	1.624 ± 0.56	6.64	1.000
Rank deficient underlearning (linear AE, 1D)	1.364 ± 0.51	1.386 ± 0.44	1.435 ± 0.90	1.04	0.422

Toy data verification: Table 9 reports reconstruction scores for the toy data illustrated in Figure 1. The join model reconstructs nominal and anomalous samples similarly, demonstrating the loss of selectivity shown in Figure 1(b). The rank deficient model assigns large errors to both groups, demonstrating the loss of nominal fidelity shown in Figure 1(d). Because the two nominal subspaces have a trivial meet, this case is not an instantiation of Proposition 10. Only the compact union model preserves low nominal error while assigning substantially larger scores to anomalies, as shown in Figure 1(c).

A.6 Proof of the separate pull and push guarantees

Proof. Since $P_\theta(\widetilde{s}_n) \in T$, the pull bound implies

$$\mathrm{dist}_d(s_n, T) \le d(P_\theta(\widetilde{s}_n), s_n) \le \varepsilon_{\mathrm{pull}}.$$

Moreover, $[z]_+ \ge z$ gives

$$\rho_{\theta,d}(\widetilde{s}_n) \ge \alpha\delta_e - \varepsilon_{\mathrm{push}}.$$

Under P2, $\rho_{\theta,d}(\widetilde{s}_n) = \mathrm{dist}_d(\widetilde{s}_n, T)$. The triangle inequality also gives

$$\delta_e \le \mathrm{dist}_d(\widetilde{s}_n, T) + d(P_\theta(\widetilde{s}_n), s_n) \le \mathrm{dist}_d(\widetilde{s}_n, T) + \varepsilon_{\mathrm{pull}}.$$

Together with nonnegativity, these inequalities prove Eq. (5).

Finally, if $\mathcal{U} \subseteq T$, then

$$\mathrm{dist}_d(\widetilde{s}_n, T) \le \mathrm{dist}_d(\widetilde{s}_n, \mathcal{U}) \le \delta_e + \eta(s_n).$$

Combining this upper bound with $\mathrm{dist}_d(\widetilde{s}_n, T) \ge \alpha\delta_e - \varepsilon_{\mathrm{push}}$ and dividing by $\delta_e > 0$ proves Eq. (6). □

B Detailed construction of nested manifold carving

This appendix gives the stagewise construction and capacity conditions behind Section 3.2. Following P4, nested carving applies dynamic push–pull recursively in learned latent spaces. Let $x^{(0)} = s$. At stage ℓ, the previously learned encoder–decoder pairs are frozen, and a new pair is trained on the current nominal representation $x^{(\ell-1)}$:

$$x^{(\ell)} = f_E^{(\ell)}(x^{(\ell-1)}), \qquad \widehat{x}^{(\ell-1)} = f_D^{(\ell)}(x^{(\ell)}). \tag{21}$$

Before training, the new latent dimension is chosen to preserve the current nominal union. In the local linearized model, write $\mathcal{U}^{(\ell-1)} = \bigcup_i S_i^{(\ell-1)}$ and let $f_E^{(\ell)} : \mathbb{R}^{n_{\ell-1}} \to \mathbb{R}^{n_\ell}$. As reviewed in PoS Yamac et al. (2026) and classical union-of-subspaces theory Blumensath & Davies (2009); Lu & Do (2008), a linear encoder is injective on $\mathcal{U}^{(\ell-1)}$ precisely when

$$\ker f_E^{(\ell)} \cap \big(S_i^{(\ell-1)} + S_j^{(\ell-1)}\big) = 0 \quad \text{for all } i, j. \tag{22}$$

For a finite family, this condition is generically feasible when

$$n_\ell \ge k_{\max}^{(\ell-1)} := \max_{i,j} \dim\big(S_i^{(\ell-1)} + S_j^{(\ell-1)}\big). \tag{23}$$

Below this width, some pairwise difference is necessarily annihilated, so distinct nominal points become indistinguishable and exact reconstruction of the union is impossible.

Equations (22)–(23) preserve nominal distinctions but do not themselves ensure compact-union carving. If the complete stage map is linear and faithful on a non-subspace union, Proposition A.2.1 forces its range to contain the corresponding join. Exact union carving therefore additionally requires nonlinearity and informative perturbations.

When Eq. (22) holds and

$$S_i^{(\ell)} := f_E^{(\ell)}(S_i^{(\ell-1)}),$$

the dimensions of all pairwise sums are preserved, giving $k_{\max}^{(\ell)} = k_{\max}^{(\ell-1)}$. Hence the capacity floor in Eq. (23) persists across faithful linearized levels. A simple sufficient architectural choice is $n_\ell \geq n_{\ell-1}$, for which a generic full-column-rank encoder has zero kernel. These are local linearized capacity conditions; nominal fidelity must still be learned by the pull objective.

Finally, a perturbation $e^{(\ell)}$ is applied to $x^{(\ell-1)}$, and Eq. (2) is used at level ℓ with

$$s_n = x^{(\ell-1)}, \qquad e = e^{(\ell)}, \qquad P_\theta = P_{\theta_\ell} := f_D^{(\ell)} \circ f_E^{(\ell)}. \tag{24}$$

Thus, the pull term preserves the nominal latent representation, while the push term discourages the new stage from reproducing its perturbed counterpart. After optimization, the new pair is frozen and the procedure is repeated at the next latent level. For compact notation, define the encoder and decoder stacks

$$F_\ell := f_E^{(\ell)} \circ \cdots \circ f_E^{(1)}, \qquad G_\ell := f_D^{(1)} \circ \cdots \circ f_D^{(\ell)},$$

with $F_0 = G_0 = I$. The complete reconstruction map through level ℓ is therefore

$$\overline{P}^{(\ell)} := G_{\ell-1} \circ P_{\theta_\ell} \circ F_{\ell-1} = G_\ell \circ F_\ell, \qquad \overline{T}^{(\ell)} := \operatorname{range}\big(\overline{P}^{(\ell)}\big), \tag{25}$$

where, $\overline{T}^{(\ell)}$ is the range of the level-ℓ reconstruction after decoding it back to the input space.

Range nesting requires the new stage to remain compatible with the representation inherited from the preceding encoder stack:

$$P_{\theta_\ell}\left(\operatorname{range}(F_{\ell-1})\right) \subseteq \operatorname{range}(F_{\ell-1}). \tag{26}$$

Indeed, Eq. (26) gives $\overline{T}^{(\ell)} \subseteq G_{\ell-1}\left(\operatorname{range}(F_{\ell-1})\right) = \overline{T}^{(\ell-1)}$. Nominal coverage is preserved if the preceding encoder–decoder stack and the new stage are faithful on $\mathcal{U}$, namely,

$$G_{\ell-1}(F_{\ell-1}(u)) = u, \qquad P_{\theta_\ell}(F_{\ell-1}(u)) = F_{\ell-1}(u), \qquad u \in \mathcal{U}.$$

These conditions imply $\overline{P}^{(\ell)}(u) = G_{\ell-1}\left(P_{\theta_\ell}(F_{\ell-1}(u))\right) = u$, and hence $\mathcal{U} \subseteq \overline{T}^{(\ell)}$. Applying compatibility and nominal fidelity at every level yields

$$\overline{T}^{(1)} \supseteq \overline{T}^{(2)} \supseteq \cdots \supseteq \overline{T}^{(L)} \supseteq \mathcal{U}. \tag{27}$$

C Detailed MNIST ablation study

This appendix provides the complete results supporting the controlled MNIST study in Section 4.1. We consider four nominal-digit sets and treat all remaining digits as anomalies. All models use the same ConvAE backbone, with Stage-0 latent dimension 9 and Stage-1 expanded dimension 60. The ordinary autoencoder objective uses MSE, whereas push–pull uses the Euclidean metric. For every stage ℓ, the anomaly score is always computed in the input space as

$$\rho^{(\ell)}(s) = \left\| s - \widehat{s}^{(\ell)} \right\|_2,$$

where $\widehat{s}^{(\ell)}$ is the complete reconstruction through that stage. Stage-0 and Stage-1 scores are reported separately; no latent-space or multistage score fusion is used. Results are averaged over seeds 42–46, and F1 uses the 95th percentile of the fit-split nominal scores as its threshold.

Per-stage results: Table 10 reports the mean and sample standard deviation of ROC-AUC and F1 at both stages for every nominal-digit set and training configuration.

When nested carving contributes: The benefit of Stage 1 depends on which anomalous directions remain unresolved after Stage 0. The clearest improvement occurs for the nominal set $\{0, 2, 4, 5, 6, 8, 9\}$, for which digit 1 is anomalous. Without masking or sparsity, AE+PP increases from 0.722 to 0.816 ROC-AUC and from 0.436 to 0.489 F1 between Stage 0 and Stage 1. This behavior is consistent with the geometry in Figure 3: digit 1 can remain close to the Stage-0 join while lying outside the nominal union, and therefore benefits from additional carving in the latent space.

When digit 1 is itself nominal, as in the set $\{1, 2, 5, 7, 8\}$, the same configuration changes only from 0.921 to 0.922 ROC-AUC and from 0.748 to 0.749 F1. Here, Stage 0 already separates the remaining anomalies effectively, leaving little unresolved off-union structure for Stage 1 to remove. Intermediate gains appear for the other two settings in which digit 1 is anomalous. With sparsity and no masking, Stage 1 raises ROC-AUC from 0.848 to 0.875 for nominal digits $\{2, 3, 6, 9\}$, and from 0.903 to 0.928 for nominal digits $\{4, 5, 7\}$. These results indicate that nesting is most useful when the preceding stage retains anomaly directions within its learned coarse range, rather than when the first projection has already resolved them.

Masked reconstruction versus push–pull: Masking alone provides no consistent carving advantage. For nominal digits $\{4, 5, 7\}$, the unmasked Stage-0 AE obtains 0.893 ROC-AUC and 0.780 F1, while masked reconstruction decreases these values to 0.881 and 0.766. In contrast, the unmasked AE+PP reaches 0.897 ROC-AUC and 0.792 F1 at Stage 0, and nested push–pull further increases them to 0.909 and 0.798 at Stage 1. The same general pattern appears across the other nominal-digit sets: push–pull improves over the corresponding AE objective more consistently than masking alone, and its gain can increase when applied recursively.

Masking can nevertheless complement push–pull when combined with an appropriate nested objective. For nominal digits $\{4, 5, 7\}$, masked AE+PP with sparsity reaches 0.929 ROC-AUC and 0.815 F1 at Stage 1, the best result for this setting. Thus, the results do not show that masking is intrinsically ineffective. Rather, masking by itself does not specify which perturbation directions should be excluded, whereas push–pull explicitly separates the reconstruction from the perturbed input while anchoring it to the nominal observation.

Complete structural ablation: Table 11 reorganizes the $\{4, 5, 7\}$ results to isolate the effects of hierarchy, push–pull, masking, and sparsity. A reconstruction-only nested module leaves performance essentially unchanged because the expanded latent autoencoder admits the identity map and receives no explicit carving signal. Push–pull supplies this signal, while sparsity and masking act as optional complementary regularizers. Their effects are not uniformly additive: sparsity strengthens nested push–pull in the $\{4, 5, 7\}$ experiment but can reduce performance for other nominal sets. The principal consistent gain therefore comes from the push–pull objective, with nesting contributing when Stage 0 retains off-union anomaly directions.

Table 10: **Per-stage MNIST anomaly-detection performance.** At each stage ℓ, the anomaly score is the input-space reconstruction error $d(s, \widehat{s}^{(\ell)})$, where $\widehat{s}^{(\ell)}$ is the complete reconstruction through that stage. No latent-space or multistage score fusion is used. Each entry reports the mean $\pm$ sample standard deviation over seeds 42–46. AE denotes the nominal reconstruction objective $\mathcal{L}_{\mathrm{AE}}$, while AE+PP additionally includes $\lambda_{\mathrm{PP}}\mathcal{L}_{\mathrm{PP}}$. The best result within each nominal-digit set and stage is shown in **bold**.

Normal digits	Aug.	Objective	Sparse	Stage 0		Stage 1	
				ROC-AUC	F1	ROC-AUC	F1
0,2,4,5,6,8,9	mask	AE	–	0.700 ± 0.015	0.401 ± 0.038	0.702 ± 0.017	0.402 ± 0.040
0,2,4,5,6,8,9	mask	AE	yes	0.685 ± 0.008	0.373 ± 0.045	0.688 ± 0.012	0.382 ± 0.038
0,2,4,5,6,8,9	mask	AE+PP	–	0.711 ± 0.011	0.427 ± 0.029	$\mathbf{0.822 \pm 0.016}$	0.493 ± 0.028
0,2,4,5,6,8,9	mask	AE+PP	yes	0.708 ± 0.020	0.400 ± 0.030	0.782 ± 0.024	0.472 ± 0.048
0,2,4,5,6,8,9	none	AE	–	0.704 ± 0.016	0.410 ± 0.032	0.710 ± 0.017	0.411 ± 0.035
0,2,4,5,6,8,9	none	AE	yes	0.698 ± 0.023	0.394 ± 0.043	0.704 ± 0.023	0.401 ± 0.042
0,2,4,5,6,8,9	none	AE+PP	–	$\mathbf{0.722 \pm 0.015}$	$\mathbf{0.436 \pm 0.050}$	0.816 ± 0.017	0.489 ± 0.025

Continued on next page

Table 10 – continued

Normal digits	Aug.	Objective	Sparse	Stage 0 ROC-AUC	Stage 0 F1	Stage 1 ROC-AUC	Stage 1 F1
0,2,4,5,6,8,9	none	AE+PP	yes	0.711 ± 0.013	0.417 ± 0.049	0.782 ± 0.038	$\mathbf{0.494 \pm 0.058}$
1,2,5,7,8	mask	AE	–	0.900 ± 0.002	0.664 ± 0.028	0.899 ± 0.002	0.660 ± 0.031
1,2,5,7,8	mask	AE	yes	0.894 ± 0.012	0.653 ± 0.049	0.895 ± 0.011	0.657 ± 0.048
1,2,5,7,8	mask	AE+PP	–	0.919 ± 0.002	0.731 ± 0.012	0.918 ± 0.004	0.724 ± 0.026
1,2,5,7,8	mask	AE+PP	yes	0.919 ± 0.005	0.710 ± 0.038	0.918 ± 0.006	0.714 ± 0.033
1,2,5,7,8	none	AE	–	0.910 ± 0.008	0.689 ± 0.036	0.911 ± 0.008	0.690 ± 0.036
1,2,5,7,8	none	AE	yes	0.903 ± 0.006	0.672 ± 0.041	0.904 ± 0.007	0.674 ± 0.040
1,2,5,7,8	none	AE+PP	–	$\mathbf{0.921 \pm 0.004}$	$\mathbf{0.748 \pm 0.025}$	$\mathbf{0.922 \pm 0.003}$	$\mathbf{0.749 \pm 0.025}$
1,2,5,7,8	none	AE+PP	yes	0.919 ± 0.003	0.727 ± 0.021	0.919 ± 0.004	0.738 ± 0.026
2,3,6,9	mask	AE	–	0.814 ± 0.004	0.557 ± 0.023	0.815 ± 0.003	0.561 ± 0.017
2,3,6,9	mask	AE	yes	0.823 ± 0.003	0.564 ± 0.036	0.820 ± 0.004	0.572 ± 0.036
2,3,6,9	mask	AE+PP	–	0.845 ± 0.010	0.613 ± 0.025	0.847 ± 0.006	0.633 ± 0.024
2,3,6,9	mask	AE+PP	yes	0.841 ± 0.006	0.600 ± 0.034	0.869 ± 0.008	0.646 ± 0.055
2,3,6,9	none	AE	–	0.829 ± 0.009	0.601 ± 0.031	0.828 ± 0.010	0.603 ± 0.029
2,3,6,9	none	AE	yes	0.830 ± 0.009	0.580 ± 0.033	0.830 ± 0.011	0.584 ± 0.033
2,3,6,9	none	AE+PP	–	0.842 ± 0.005	0.614 ± 0.035	0.850 ± 0.008	0.626 ± 0.022
2,3,6,9	none	AE+PP	yes	$\mathbf{0.848 \pm 0.003}$	$\mathbf{0.626 \pm 0.027}$	$\mathbf{0.875 \pm 0.010}$	$\mathbf{0.665 \pm 0.036}$
4,5,7	mask	AE	–	0.881 ± 0.005	0.766 ± 0.017	0.882 ± 0.003	0.766 ± 0.018
4,5,7	mask	AE	yes	0.884 ± 0.005	0.769 ± 0.009	0.886 ± 0.005	0.770 ± 0.008
4,5,7	mask	AE+PP	–	0.902 ± 0.003	0.790 ± 0.016	0.913 ± 0.002	0.794 ± 0.019
4,5,7	mask	AE+PP	yes	0.899 ± 0.006	0.788 ± 0.018	$\mathbf{0.929 \pm 0.007}$	$\mathbf{0.815 \pm 0.006}$
4,5,7	none	AE	–	0.893 ± 0.008	0.780 ± 0.012	0.893 ± 0.008	0.780 ± 0.011
4,5,7	none	AE	yes	0.890 ± 0.011	0.772 ± 0.012	0.889 ± 0.011	0.772 ± 0.009
4,5,7	none	AE+PP	–	0.897 ± 0.004	0.792 ± 0.008	0.909 ± 0.008	0.798 ± 0.010
4,5,7	none	AE+PP	yes	$\mathbf{0.903 \pm 0.004}$	$\mathbf{0.793 \pm 0.011}$	0.928 ± 0.004	0.808 ± 0.029

Table 11: **Full factorial ablation on MNIST with nominal digits** $4, 5, 7$**.** The first row reproduces the two-dimensional autoencoder baseline of Bouman & Heskes (2025). Rows without a hierarchy tick use the Stage-0 reconstruction score; ticked rows use the complete Stage-1 reconstruction decoded to the input space. Values are means over different seeds. Best results are shown in **bold** and second-best results are underlined.

Normal digits	Hierarchy level	Push pull	Mask	Sparsity	F1	ROC-AUC	PR-AUC
Baseline (4,5,7)					0.693	0.866	0.936
4,5,7					0.780	0.893	0.962
4,5,7	✓				0.780	0.893	0.962
4,5,7				✓	0.772	0.890	0.960
4,5,7	✓			✓	0.772	0.889	0.960
4,5,7		✓			0.792	0.897	0.963
4,5,7	✓	✓			0.798	0.909	0.967
4,5,7		✓		✓	0.793	0.903	0.965
4,5,7	✓	✓		✓	<u>0.808</u>	<u>0.928</u>	<u>0.972</u>
4,5,7			✓		0.766	0.881	0.957
4,5,7	✓		✓		0.766	0.882	0.957
4,5,7			✓	✓	0.769	0.884	0.958
4,5,7	✓		✓	✓	0.770	0.886	0.959
4,5,7		✓	✓		0.790	0.902	0.965
4,5,7	✓	✓	✓		0.794	0.913	0.968
4,5,7		✓	✓	✓	0.788	0.899	0.964
4,5,7	✓	✓	✓	✓	**0.815**	**0.929**	**0.973**

Denoising versus Push and Pull: A denoising autoencoder (DAE) receives a perturbed nominal sample $\widetilde{s}_n = s_n + e$ and learns to reconstruct its clean counterpart s_n. This corresponds to retaining only the Pull term. Indeed, setting $\alpha = 0$ gives

$$\mathcal{L}_{\mathrm{PP}} = d(\widehat{s}_n, s_n) + \left[-d(\widehat{s}_n, \widetilde{s}_n)\right]_+ = d(\widehat{s}_n, s_n),$$

Table 12: Matched comparison of reconstruction, denoising, and Push and Pull on MNIST with nominal digits $4, 5, 7$. Masking and sparsity are disabled to isolate the training objective. Stage 1 uses the complete nested reconstruction decoded to input space, without score fusion.

Objective	Stage	F1	ROC-AUC	PR-AUC
AE	0	0.780	0.893	0.962
AE	1	0.780	0.893	0.962
DAE	0	0.783	0.895	0.962
DAE	1	0.783	0.895	0.962
AE with Push and Pull	0	0.792	0.897	0.963
AE with Push and Pull	1	**0.798**	**0.909**	**0.967**

which is the DAE objective, up to whether the reconstruction discrepancy is implemented as Euclidean distance or its square. The comparison in Table 12 therefore isolates the contribution of the positive Push margin from ordinary denoising. DAE changes the Stage 0 results only marginally relative to AE, from 0.780 to 0.783 F1 and from 0.893 to 0.895 ROC-AUC, and its Stage 1 results remain unchanged. In contrast, Push and Pull reaches 0.792 F1 and 0.897 ROC-AUC at Stage 0, followed by 0.798 F1 and 0.909 ROC-AUC after nested carving. This behavior agrees with Lemma 1 and Figure 2. In the ideal case $s_n \in \mathcal{U}$, Pull can recover s_n, and for $\alpha \leq 1$ the Push hinge is inactive at that optimum, although it can still supply a gradient while the margin is violated during learning. For realistic nominal data lying near rather than exactly on $\mathcal{U}$, Pull alone only attracts the reconstruction toward s_n, whereas a positive Push margin additionally requires separation from $\widetilde{s}_n$. The table consequently indicates that exposure to perturbations through denoising is insufficient to explain the observed gain; the explicit separation constraint contributes additional carving, particularly when it is applied at the nested stage.

C.1 Geometric contribution of nested latent carving

Aggregate detection metrics show whether Stage 1 improves performance, but not which geometric cases produce that improvement. We therefore examine the image and latent reconstruction errors separately for each digit. This provides a direct empirical test of the geometry proposed in Figure 3 and of the representation dynamics described in Section 4.2.

Let T_{in} denote the frozen Stage 0 encoder, let the nested module map x to $\widehat{x}$, and let T_{out} denote the complete decoder. For an input image s,

$$x = T_{\text{in}}(s), \qquad \widehat{x} = T_{\text{nested}}(x), \qquad \widehat{s} = T_{\text{out}}(\widehat{x}).$$

We measure the change made by the complete model in the image domain and the change made specifically by Stage 1 in the Stage 0 latent space:

$$e^{\text{rec}}(s) = \frac{1}{|s|} \|\widehat{s} - s\|_2^2, \qquad \text{image reconstruction error,} \tag{28}$$

$$e^{\text{sub}}(s) = \frac{1}{|x|} \|\widehat{x} - x\|_2^2, \qquad \text{latent reconstruction error.} \tag{29}$$

Each quantity is computed per image by averaging over its elements. No averaging across images or batches is used. The latent error is introduced only for this diagnostic analysis. The principal results in the main paper continue to use the complete input space reconstruction score without score fusion.

The two errors distinguish the geometric cases illustrated in Figure 3. If the anomalous component of s lies largely outside the coarse Stage 0 range $\mathcal{V}$, Stage 0 removes it before x is formed. The resulting latent representation can therefore already resemble a nominal representation, giving a large image error but a small Stage 1 latent error. Digit 8 is the clearest example of this case: its anomalous direction lies predominantly outside the learned Stage 0 range, and its retained representation becomes close to that of nominal digit 5. By contrast, an anomaly lying inside $\mathcal{V} \setminus \mathcal{U}$ can be reconstructed well at Stage 0 and remain difficult to detect in the image domain. Its retained off-union direction is still present in x, however, allowing Stage 1 to carve it and produce a large discrepancy between x and $\widehat{x}$. Figure 3 predicts digit 1 as the representative case.

Table 13: Per-image flag decomposition of the two-channel or-rule. Entries are fractions of test images, with thresholds calibrated on held-out nominal images to a joint 5% false-positive rate ($q \approx 97$). Results are mean ± standard deviation over 12 runs. Nominal digits are the negative class; hence, their detection share is undefined.

		Channel fires		Disjoint outcomes			
Digit	Role	Recon.	Subsp.	Recon. only	Subsp. only	Neither	Subsp.-only share of det.
0	anomaly	0.952±0.044	0.325±0.113	0.632±0.127	0.006±0.008	0.042±0.038	0.6%
1	anomaly	**0.123±0.094**	**0.576±0.137**	0.028±0.015	**0.481±0.101**	0.396±0.126	**80.8%**
2	anomaly	0.965±0.008	0.444±0.098	0.531±0.099	0.010±0.005	0.024±0.005	1.1%
3	anomaly	0.827±0.033	0.129±0.077	0.705±0.093	0.008±0.004	0.165±0.030	0.9%
6	anomaly	0.581±0.062	0.388±0.080	0.249±0.081	0.056±0.031	0.364±0.050	8.8%
8	anomaly	0.876±0.031	0.124±0.105	0.758±0.106	0.006±0.008	0.117±0.032	0.7%
9	anomaly	0.101±0.014	0.049±0.026	0.081±0.020	0.030±0.017	0.870±0.023	21.6%
4	nominal	0.017±0.004	0.026±0.008	0.012±0.003	0.021±0.008	0.962±0.009	—
5	nominal	0.050±0.011	0.049±0.013	0.034±0.009	0.033±0.007	0.917±0.015	—
7	nominal	0.023±0.005	0.034±0.008	0.016±0.002	0.027±0.006	0.950±0.007	—

To quantify this complementarity, we calibrate one threshold for each score using a held out set $\mathcal{D}_{\text{fit}}$ of $n_{\text{fit}} = 435$ nominal images that are used neither for training nor for evaluation. For $c \in \{\text{rec}, \text{sub}\}$, let $t_c(q)$ be the qth percentile of score c on $\mathcal{D}_{\text{fit}}$. Define

$$A(s) = [e^{\text{rec}}(s) > t_{\text{rec}}(q)], \qquad B(s) = \left[e^{\text{sub}}(s) > t_{\text{sub}}(q)\right].$$

For this auxiliary analysis, an image is flagged when either channel fires:

$$\text{flag}(s) = A(s) \vee B(s). \tag{30}$$

Because Eq. (30) combines two tests, assigning the 95th percentile independently to both channels would produce a joint false positive rate above 5%. We therefore select a common percentile $q^\star$ on $\mathcal{D}_{\text{fit}}$ such that the combined rule has an approximately 5% false positive rate. Across the evaluated runs, $q^\star \in [97.0, 97.5]$, producing realized rates between 0.046 and 0.051.

For each digit, Table 13 reports the marginal firing probabilities $\Pr(A)$ and $\Pr(B)$, together with the disjoint outcomes

$$A \wedge \neg B, \qquad \neg A \wedge B, \qquad \neg A \wedge \neg B.$$

The unreported intersection can be recovered as

$$\Pr(A \wedge B) = \Pr(A) - \Pr(A \wedge \neg B) = \Pr(B) - \Pr(\neg A \wedge B).$$

We additionally report

$$r_{\text{latent}} = \frac{\Pr(\neg A \wedge B)}{\Pr(A \vee B)}, \tag{31}$$

which measures the fraction of detected samples that would be lost if only the image reconstruction score were retained.

The per digit decomposition closely matches the proposed geometry. For digit 8, the image channel fires for 87.6% of samples, whereas the latent channel fires for only 12.4%, and only 0.6% of its detections depend exclusively on the latent score. Thus, most of the anomalous information in digit 8 is removed before Stage 1, leaving a Stage 0 representation that the nested module changes little. This supports the first case in Figure 3: an anomaly outside the coarse range is detected through its large image reconstruction error, but its removed directions are unavailable to deeper latent processing.

Digit 1 exhibits the opposite behavior. The image channel fires for only 12.3% of its samples, consistent with digit 1 lying close to the coarse range $\mathcal{V}$ and therefore remaining a Stage 0 blind case. The latent channel fires for 57.6%, with 48.1% of all digit 1 samples detected exclusively in the latent space. Consequently, 80.8% of the detected digit 1 samples would be missed without the Stage 1 latent score. This result supports the second geometric case: Stage 0 retains the digit 1 direction inside its coarse range, while Stage 1 differentiates it from the compact nominal union.

The reconstructions in Figure 5 provide complementary visual evidence. After nested carving, digit 1 no longer preserves its original structure and frequently becomes 7-like, with some reconstructions

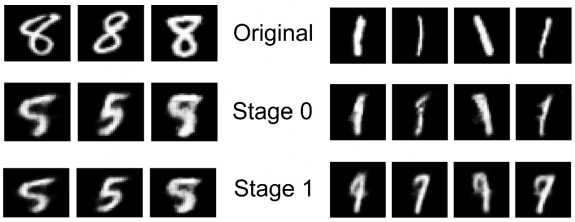

Figure 5: Representative reconstructions at successive stages. Anomalous directions removed at Stage 0 cannot be recovered from the Stage 0 latent representation. Digit 1, which remains inside the coarse Stage 0 range, is modified by nested carving and frequently becomes 7-like, consistent with projection toward a nearby nominal component.

remaining visually ambiguous between 7 and 9. The 7-like cases are consistent with the predicted representation dynamics: as the learned range becomes more compact, an off-union anomaly is moved toward a nearby nominal component. The experiment therefore does more than show an aggregate benefit from nesting. It identifies the anomalies removed before the latent stage, the anomalies exposed by latent carving, and the loss of anomalous information after projection. This agreement between the predicted geometry, the latent discrepancies, and the reconstructed images demonstrates the explanatory value of the PoS based geometric framework.

D METRICS

The following metrics in this section are mainly introduced to provide additional quantitative information regarding the claims that have been asserted throughout the paper. It should be noted that all introduced metrics are always taken between x and $\hat{x}$ which are the outputs of the frozen encoder and the outputs of the subspace, namely the network in between the frozen encoder decoder pair, respectively. Before any metric is computed, both spaces are standardised per dimension using the mean and the standard deviation of the normal x samples (with $\varepsilon = 10^{-8}$ added to the standard deviation and this value is used everywhere), so that the same map is applied to x and $\hat{x}$. All metrics are computed in the full latent space and since the results are a comparison of the results taken from two latent spaces having the same dimension, their difference is calculated using the relative change, for any scalar quantity Q, as following,

$$\Delta Q \;=\; 100 \cdot \frac{Q_{\hat{x}} - Q_x}{Q_x + \varepsilon} \quad [\%]$$

Additionally, there exists some commonly used quantities in all the metrics, namely the center of the normal clusters and their radii. Let $z \in \{x, \hat{x}\}$ index the two spaces mentioned above, and write z_i for the representation of point i in that space. The normal classes are $c \in \mathcal{C}$ with $K = |\mathcal{C}|$ and index sets $\mathcal{H}_c$, which are the same in both spaces, and the anomalies are $a \in \mathcal{A}$. The anomaly classes are $b \in \mathcal{B}$ with $M = |\mathcal{B}|$ and index sets $\mathcal{A}_b$, again the same in both spaces. For each normal class $c \in \mathcal{C}$, we, first, get the center points and the radii of the formed clusters in the latent space according to their labels. The centers of the normal clusters, denoted as μ_c^z, are calculated by taking the means of these points, and the radii, denoted as R_c^z, are calculated by taking the distance of each point in the class to the center and averaging over the class, i.e.,

$$\mu_c^z = \frac{1}{|\mathcal{H}_c|} \sum_{i \in \mathcal{H}_c} z_i, \qquad R_c^z = \frac{1}{|\mathcal{H}_c|} \sum_{i \in \mathcal{H}_c} \left\| z_i - \mu_c^z \right\| \tag{32}$$

where $\|\cdot\| = \|\cdot\|_2$ for all cases.

D.1 SEPARATION RATIO

To quantify the possible separation in between the normal clusters, we have used a metric we call the Separation Ratio, S^z. This metric is calculated by taking the means of the inter-distances of the normal clusters $\bar{D}^z_{\text{inter}}$ and dividing the result by the mean of the radii of all normal clusters $\bar{R}^z$. The relevant equations are as follows

Table 14: **Do the normal classes separate from one another?** Every column is a relative change from x to $\hat{x}$. $\bar{D}_{\text{inter}}$ is how far apart the classes sit; $\bar{S}_{\text{intra}}$ is how wide the classes are; $S = \bar{D}_{\text{inter}}/\bar{S}_{\text{intra}}$ combines the two into a single separation score, so it can rise either because the classes move apart or because they contract in place. Under PushPull the rise in S is driven by $\bar{S}_{\text{intra}}$ falling by roughly 8–10% while $\bar{D}_{\text{inter}}$ also falls, only more slowly. MSE leaves all three unchanged. The arrow marks the direction of improvement, the best value is **bold** and the second best is underlined.

normal set	mask	loss	sparsity	$\Delta S \uparrow$ (%)	$\Delta\bar{D}_{\text{inter}}$ (%)	$\Delta\bar{S}_{\text{intra}}$ (%)
$\{4,5,7\}$	none	MSE	–	+0.1	+0.0	-0.0
	none	MSE	yes	-0.2	-0.1	+0.1
	mask	MSE	–	-0.1	+0.0	+0.1
	mask	MSE	yes	-0.8	-0.9	-0.1
	none	PP	–	+7.6	-3.1	-10.0
	none	PP	yes	+2.6	-0.4	-3.0
	mask	PP	–	**+7.7**	-2.0	-9.0
	mask	PP	yes	+3.5	-0.9	-4.3
$\{2,3,6,9\}$	none	MSE	–	+0.2	+0.0	-0.2
	none	MSE	yes	-0.4	+0.4	+0.8
	mask	MSE	–	-0.3	-0.2	+0.1
	mask	MSE	yes	-0.8	-1.1	-0.3
	none	PP	–	+5.2	-4.0	-8.7
	none	PP	yes	+1.9	-2.8	-4.6
	mask	PP	–	**+7.2**	-1.9	-8.5
	mask	PP	yes	+1.8	-1.2	-2.9
$\{1,2,5,7,8\}$	none	MSE	–	-0.1	+0.2	+0.3
	none	MSE	yes	+0.0	+0.8	+0.8
	mask	MSE	–	+0.1	-0.2	-0.4
	mask	MSE	yes	-0.1	-1.6	-1.5
	none	PP	–	**+5.4**	-2.9	-7.9
	none	PP	yes	+1.9	-0.9	-2.7
	mask	PP	–	+5.3	-2.7	-7.6
	mask	PP	yes	+1.4	-0.1	-1.5
$\{0,2,4,5,6,8,9\}$	none	MSE	–	+0.3	-0.0	-0.3
	none	MSE	yes	+0.7	-0.0	-0.7
	mask	MSE	–	+0.1	-0.0	-0.1
	mask	MSE	yes	+0.1	-1.3	-1.5
	none	PP	–	+6.5	-2.0	-8.0
	none	PP	yes	+1.6	-0.7	-2.3
	mask	PP	–	**+7.3**	-1.7	-8.3
	mask	PP	yes	+0.7	-0.4	-1.1

$$\bar{D}^z_{\text{inter}} = \binom{K}{2}^{-1} \sum_{\substack{c,c' \in \mathcal{C} \\ c<c'}} \left\|\mu^z_c - \mu^z_{c'}\right\|, \qquad \bar{R}^z = \frac{1}{K}\sum_{c\in\mathcal{C}} R^z_c, \qquad S^z = \frac{\bar{D}^z_{\text{inter}}}{\bar{R}^z + \varepsilon}, \tag{33}$$

where $\varepsilon = 10^{-8}$. The division by $\bar{R}^z$ makes S^z invariant to a uniform rescaling of the latent space. What it does *not* do is isolate separation from compactness. S^z rises whenever $\bar{D}^z_{\text{inter}}$ grows relative to $\bar{R}^z$, which happens both when the normal classes move apart and when they contract in place. So we report $\bar{D}^z_{\text{inter}}$ and $\bar{R}^z$ separately in the results, and read S^z as a statement about the relative movement of these two terms. For observing shrinkage in normal clusters, we expect the $\bar{R}^z$ to be negative, and for normal clusters moving out, we expect $\bar{D}^z_{\text{inter}}$ to be positive, and as the combination of both of these metrics we expect absolute of S^z to be a high number indicating a separation.

D.2 Proximity Ratio

As for showing the predicted behavior of the anomalous data approach to the normal clusters in latent space, we have used another metric called Proximity Ratio denoted as ρ^z. For an anomalous point a,

Table 15: **Do the anomalies move into the normal clusters?** All columns are relative changes from x to $\hat{x}$, and all are built from the same quantity $d(a)$, the distance from an anomalous point to its nearest normal centroid. Δd_{samp} is the median of $d(a)$ over all anomalous points, while Δd_{pool} takes the median within each anomaly class and then the median across classes. ΔIC reports whether the anomalies approached faster than the space itself contracted. $\Delta\rho$ divides instead by the radius of each point's own nearest cluster. The arrows mark the direction of improvement, and the best value in each column is set in bold and the second best is underlined.

normal set	mask	loss	sparsity	$\Delta d_{\text{samp}} \downarrow$ (%)	$\Delta d_{\text{pool}} \downarrow$ (%)	ΔIC $\downarrow$ (%)	$\Delta\rho \downarrow$ (%)
$\{4,5,7\}$	none	MSE	–	-0.4	-0.1	-0.5	-0.1
	none	MSE	yes	-0.5	-0.3	-0.4	-0.6
	mask	MSE	–	-1.0	-0.8	-1.0	-0.6
	mask	MSE	yes	-1.1	-1.7	-0.2	-1.0
	none	PP	–	-12.4	**-15.3**	-9.6	-2.2
	none	PP	yes	-12.1	-13.4	-11.8	-9.2
	mask	PP	–	-11.9	-12.7	-10.1	-3.3
	mask	PP	yes	**-15.5**	-14.8	**-14.7**	**-10.4**
$\{2,3,6,9\}$	none	MSE	–	-0.7	-1.1	-0.8	-0.7
	none	MSE	yes	-1.0	-1.4	-1.4	-2.3
	mask	MSE	–	-0.8	-1.6	-0.5	-0.9
	mask	MSE	yes	-2.4	-2.5	-1.3	-2.6
	none	PP	–	-15.3	-19.7	-11.8	-7.2
	none	PP	yes	**-18.1**	**-20.6**	**-15.7**	**-13.0**
	mask	PP	–	-12.7	-15.7	-11.1	-3.9
	mask	PP	yes	-13.6	-14.1	-12.6	-8.7
$\{1,2,5,7,8\}$	none	MSE	–	-0.1	-0.2	-0.3	-0.3
	none	MSE	yes	-0.3	-1.0	-1.1	-1.2
	mask	MSE	–	-0.3	-0.4	-0.1	-0.3
	mask	MSE	yes	-2.7	-2.5	-1.1	-2.3
	none	PP	–	**-11.7**	**-12.0**	**-9.0**	**-6.3**
	none	PP	yes	-5.2	-2.3	-4.3	-3.5
	mask	PP	–	-9.9	-11.1	-7.4	-3.1
	mask	PP	yes	-2.4	-1.2	-2.2	-2.0
$\{0,2,4,5,6,8,9\}$	none	MSE	–	-0.4	-0.1	-0.3	-0.6
	none	MSE	yes	+1.0	+1.9	+1.0	+0.8
	mask	MSE	–	-0.9	+0.1	-0.8	-0.8
	mask	MSE	yes	-2.4	-1.1	-1.1	-1.2
	none	PP	–	**-21.0**	**-17.1**	**-19.4**	-8.8
	none	PP	yes	-10.8	-13.2	-10.1	-5.7
	mask	PP	–	-20.0	-15.1	-18.6	**-11.1**
	mask	PP	yes	-8.7	-9.3	-8.3	-6.9

let $c^* = \arg\min_{c\in\mathcal{C}} \|a - \mu_c^z\|$ be its closest normal cluster and $d^z(a)$ the distance to that cluster's center. The score is then the median, over the anomalous points, of $d^z(a)$ divided by the radius of that closest cluster:

$$d^z(a) = \left\|a - \mu_{c^*}^z\right\|, \qquad \rho^z = \operatorname*{med}_{a\in\mathcal{A}} \frac{d^z(a)}{R_{c^*}^z + \varepsilon}. \tag{34}$$

Note that the denominator here is not a single constant per space: $R_{c^*}^z$ depends on which cluster is closest and therefore varies from one anomalous point to the next. This metric basically tells us whether the anomalies move faster toward the normal clusters than the rate of shrinkage of their closest normal clusters. This, along with Δd columns can tell us whether the anomaly points are actually approaching to the normal cluster, and gives the rate of the approaching behavior.

But since this metric is susceptible to the behavior of the normal cluster as well -a shrink in the normal cluster can be interpreted as the anomalies moving away, inaccurately- we have devised some other ways to understand the behavior of the anomalous data. Firstly, we report the relative change of the median of d^z itself, taken over all anomalous points at once, indicated as Δd_{samp}. Since the

anomaly classes are not equally sized, we then take the median within each anomaly class and the median across those class values, indicated as Δd_{pool}, so that no single class dominates the number,

$$d^z_{\text{samp}} = \operatorname*{med}_{a\in\mathcal{A}} d^z(a), \qquad d^z_{\text{pool}} = \operatorname*{med}_{b\in\mathcal{B}} \Big(\operatorname*{med}_{a\in\mathcal{A}_b} d^z(a)\Big). \tag{35}$$

What the pair of them diagnoses is the effect of the class sizes: d^z_{samp} lets the larger anomaly classes weigh more, while d^z_{pool} does not. It should be noted that with M small and odd the outer median selects one class value rather than combining them, so d^z_{pool} can coincide with a single anomaly class throughout.

And lastly, we report a second normalised score, in which $d^z(a)$ is divided by the mean inter-normal-centroid distance $\bar{D}^z_{\text{inter}}$ rather than by a cluster radius. Unlike the denominator of ρ^z, this one is a single constant per space and is independent of which cluster is closest, so it acts as a scale reference: IC^z is unchanged under a uniform rescaling of the latent space and asks whether the anomalies approached the normal classes faster than the space as a whole contracted.

$$\mathrm{IC}^z = \operatorname*{med}_{a\in\mathcal{A}} \frac{d^z(a)}{\bar{D}^z_{\text{inter}} + \varepsilon}. \tag{36}$$

The expectation from ρ^z and IC^z is for them to drop to show an approaching behavior of the anomalous data to their closest overall normal cluster, and thus the $\Delta\rho$ and $\Delta\mathrm{IC}$ results are expected to be negative.

E Standard One-Class Anomaly Detection

This section gives the protocol, competing methods, implementation details, and ablations for the experiments in Section 4.3. It also reports the class-wise results omitted from the main paper.

E.1 Benchmarks and protocol

We use MNIST (Deng, 2012), Fashion-MNIST (Xiao et al., 2017), CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), and Arrhythmia (Guvenir et al., 1997). MNIST, Fashion-MNIST, and CIFAR-10 define ten one-class problems, one for each nominal class. CIFAR-100 defines twenty problems using its coarse superclasses. Arrhythmia defines one problem; following Ghafourian et al. (2025), classes 3, 4, 5, 7, 8, 9, 14, and 15 are anomalous and the remaining classes are nominal.

For each image problem, the training split is restricted to the nominal class. The validation and test splits retain all classes and are relabelled as nominal or anomalous. We use ten seeds for each nominal class. A seed controls model initialisation and stochastic operations but does not change the data split. Validation is performed every 20 iterations, and the reported value is the test score at the checkpoint selected on the corresponding validation metric. We use AUC-ROC for the image datasets. For Arrhythmia, the 15% of test samples with the highest anomaly scores are labelled anomalous before computing F_1, as in Ghafourian et al. (2025). Class-wise results are means and standard deviations over seeds; dataset averages are unweighted means over nominal classes.

MNIST and Fashion-MNIST inputs are standardised to zero mean and unit variance. CIFAR-10 and CIFAR-100 use global contrast normalisation with the ℓ_1 norm, followed by min–max scaling to $[0, 1]$, as in Ghafourian et al. (2025). No anomalous sample is used for training. Table 16 summarises the data used before restricting the training pool to a nominal class.

E.2 Competing methods

We compare against the following competing methods: OC-SVM (Schölkopf et al., 1999), KDE (Parzen, 1962), Isolation Forest (Liu et al., 2008), Deep SVDD (Ruff et al., 2018), the winner-take-all autoencoder (DCAE) (Makhzani & Frey, 2015), VAE (Kingma & Welling, 2013), MemAE (Gong et al., 2019), MemGAN (Yang et al., 2021), deep structured energy-based models (DSEBM and DSEBM-e) (Zhai et al., 2016), DAGMM (Zong et al., 2018), AnoGAN (Schlegl et al., 2017), ALAD (Zenati et al., 2018), ADGAN (Deecke et al., 2018), OCGAN (Perera et al.,

Table 16: One-class benchmark statistics. For the image datasets, the training count denotes the full pool before restriction to the nominal class. All our variants and seeds use the same fixed splits.

Dataset	Training pool	Validation	Test	Input shape	Problems
MNIST	50 000	10 000	10 000	1×28×28	10
Fashion-MNIST	50 000	10 000	10 000	1×28×28	10
CIFAR-10	40 000	10 000	10 000	3×32×32	10
CIFAR-100	40 000	10 000	10 000	3×32×32	20
Arrhythmia	452 records (fixed stratified split)			274 features	1

Table 17: Plain AE, AE-PP and AE-PoS on the one-class benchmarks.

Dataset	AE	AE-PP	AE-PoS
MNIST	94.74±4.65	**99.38±0.42**	99.35±0.51
Fashion-MNIST	91.60±5.34	94.83±3.95	**94.87±4.04**
CIFAR-10	68.96±7.29	71.30±6.70	**71.73±6.60**
CIFAR-100	65.42±9.09	68.56±8.89	**68.87±8.97**
Arrhythmia (F1)	60.77±7.26	67.69±5.75	**70.77±5.75**

2019), GANomaly (Akcay et al., 2019), RCGAN (Yang et al., 2020), GOAD (Bergman & Hoshen, 2020), geometric-transformation classification (GeoTrans) (Golan & El-Yaniv, 2018), the attribute-restoration framework (ARNet) (Ye et al., 2022), and targeted collapse (TOLL) (Ghafourian et al., 2025).

TOLL adds a latent-norm compactness term to this autoencoder family; its training loss and anomaly score include $\beta\|x\|_2$. The TOLL column denotes that regularised method, whereas AE denotes the $\beta = 0$ reconstruction-only baseline. AE-PP and AE-PoS also set $\beta = 0$: AE-PP adds Dynamic Push and Pull to the common backbone, and AE-PoS further applies Nested Manifold Carving to the frozen Stage-0 representation. Thus, the methods share TOLL's architecture and optimisation settings, but not its compactness objective.

E.3 Models and training

Common backbone: AE, AE-PP, and the Stage-0 component of AE-PoS use the autoencoder architecture, latent width, optimiser, learning rate, and batch size specified by Ghafourian et al. (2025). This holds model capacity and the basic optimisation budget fixed while changing the objective. Stage 0 trains for 2,000 iterations with Adam and no weight decay. Stage 1 freezes the Stage-0 encoder and decoder from the same nominal class and seed, then trains a nested MLP encoder–decoder for 6,000 iterations. The nested MLP uses ReLU, with LayerNorm on MNIST and Fashion-MNIST and BatchNorm on Arrhythmia, CIFAR-10, and CIFAR-100. Its dimensions and dataset-specific settings are given in Tables 18 and 19.

Objectives and score: AE minimises clean reconstruction error. AE-PP retains that term and adds the Dynamic Push and Pull loss using the negative views in Table 20. AE-PoS trains Stage 1 inside the frozen Stage-0 model. All reconstruction and push-pull distances use the ℓ_2 norm. At test time, AE, AE-PP, and AE-PoS use reconstruction error as the anomaly score.

E.4 Push-pull negative views

The negative view is generated from a nominal training sample and is used only to construct the push-pull objective; it is not labelled anomaly data. At Stage 1, MNIST, Fashion-MNIST, and Arrhythmia reuse their Stage-0 transformation, and the CIFAR models perturb the frozen Stage-0 latent rather than an image. Table 20 gives the complete configuration.

Table 18: Stage-0 AE-PP settings. Latent width, learning rate, and batch size follow Ghafourian et al. (2025); scale and weight specify Dynamic Push and Pull. All runs use Adam without weight decay for 2,000 iterations and validate every 20 iterations.

				Push-pull	
Dataset	Latent	LR	Batch	Scale	Weight
Arrhythmia	16	10^{-3}	100	3.00	1.00
MNIST	128	10^{-4}	1000	3.00	1.00
Fashion-MNIST	256	10^{-4}	500	3.00	1.00
CIFAR-10	64	10^{-2}	50	2.00	1.00
CIFAR-100	64	10^{-3}	100	2.00	1.00

Table 19: Stage-1 AE-PoS settings. Stage 1 is trained within the frozen Stage 0 model for each nominal class and seed. The decoder mirrors the listed MLP encoder path.

				Push-pull		
Dataset	Encoder path	LR	Batch	Space	Scale	Weight
Arrhythmia	$16 \to 8 \to 2$	10^{-3}	100	latent	1.00	1.00
MNIST	$128 \to 256 \to 256$	10^{-4}	1000	pixel	2.00	1.00
Fashion-MNIST	$256 \to 384 \to 384$	10^{-4}	500	pixel	2.00	4.00
CIFAR-10	$64 \to 32 \to 32$	10^{-2}	50	latent	2.00	1.00
CIFAR-100	$64 \to 96 \to 96$	10^{-3}	100	latent	2.00	1.00

E.5 Ablation and complete results

The controlled ablation uses the same backbone, data splits, and training budget for AE, AE-PP, and AE-PoS. AE isolates reconstruction, AE-PP adds Dynamic Push and Pull, and AE-PoS further adds Stage 1.

Tables 21, 22, 23, and 24 report the complete class-wise results for the image datasets. In particular, Table 23 expands the average CIFAR-10 comparison reported in Table 4 of the main paper. Arrhythmia, which comprises a single one-class task, is reported directly in Table 3 of the main paper.

E.6 Push–Pull Sensitivity

Table 25 evaluates the sensitivity of AE-PP to the margin scale α, with the Push–Pull weight fixed at $w = 1$. The AE column is the reconstruction-only baseline obtained using the common TOLL backbone with its compactness coefficient set to $\beta = 0$, and is therefore distinct from the AE-PP configurations. When $\alpha = 0$, the Push term vanishes, leaving only the Pull objective, which maps a perturbed nominal view to its clean counterpart. This is exactly a denoising autoencoder when the perturbation is additive Gaussian noise, as in our controlled multi-component MNIST experiment, and more generally subsumes perturbation-to-clean autoencoders based on CutPaste, shuffling, rotations, and related transformations used in the other settings. For $\alpha > 0$, the same dataset-specific perturbations additionally activate Push and impose an explicit separation margin. The consistent improvement from $\alpha = 0$ to $\alpha > 0$ isolates the benefit of the proposed Push term beyond denoising and supports its intended role in separating overlapping nominal submanifolds. Most gains are obtained by $\alpha = 2$ or 3, after which performance remains broadly stable. Validation selects $\alpha = 3$ for Arrhythmia, MNIST, and Fashion-MNIST, and $\alpha = 2$ for CIFAR-10 and CIFAR-100, indicating limited sensitivity once a sufficient Push margin is imposed. Arrhythmia is evaluated using F1, while all other datasets use AUC-ROC.

Under the conditions of Lemma 1, Eq. (6) also predicts saturation with respect to α. Once α exceeds the zero-Push-error bound $1 + \eta(s_n)/\delta_e$ for most training pairs, the hinge remains active, and its gradient, $-\nabla_\theta d(\hat{s}_n, \tilde{s}_n)$, no longer depends on α. Consistent with this prediction, performance plateaus at $\alpha = 2$ or 3; on Arrhythmia, all settings $\alpha \in \{3, 4, 5, 6\}$ produce identical results.

Table 20: Negative views used by Dynamic Push and Pull. *OneOf* samples one transformation per input with weights 3:1:1:1 in the listed order.

Dataset	Stage 0	Stage 1
MNIST	CutPaste with three rotated patches	Same CutPaste configuration
Fashion-MNIST	*OneOf*: CutPaste (area 0.15–0.4, two patches), patch shuffle (grid 2), 90° rotation, and phase scramble (0.3)	Same *OneOf* configuration
CIFAR-10	*OneOf*: phase scramble (0.3), patch shuffle (grid 2), CutPaste, and channel shuffle	Gaussian noise on the Stage-0 latent, with relative $\sigma = 0.03$ and $p = 0.8$
CIFAR-100	*OneOf*: phase scramble (0.3), patch shuffle (grid 2), CutPaste, and channel shuffle	Gaussian noise on the Stage-0 latent, with relative $\sigma = 0.03$ and $p = 0.8$
Arrhythmia	Feature shuffle with fraction 0.2	Same feature-shuffle configuration

Table 21: Mean and standard deviation of AUC-ROC in % on MNIST (Deng, 2012) classes over 10 seeds. The highest value in each row is in bold.

Normal	OC-SVM	OCGAN	VAE	DCAE	IF	AnoGAN	KDE	DSVDD	MEMAE	MEMGAN	Toll	AE-PP	AE-PoS
0	98.6±0.0	99.8	99.7	97.6±0.7	98.0±0.3	96.6±1.3	97.1±0.0	98.0±0.7	99.3±0.1	99.3±0.1	99.8±0.1	**99.91±0.02**	99.90±0.02
1	99.5±0.0	99.9	99.9	98.3±0.6	97.3±0.4	99.2±0.6	98.9±0.0	99.7±0.1	99.8±0.0	99.9±0.0	99.9±0.0	99.94±0.01	**99.95±0.01**
2	82.5±0.1	94.2	93.6	85.4±2.4	88.6±0.5	85.0±2.9	79.0±0.0	91.7±0.8	90.6±0.8	94.5±0.1	96.2±1.6	99.24±0.05	**99.28±0.05**
3	88.1±0.0	96.3	95.9	86.7±0.9	89.9±0.4	88.7±2.1	86.2±0.0	91.9±1.5	94.7±0.6	95.7±0.4	97.9±0.6	99.14±0.11	**99.15±0.12**
4	94.9±0.0	97.5	97.3	86.5±2.0	92.7±0.6	89.4±1.3	87.9±0.0	94.9±0.8	94.5±0.4	96.1±0.4	97.8±0.4	**99.06±0.03**	99.03±0.05
5	77.1±0.0	98.0	96.4	78.2±2.7	85.5±0.8	88.3±2.9	73.8±0.0	88.5±0.9	95.1±0.1	93.6±0.3	98.1±0.6	99.33±0.03	**99.34±0.04**
6	96.5±0.0	99.1	99.3	94.6±0.5	95.6±0.3	94.7±2.7	87.6±0.0	98.3±0.5	98.4±0.5	98.6±0.1	99.5±0.1	**99.84±0.01**	**99.84±0.02**
7	93.7±0.0	98.1	97.6	92.3±1.0	92.0±0.4	93.5±1.8	91.4±0.0	94.6±0.9	95.4±0.2	96.2±0.2	98.7±0.2	**99.28±0.06**	**99.28±0.07**
8	88.9±0.0	93.9	92.3	86.5±1.6	89.9±0.4	84.9±2.1	79.2±0.0	93.9±1.6	86.9±0.5	93.5±0.1	97.3±0.5	**98.54±0.23**	98.17±0.21
9	93.1±0.0	98.1	97.6	90.4±1.8	93.5±0.3	92.4±1.1	88.2±0.0	96.5±0.3	97.3±0.2	95.9±0.1	98.5±0.3	99.49±0.04	**99.56±0.03**
Average	91.3	97.5	97.0	89.7	92.3	91.4	87.0	94.8	95.2	96.5	98.4	**99.38**	99.35

E.7 Detecting Unseen Degradations in High-Quality Images

E.7.1 Dataset Construction

Training data: All models are trained on the same set of 8,693 clean images: 800 from DIV2K (Agustsson & Timofte, 2017), 2,649 from Flickr2K (Lim et al., 2017), 4,744 from the Waterloo Exploration Database (Ma et al., 2017), and 500 from BSDS500 (Arbeláez et al., 2011).

Synthetic degraded views during training: During training, each 256×256 clean crop is paired with a degraded view generated online for the Push–Pull loss. Defocus blur ($r \in [1, 6]$), motion blur ($\ell \in [3, 25]$), and Gaussian blur ($\sigma \in [0.2, 1.0]$) are selected independently with probabilities 0.55, 0.40, and 0.30, respectively. When n blur types are selected, their strengths are scaled by $1/\sqrt{n}$. Their kernels are combined with an optional, unattenuated sinc kernel ($p = 0.10$) and applied once in linear-light space using $\gamma = 2.2$. Spatially varying defocus is used for 20% of the defocus samples. The blurred image is then downsampled by a random factor of at most 2 and resized to its original resolution with probability 0.5. Gaussian sensor noise ($\sigma \in [0, 6]$) is always added, followed by JPEG compression with quality 70–95 with probability 0.8.

REDS blur set: The REDS blur set contains five aligned degraded versions of each ground-truth video frame, with blur levels 0.2, 0.4, 0.6, 0.8, and 1.0. For evaluation, we use the same provided subset of 200 frames at every level, with one frame selected from each video clip. The degradation increases monotonically with the level: reconstruction error relative to the ground truth rises while image sharpness decreases. The blur follows local scene motion, leaving static regions nearly unchanged while increasingly blurring moving regions, rather than applying a single spatially invariant kernel to the whole image. Empirically, the mean squared deviation from the ground truth grows approximately quadratically with the blur level.

Underwater degradation: We use underwater images from the LSUI dataset (Peng et al., 2023), grouped by degradation severity following Devecioglu et al. (2025). The groups are defined using PSNR between each underwater image and its paired LSUI reference: high degradation corresponds to 8.79-16.57 dB, medium degradation to 16.57-19.79 dB, and low degradation to 19.79-31.53 dB. Our evaluation contains 144, 143, and 143 images from these groups, respectively. The paired LSUI

Table 22: One-class anomaly detection on Fashion-MNIST (Xiao et al., 2017), measured by AUC-ROC (%). Values with uncertainty are reported as mean ± standard deviation over 10 seeds. Bold indicates the highest mean in each row.

	Baselines							Ours	
Nominal	DAGMM	DSEBM	ADGAN	GANomaly	GeoTrans	ARNet	TOLL	AE-PP	AE-PoS
T-shirt/top	42.1	91.6	89.9	80.3	**99.4**	92.7	91.9±0.2	95.12±0.09	94.92±0.10
Trouser	55.1	71.8	81.9	83.0	97.6	**99.3**	98.7±0.0	98.91±0.06	99.24±0.08
Pullover	50.4	88.3	87.6	75.9	91.1	89.1	91.3±0.2	**92.47±0.20**	92.14±0.12
Dress	57.0	87.3	91.2	87.2	89.9	93.6	92.4±0.2	95.24±0.19	**95.40±0.11**
Coat	26.9	85.2	86.5	71.4	92.1	90.8	91.5±0.1	92.92±0.22	**93.25±0.27**
Sandal	70.5	87.1	89.6	92.7	93.4	93.1	93.1±0.2	**94.28±0.39**	93.07±0.46
Shirt	48.3	73.4	74.3	81.0	83.3	85.0	84.6±0.3	85.13±0.20	**85.37±0.14**
Sneaker	83.5	98.1	97.2	88.3	98.9	98.4	99.0±0.0	**99.15±0.09**	99.13±0.03
Bag	49.9	86.0	89.0	69.3	90.8	**97.8**	90.8±0.6	96.56±0.48	97.46±0.13
Ankle boot	34.0	97.1	97.1	80.3	**99.2**	98.4	98.9±0.1	98.51±0.12	98.71±0.08
Average	51.8	86.6	88.4	80.9	93.5	93.9	93.2	94.83	**94.87**

Table 23: One-class anomaly detection on CIFAR-10 (Krizhevsky, 2009), measured by AUC-ROC (%). AE-PP and AE-PoS results are reported as mean ± standard deviation over 10 seeds. Bold indicates the highest mean in each row.

	Baselines											Ours	
Nominal	OCGAN	VAE	DSVDD	DSEBM	DAGMM	IF	AnoGAN	ALAD	MEMAE	MEMGAN	TOLL	AE-PP	AE-PoS
Airplane	75.7	70.0	61.7±4.1	41.4±2.3	56.0±6.9	60.1±0.7	67.1±2.5	64.7±2.6	66.5±0.9	73.0±0.8	75.0±4.3	75.53±2.18	**75.99±1.94**
Automobile	53.1	38.6	65.9±2.1	57.1±2.0	48.3±1.8	50.8±0.6	54.7±3.4	38.7±0.8	46.4±0.1	52.5±0.7	67.6±3.2	68.65±1.46	**69.10±1.32**
Bird	64.0	**67.9**	50.8±0.8	61.9±0.1	53.8±4.0	49.2±0.4	52.9±3.0	67.0±0.7	66.0±0.1	67.2±0.1	56.2±2.8	59.47±1.95	59.99±1.94
Cat	62.0	53.5	59.1±1.4	50.1±0.4	51.2±0.8	55.1±0.4	54.5±1.9	59.2±0.3	52.9±0.1	57.3±0.2	**64.6±1.6**	63.41±0.47	63.91±0.58
Deer	72.3	**74.8**	60.9±1.1	73.3±0.2	52.2±7.3	49.8±0.4	65.1±3.2	72.7±0.6	72.8±0.1	73.9±0.9	67.7±1.4	69.47±0.81	70.01±0.99
Dog	62.0	52.3	65.7±0.8	60.5±0.3	49.3±3.6	58.5±0.4	60.3±2.6	52.8±1.2	52.9±0.2	65.0±0.2	68.1±1.6	68.85±0.71	**69.36±0.43**
Frog	72.3	68.7	67.7±2.6	68.4±0.3	64.9±1.7	42.9±0.6	58.5±1.4	69.5±1.1	63.7±0.4	72.8±0.7	74.7±2.5	76.77±0.51	**77.20±0.75**
Horse	57.5	49.3	67.3±0.9	53.3±0.7	55.3±0.8	55.1±0.7	62.5±0.8	44.8±0.4	45.9±0.1	52.5±0.5	65.8±1.3	71.24±1.47	**71.71±1.42**
Ship	82.0	69.6	75.9±1.2	73.9±0.3	51.9±2.4	74.2±0.6	75.8±4.1	73.4±0.4	70.1±0.1	74.4±0.3	80.0±1.1	83.12±0.63	**83.33±0.62**
Truck	55.4	38.6	73.1±1.2	63.6±3.1	54.2±5.8	58.9±0.7	66.5±2.8	39.2±1.3	48.2±0.2	65.6±1.6	76.2±0.5	76.53±0.51	**76.69±0.60**
Average	65.6	58.3	64.8	60.4	54.4	55.5	61.8	59.3	58.5	65.3	69.6	71.30	**71.73**

Table 24: One-class anomaly detection on the CIFAR-100 coarse superclasses (Krizhevsky, 2009), measured by AUC-ROC (%). Values with uncertainty are reported as mean ± standard deviation over 10 seeds. Bold indicates the highest mean in each row.

	Baselines							Ours	
Nominal	DAGMM	DSEBM	ADGAN	GANomaly	GeoTrans	ARNet	TOLL	AE-PP	AE-PoS
Aquatic mammals	43.4	64.0	63.1	57.9	74.7	**77.5**	70.3±3.4	69.30±1.55	69.62±1.07
Fish	49.5	47.9	54.9	51.9	68.5	70.0	66.2±1.1	69.99±1.49	**70.38±1.41**
Flowers	66.1	53.7	41.3	36.0	74.0	62.4	80.3±0.7	80.13±1.06	**80.85±1.06**
Food containers	52.6	48.4	50.0	46.5	**81.0**	76.2	58.2±1.5	66.80±3.91	67.27±3.95
Fruit and vegetables	56.9	59.7	40.6	46.6	78.4	77.7	80.1±1.3	79.78±0.72	**80.36±0.62**
Household electrical devices	52.4	46.6	42.8	42.9	59.1	**64.0**	61.4±2.6	61.56±1.92	61.56±1.89
Household furniture	55.0	51.7	51.1	53.7	81.8	**86.9**	64.8±3.0	69.64±1.51	69.78±1.44
Insects	52.8	54.8	55.4	59.4	65.0	65.6	70.5±1.3	72.51±0.93	**73.16±1.37**
Large carnivores	53.2	66.7	59.2	63.7	**85.5**	82.7	62.1±2.1	59.37±1.98	59.59±1.97
Large man-made outdoor things	42.5	71.2	62.7	68.0	**90.6**	90.2	79.6±1.4	81.53±1.19	81.99±0.98
Large natural outdoor scenes	52.7	78.3	79.8	75.6	**87.6**	85.9	81.6±1.1	83.89±1.02	84.15±0.98
Large omnivores and herbivores	46.4	62.7	53.7	57.6	**83.9**	83.5	62.6±1.2	61.04±0.84	61.19±0.77
Medium-sized mammals	42.7	66.8	58.9	58.7	83.2	**84.6**	61.2±1.9	61.21±1.06	61.40±1.33
Non-insect invertebrates	45.4	52.6	57.4	59.9	58.0	**67.6**	58.6±2.0	55.65±2.03	55.95±2.18
People	57.2	44.0	39.4	43.9	**92.1**	84.2	65.4±4.5	65.01±1.74	65.11±1.75
Reptiles	48.8	56.8	55.6	59.9	68.3	**74.1**	56.1±1.5	53.48±1.68	53.77±1.67
Small mammals	54.4	63.1	63.3	64.4	73.5	**80.3**	62.3±1.1	63.58±1.20	63.94±1.27
Trees	36.4	73.0	66.7	71.8	**93.8**	91.0	80.9±1.8	81.19±0.82	81.31±0.69
Vehicles 1	52.4	57.7	44.3	54.9	**90.7**	85.3	65.2±1.4	65.77±1.19	66.20±1.45
Vehicles 2	50.3	55.5	53.0	56.8	85.0	**85.4**	63.6±1.6	69.82±1.03	69.85±1.04
Average	50.5	58.8	54.7	56.5	78.7	**78.8**	67.6	68.56	68.87

Table 25: Effect of the Push–Pull margin scale α. AE is the reconstruction-only baseline obtained using the common TOLL backbone with its compactness coefficient set to $\beta = 0$. All AE-PP configurations use $w = 1$; $\alpha = 0$ retains Pull while disabling Push. Results are reported as mean $\pm$ standard deviation. Bold indicates the highest mean in each row, and the dagger marks the value of α selected on the validation set.

	Baseline	AE-PP ($w = 1$)						
Dataset	AE ($\beta = 0$)	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$	$\alpha = 6$
Arrhythmia (F1)	60.77±7.26	63.08±3.24	63.08±3.24	65.38±4.05	**67.69±6.07**†	**67.69±6.07**	**67.69±6.07**	**67.69±6.07**
MNIST	94.74±4.65	98.54±1.38	99.04±0.77	99.37±0.44	**99.38±0.42**†	99.32±0.45	99.28±0.48	99.26±0.50
Fashion-MNIST	91.60±5.34	93.75±4.60	94.12±4.43	94.74±4.02	**94.83±3.97**†	94.74±3.99	94.67±3.94	94.69±3.95
CIFAR-10	68.96±7.29	69.32±7.36	69.48±7.13	**71.30±6.73**†	71.16±6.81	71.05±6.78	71.23±6.97	71.29±7.06
CIFAR-100	65.42±9.09	66.78±9.26	67.18±9.17	68.56±8.91†	68.45±8.98	68.59±8.69	**68.69±8.53**	68.57±8.72

Table 26: Detection of unseen image degradations. The clean reference defines the nominal class: paired clean images (Own GT) are used when available; otherwise, a fixed set of clean DIV8K crops is used. AUC-ROC and PR-AUC are computed using reconstruction error as the anomaly score. Bold indicates the best result in each row and metric.

Dataset	Clean reference	# Images	AE		VAE		AE-PP		AE-PoS	
			ROC	PR	ROC	PR	ROC	PR	ROC	PR
Additive noise										
Kodak24 ($\sigma = 25$)	Own GT	24	**1.0000**	**1.0000**	0.9983	0.9983	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Blur										
REDS blur 0.2	Own GT	200	**0.7268**	**0.6589**	0.3552	0.4024	0.6083	0.5725	0.6002	0.5766
REDS blur 0.4	Own GT	200	0.6581	0.5683	0.2168	0.3484	0.7597	0.7239	**0.7808**	**0.7574**
REDS blur 0.6	Own GT	200	0.6726	0.5699	0.1111	0.3245	0.9211	0.9222	**0.9554**	**0.9567**
REDS blur 0.8	Own GT	200	0.7703	0.6713	0.0528	0.3151	0.9920	0.9930	**0.9972**	**0.9975**
REDS blur 1.0	Own GT	200	0.5724	0.5215	0.0346	0.3119	0.9985	0.9986	**0.9996**	**0.9996**
GoPro	Own GT	200	0.4114	0.4217	0.0569	0.3137	0.9471	0.9568	**0.9846**	**0.9862**
GoPro	DIV8K	200	0.1621	0.3378	0.1018	0.3231	0.9983	0.9982	**0.9997**	**0.9997**
Rain										
Test100	DIV8K	98	0.3284	0.3863	0.7842	0.6766	0.9481	0.9418	**0.9594**	**0.9470**
Rain100L	DIV8K	100	0.4709	0.4461	0.7467	0.6432	0.8708	0.8366	**0.8925**	**0.8479**
Rain100H	DIV8K	100	0.8196	0.7307	0.9051	0.8269	0.9436	0.9131	**0.9570**	**0.9226**
Test1200	DIV8K	200	0.2025	0.3458	0.7151	0.5944	0.9885	0.9864	**0.9919**	**0.9881**
Test2800	DIV8K	200	0.3801	0.4016	0.8017	0.6917	0.9798	0.9764	**0.9879**	**0.9825**
Underwater										
Underwater high	DIV8K	144	0.3673	0.4010	0.5254	0.4684	0.9846	0.9848	**0.9891**	**0.9888**
Underwater medium	DIV8K	143	0.3555	0.3974	0.5308	0.4769	0.9750	0.9753	**0.9866**	**0.9857**
Underwater low	DIV8K	143	0.1630	0.3359	0.2918	0.3729	0.9412	0.9246	**0.9798**	**0.9677**

references are used only to determine degradation severity; clean DIV8K images define the nominal class in our anomaly-detection evaluation.

Anomalous test data: The evaluation covers 16 settings across five degradation families. Additive noise uses the 24 Kodak24 images (Franzen, 1999) corrupted with Gaussian noise at $\sigma = 25$. The REDS blur set (Nah et al., 2019) provides five aligned severity levels from 0.2 to 1.0, evaluated on the same 200 frames at every level. GoPro (Nah et al., 2017) contributes 200 motion-blurred images and is evaluated under two nominal-reference protocols. Rain comprises Test100 (98 images) (Zhang et al., 2020), Rain100L and Rain100H (100 each) (Yang et al., 2017), Test1200 (200) (Zhang & Patel, 2018), and Test2800 (200) (Fu et al., 2017). Finally, LSUI (Peng et al., 2023) contributes high-, medium-, and low-degradation underwater groups containing 144, 143, and 143 images, respectively, following the PSNR-based partition of Devecioglu et al. (2025).

Nominal test data: For Kodak24, REDS, and the paired GoPro protocol, we use the clean counterpart of each degraded image, so the two classes differ only in degradation. For rain, underwater, and the second GoPro protocol, the nominal class is drawn from DIV8K (Gu et al., 2019). These nominal sets are matched to the corresponding anomalous sets in size and centre-cropped at native resolution

Table 27: Average reconstruction PSNR (dB), computed between each input s and its reconstruction $\hat{s}$, for clean and degraded images. Higher PSNR indicates lower reconstruction error. The # Images column reports the number of images in each evaluation setting. "Own GT" denotes the paired clean ground truth associated with each degraded image, whereas "DIV8K" denotes the fixed clean DIV8K evaluation subset.

Dataset	Clean reference	# Images	AE		VAE		AE-PP		AE-PoS	
			Clean	Degr.	Clean	Degr.	Clean	Degr.	Clean	Degr.
Additive noise										
Kodak24 ($\sigma = 25$)	Own GT	24	57.15	26.49	26.49	18.86	44.58	23.64	43.65	23.55
Synthetic blur										
REDS blur 0.2	Own GT	200	63.04	62.06	27.59	29.11	43.42	42.55	43.07	42.38
REDS blur 0.4	Own GT	200	63.04	62.41	27.59	30.84	43.42	41.00	43.07	40.88
REDS blur 0.6	Own GT	200	63.04	62.37	27.59	32.85	43.42	38.12	43.07	37.97
REDS blur 0.8	Own GT	200	63.04	61.95	27.59	35.22	43.42	33.26	43.07	33.60
REDS blur 1.0	Own GT	200	63.04	62.72	27.59	37.01	43.42	30.51	43.07	31.79
Motion blur										
GoPro	Own GT	200	62.66	63.12	31.15	39.17	33.80	28.11	34.50	28.55
GoPro	DIV8K	200	59.14	63.12	29.58	39.17	44.28	28.11	43.54	28.55
Rain										
Test100	DIV8K	98	58.43	61.11	29.86	23.51	44.59	34.35	43.53	34.58
Rain100L	DIV8K	100	58.73	59.85	30.53	24.54	44.89	38.60	43.77	38.30
Rain100H	DIV8K	100	58.73	54.50	30.53	20.75	44.89	37.08	43.77	36.77
Test1200	DIV8K	200	59.41	63.13	30.33	25.31	44.67	31.72	43.77	32.80
Test2800	DIV8K	200	59.44	60.93	30.22	23.47	44.58	32.12	43.69	32.90
Underwater										
Underwater high	DIV8K	144	58.91	61.03	29.64	29.01	44.19	29.61	43.43	30.76
Underwater medium	DIV8K	143	58.91	61.00	29.53	29.01	44.10	30.07	43.42	31.06
Underwater low	DIV8K	143	58.91	63.02	29.53	33.71	44.10	35.30	43.42	33.94

to the largest image dimensions in each test set. The DIV8K protocols are therefore corpus-level rather than scene-paired comparisons and may also reflect differences in image content and native resolution; in particular, crops from 8K images generally contain finer per-pixel detail than equally sized crops from lower-resolution images.

E.7.2 Compared Models

Autoencoder architecture (AE, AE-PP, AE-PoS): The Stage-0 encoder maps the RGB input to 128 channels using a stride-2 patch convolution, applies four residual blocks, and projects the result to a 9-channel latent representation at half the input resolution. This gives a latent-to-input dimensionality ratio of $9(H/2)(W/2)/(3HW) = 0.75$. The decoder mirrors the encoder and restores the original resolution using pixel shuffle. Each residual block consists of LayerNorm, a 1×1 expansion to twice the channel width, a 3×3 depthwise convolution, a ReLU-gated elementwise product between the two channel halves, and a 1×1 projection back to the original width. All convolutions are bias-free, and the network uses neither attention nor encoder–decoder skip connections. This is the architecture used in the AE and AE-PP baselines. Stage 1 applies the same design to the frozen 9-channel Stage-0 latent, using an overcomplete $9 \to 27 \to 9$ mapping without changing the spatial resolution.

Optimisation: Stage 0 is trained for 250k iterations using AdamW with a learning rate of 10^{-3}, zero weight decay, a batch size of 4, and gradient clipping at 1.0. We use a two-cycle cosine schedule with periods $[150\text{k}, 100\text{k}]$ and relative weights $[1, 0.5]$. Stage 1 follows the same training schedule with a learning rate of 10^{-4}.

VAE: As a pretrained baseline, we use the frozen Stable Diffusion KL-$f8$ autoencoder (Rombach et al., 2022) with the MSE-fine-tuned decoder released as `sd-vae-ft-mse` (Stability AI, 2022). It is evaluated deterministically through its 4-channel latent representation, without training on our data. Because its decoder was fine-tuned using mean squared error, the checkpoint is well matched to our reconstruction-error score. The VAE contains $83{,}653{,}863$ parameters, compared with $830{,}720$

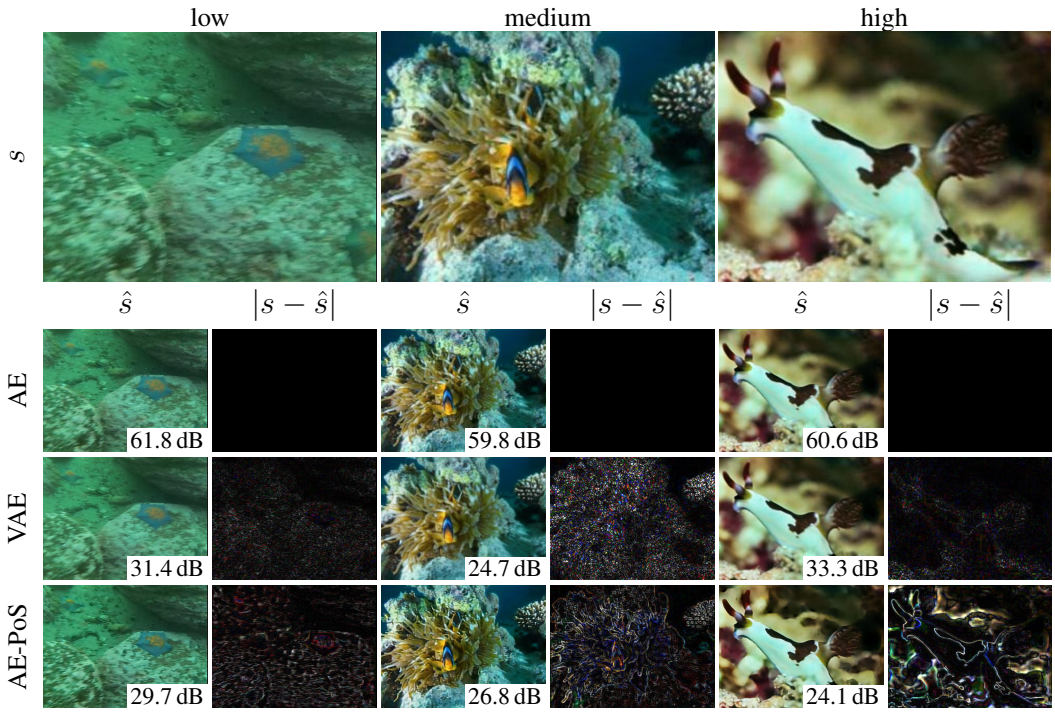

Figure 6: Underwater degradation at three severities (columns). Top row: the input s; below it each network's reconstruction $\hat{s}$ with its PSNR and the absolute difference $|s - \hat{s}|$. All difference panels share one intensity gain.

for AE and AE-PP and $1{,}665{,}280$ for the complete two-stage AE-PoS. It is therefore approximately $101\times$ larger than AE and $50\times$ larger than AE-PoS.

E.8 Results

Tables 26 and 27 show that AE-PP and AE-PoS combine reliable detection of unseen degradations with high nominal reconstruction fidelity. The visual comparisons in Figures 7, 8, and 6 explain the contrasting behaviours. Plain AE preserves fine image details but also reconstructs degraded inputs, resulting in weak anomaly separation. VAE has substantially lower fidelity on nominal images and increasingly represents the low-frequency blurred component as blur strengthens; under strong blur, it even reconstructs the degraded input more accurately than its clean counterpart, reversing the desired anomaly-score ordering. AE-PoS instead preserves nominal content while excluding unseen degradations from its learned target set, causing them to appear as strong reconstruction residuals.

This combination is particularly useful as a lightweight pre-detection stage in mobile imaging pipelines. When most captured images are already of sufficient quality, only those identified as degraded need to be passed to computationally expensive restoration networks, reducing the expected latency and energy consumption.

F Transferable ECG Representations

Task and baselines: Clinical annotation of 12-lead ECG recordings requires expert interpretation, whereas large collections of unlabelled recordings are readily available. We therefore evaluate whether nested representation learning improves a self-supervised ECG encoder before downstream linear probing. We compare with the published results of ECG-CPC (Al-Masud et al., 2026), ECG-FM (McKeen et al., 2025), MERL (Liu et al., 2024), and ST-MEM (Na et al., 2024).

Depth-controlled architecture: ST-MEM learns ECG representations by reconstructing masked portions of 12-lead recordings. For a capacity-controlled comparison, Stage 0 follows the ST-MEM

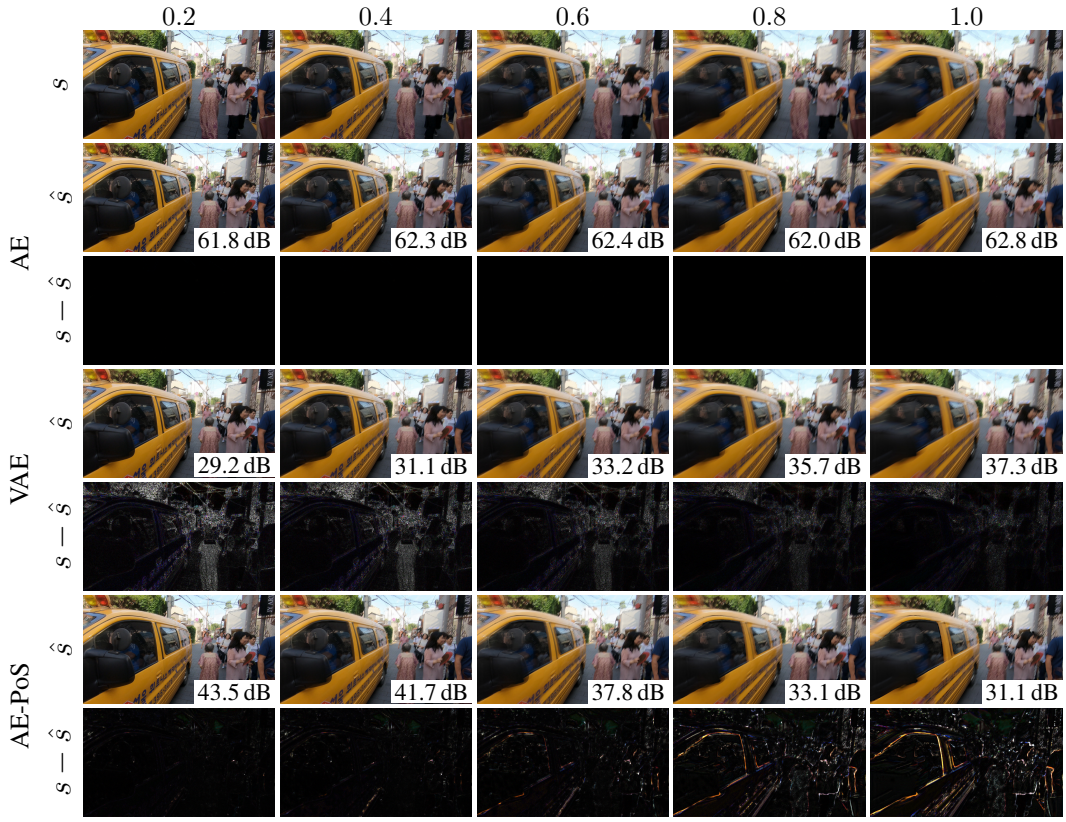

Figure 7: REDS under increasing blur. Columns show the same scene at five blur strengths; rows show the input s, each model's reconstruction $\hat{s}$ with its PSNR, and the absolute residual $|s - \hat{s}|$. The VAE increasingly represents the blurred component: its reconstruction PSNR rises from 29.2 to 37.3 dB as the blur strengthens, causing the residual to decrease. In contrast, AE-PoS excludes the blur from its learned target set, producing increasingly visible residuals and a PSNR decrease from 43.5 to 31.1 dB. All residual panels use the same intensity gain. On the corresponding clean image, the reconstruction PSNRs are 62.4, 27.6, and 43.9 dB for AE, VAE, and AE-PoS, respectively.

design but reduces its encoder from 12 to six Transformer blocks, while retaining the original four-block decoder and standard masking ratio of 75%. After pretraining, the Stage-0 encoder is frozen, and Stage 1 applies a second six-block-encoder, four-block-decoder network of the same form to its latent representation using masked latent reconstruction. The resulting nested encoder therefore has a total depth of 12 blocks, matching the original ST-MEM backbone while distributing learning across two successive stages.

Pretraining data: Both stages use the same unlabelled pretraining pool: 34,808 recordings from Ningbo (Zheng et al., 2022; Pollard et al., 2026; Zheng et al., 2020) and 130,164 recordings from CODE-15 (Ribeiro et al., 2021), giving 164,972 recordings in total. Each signal is high-pass filtered at 0.67,Hz, low-pass filtered at 40,Hz, standardized independently, resampled to 250,Hz, and cropped to 2250 samples. Stage 1 is trained only on representations of this same pretraining pool, so no additional ECG data are introduced after Stage 0.

Pretraining: For both stages, we use AdamW with learning rate 3.75×10^{-5}, weight decay 0.05, and $(\beta_1, \beta_2) = (0.9, 0.95)$. Training uses a batch size of 64, gradient clipping at norm 1.0, and 150,000 update steps. The learning rate is warmed up linearly for 15,000 steps and then follows cosine decay to 10^{-6}.

Downstream protocol: We use the data splits and frozen-encoder linear-probe protocol of Liu et al. (2024). All pretrained parameters are frozen, and only a single linear classification layer is optimized using binary cross-entropy and Adam with learning rate 10^{-3}, weight decay 10^{-4},



Figure 8: **Nominal fidelity and degradation rejection.** For each input s, the figure shows the reconstruction $\hat{s}$, its PSNR, and the absolute residual $|s - \hat{s}|$ for AE, VAE, and AE-PoS. From top to bottom, the examples are a nominal Kodak24 image, a nominal GoPro frame, its motion-blurred counterpart, and a rainy Test2800 image. AE reconstructs both nominal and degraded inputs almost identically, providing high fidelity but weak anomaly separation. VAE loses substantial detail on nominal images and represents the motion-blurred input more accurately. AE-PoS preserves nominal image content while excluding motion blur and rain from its learned target set, thereby exposing these degradations in the residual. All residual panels use the same intensity gain.

batch size 16, and 100 training epochs. The learning rate is multiplied by 0.1 after epoch 40. The checkpoint with the highest validation AUROC is selected, and its test AUROC is reported. Table 7 uses the complete labelled training split and reports the mean over five seeds. The downstream tasks are the PTB-XL superclass, subclass, form, and rhythm tasks (Wagner et al., 2020), together with ICBEB/CPSC-2018 (Perez Alday et al., 2022).

Results: As shown in Table 7, Stage 1 improves upon Stage 0 on all five downstream tasks and obtains the highest AUROC in every column among the compared methods. Because the two-stage encoder has the same total depth as the original eight-layer ST-MEM backbone, this comparison indicates that the gain arises from successive refinement of the frozen Stage-0 representation rather than from increasing total encoder depth alone.

G Expanded Limitations and Future Directions

The first limitation is that the geometry is defined relative to the reconstruction metric d, which we fix to pixel ℓ_2 distance to isolate the representation-learning question. Consequently, anomalies that remain close to the nominal union under this metric can be difficult to distinguish. Digit 9 illustrates this limitation in the $\{4, 5, 7\}$ experiment, where 87% of its samples are detected by neither score, while the gap between AE-PoS and transformation-based methods on CIFAR-100 indicates that this issue becomes more pronounced for semantically complex data. Learning task-appropriate metrics or deriving anomaly scores directly from the multilevel geometry, while preserving the metric conditions required by Lemma 1, is therefore an important direction for future work. The second limitation is that the number of nested carving levels is currently selected manually. The nested range relation suggests a label free stopping criterion based on the fraction of perturbed held out nominal samples whose reconstructions continue to violate the Push margin at level ℓ. Whether this quantity can reliably determine when the current range requires further carving, without access to anomaly examples, remains the main open question for automatically constructing the nested architecture. The present analysis characterizes representation dynamics relative to nominal data, for which nonlinear orthogonal projection onto the learned union $\mathcal{M}$ provides the ideal representation. For anomalous inputs outside the reach of $\mathcal{M}$, the nearest nominal point need not be unique, and direct projection onto $\mathcal{M}$ therefore does not generally define an ideal representation. We conjecture that important families of structured anomalies, such as specific ECG abnormalities, may lie on or near isometric images $\{\Phi_\gamma(\mathcal{M})\}_{\gamma\in\Gamma}$ of the nominal union. Future work will investigate conditions under which an identifiable inverse transformation Φ_γ^{-1} maps each anomalous input back to or within the reach of $\mathcal{M}$, so that $P_\mathcal{M} \circ \Phi_\gamma^{-1}$ provides a unique canonical representation and $\Phi_\gamma \circ P_\mathcal{M} \circ \Phi_\gamma^{-1}$ realizes nonlinear orthogonal projection onto the corresponding structured anomalous manifold $\Phi_\gamma(\mathcal{M})$.