

Noise & Pattern: Identity-Anchored Tikhonov Regularization for Robust Structural Anomaly Detection

Alexander Bauer^{1,2}

¹Machine Learning Group, TU Berlin

²BIFOLD, Berlin, Germany

alexander.bauer@tu-berlin.de

Klaus-Robert Müller^{1,2,3,4}

³Dept. of AI, Korea University, Seoul, Republic of Korea

⁴MPI for Informatics, Saarbrücken, Germany

klaus-robert.mueller@tu-berlin.de

Abstract

Anomaly detection plays a pivotal role in automated industrial inspection, aiming to identify subtle or rare defects in otherwise uniform visual patterns. As collecting representative examples of all possible anomalies is infeasible, we tackle structural anomaly detection using a self-supervised autoencoder that learns to repair corrupted inputs. To this end, we introduce a corruption model that injects artificial disruptions into training images to mimic structural defects. While reminiscent of denoising autoencoders, our approach differs in two key aspects. First, instead of unstructured i.i.d. noise, we apply structured, spatially coherent perturbations that make the task a hybrid of segmentation and inpainting. Second, and counterintuitively, we add and preserve Gaussian noise on top of the occlusions, which acts as a Tikhonov regularizer anchoring the Jacobian of the reconstruction function toward identity. This identity-anchored regularization stabilizes reconstruction and further improves both detection and segmentation accuracy. On the MVTec AD benchmark, our method achieves state-of-the-art results (I/P-AUROC: 99.9/99.4), supporting our theoretical framework and demonstrating its practical relevance for automatic inspection.

1. Introduction

Anomaly Detection (AD) aims to identify patterns that deviate from a defined notion of normality—a task of considerable importance in domains such as industrial quality control, medical imaging, and beyond [1, 9, 10, 15, 18, 21, 25, 27, 27, 36, 39, 41, 47, 49–51, 55, 58, 60, 61, 72]. A wide range of approaches has been explored for AD. Early methods rely on classical unsupervised techniques such as Principal Component Analysis (PCA) [26, 28, 52], One-Class Support Vector Machines (OC-SVM) [53], Support Vector Data Description (SVDD) [56], nearest-neighbor algorithms [35, 46], and Kernel Density Estimation (KDE) [42].

In recent years, research has shifted toward deep learning-based models, including autoencoders [16, 18, 22, 31, 32, 45, 58, 69, 73], deep one-class classifiers [23, 33, 48], generative approaches [50, 51], and self-supervised methods [24, 25, 37, 44, 54, 61, 66, 67].

In this work, we focus on AD in static images, which is typically addressed at two levels: (a) *image-level* detection, formulated as a binary classification problem, and (b) *pixel-level* detection, formulated as a segmentation task to localize anomalous regions. Specifically, we consider *structural anomalies*, which manifest as localized disruptions in texture or form (e.g., scratches, deformations, or material defects). Due to their spatially local nature at the pixel level, structural anomalies are well-suited for

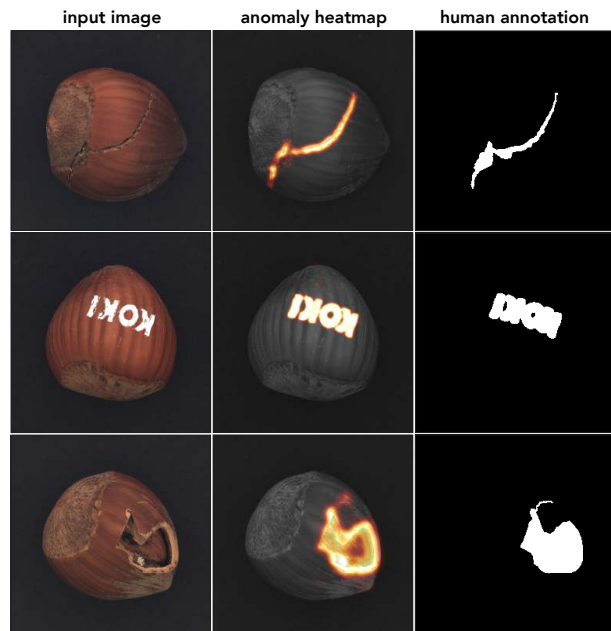


Figure 1. **Robust Structural Anomaly Detection in a Nutshell.** Our qualitative results on MVTec AD (hazelnut): input images, predicted heatmaps, and ground-truth masks.

reconstruction-based methods. In contrast, *logical anomalies* refer to semantic inconsistencies at the scene level (e.g., incorrect object counts or implausible spatial relations), which are global and not uniquely localizable, and thus fall outside the scope of reconstruction-based approaches [12].

A prominent line of AD research builds upon neural image completion [13, 29, 38, 43, 64, 65], where models are trained in a self-supervised manner to reconstruct missing or altered image regions. Representative examples include LSR [61], RIAD [66], CutPaste [37], InTra [44], DRAEM [67], and SimpleNet [40]. Despite architectural differences, these approaches share the common objective of restoring normal content from corrupted inputs.

Our work follows this reconstruction paradigm but extends it with a more expressive *corruption model*. Previous approaches simulated anomalies ranging from simple patch removal [66] to Perlin noise-based texture replacements [67], yet offered only limited structural diversity. In contrast, our model spans multiple dimensions of variation—*texture*, *geometric shape*, and *opacity*—to generate a broader range of structural perturbations. This diversity (paired with noise regularization described below) is crucial for generalization to real-world anomalies.

We train an autoencoder to repair such artificially corrupted images by minimizing the reconstruction loss between clean and corrupted counterparts. A key element of our approach is the contrasting interplay between persistent structural deviations and Gaussian noise applied across the entire image. This combination not only regularizes training but also enhances robustness in anomaly localization. It further improves detection accuracy by increasing reconstruction fidelity in both corrupted and normal regions, enabling the model to disentangle the true structure of normal patterns from unstructured pixel-wise variations.

This noise component also admits a clear theoretical interpretation: it acts as a Tikhonov-like regularization that anchors the Jacobian of the reconstruction function toward the identity. As a result, we obtain a *Filtering Autoencoder* (FAE) that removes anomalous patterns while preserving normal variability, thereby enabling robust and accurate anomaly detection. Figure 1 provides a visual illustration of the detection performance of our method.

The remainder of the paper is organized as follows. Section 2 introduces the training framework, Section 3 provides the theoretical justification linking our approach to identity-anchored Tikhonov regularization, Section 4 details the corruption model, and Section 5 presents experimental results. We conclude in Section 6.

2. Methodology

We model the autoencoder as a parameterized function

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

with learnable parameters θ and input-output tensors corresponding to RGB images of spatial size $h \times w$. Let $\mathbf{x} \in [0, 1]^{h \times w \times 3}$ denote an anomaly-free training image. A corrupted variant $\hat{\mathbf{x}}$ is created by modifying selected regions, which are indicated by a real-valued mask $\mathbf{M} \in [0, 1]^{h \times w \times 3}$. Its complement is written as $\bar{\mathbf{M}} := \mathbf{1} - \mathbf{M}$, where $\mathbf{1}$ is the all-ones tensor. Note that $\mathbf{M}$ is not binary and takes on values from a continuous range.

Training. We start with a simpler objective that will later be modified. Following [7], we minimize a loss $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}, \mathbf{M}; \theta)$ over the model weights θ defined as

$$\frac{1 - \lambda}{\|\bar{\mathbf{M}}\|_1} \|\bar{\mathbf{M}} \odot (f_{\theta}(\hat{\mathbf{x}}) - \mathbf{x})\|_2^2 + \frac{\lambda}{\|\mathbf{M}\|_1} \|\mathbf{M} \odot (f_{\theta}(\hat{\mathbf{x}}) - \mathbf{x})\|_2^2,$$

where $\odot$ denotes elementwise multiplication, $\|\cdot\|_p$ the standard ℓ^p -norm on tensors, and $\lambda \in [0, 1]$ is a hyperparameter.

The above objective is divided into two terms to penalize prediction errors on corrupted and uncorrupted regions differently. For simplicity, we omit these weights in the following discussion and adopt the simplified form:

$$\mathcal{L}_{\text{DAE}}(\hat{\mathbf{x}}, \mathbf{x}; \theta) = \|f_{\theta}(\hat{\mathbf{x}}) - \mathbf{x}\|_2^2, \quad (1)$$

which resembles the denoising autoencoder (DAE) [59].

Test-time detection. Once the autoencoder f_{θ} is trained, anomaly maps can be derived from the discrepancy between an input $\hat{\mathbf{x}}$ and its reconstruction $f_{\theta}(\hat{\mathbf{x}})$. Following [7], we define this pixel-based discrepancy as a mapping

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

There are multiple options for Δ including pixel-wise Mean Squared Error (MSE), Structural Similarity Index Measure (SSIM) [9] and Gradient Magnitude Similarity (GMS) [63]. Thresholding a corresponding output yields a binary segmentation of anomalies.

To enhance robustness, we smooth the difference map $\Delta(\cdot, \cdot)$ before thresholding. Specifically, we define

$$\text{anomap}_{f_{\theta}}^{n,k}(\hat{\mathbf{x}}) := G_k^n(\Delta(\hat{\mathbf{x}}, f_{\theta}(\hat{\mathbf{x}}))), \quad (2)$$

where G_k is a mean filter of size $k \times k$ with entries $1/k^2$, and G_k^n denotes n repeated applications (with G_k^0 being the identity). Hyperparameters $n, k \in \mathbb{N}$ control the amount of smoothing. Applying a threshold to $\text{anomap}_{f_{\theta}}^{n,k}(\hat{\mathbf{x}})$ then produces the anomaly segmentation mask.

An image-level anomaly score is obtained by aggregating pixel values of $\text{anomap}_{f_{\theta}}^{n,k}(\hat{\mathbf{x}})$, typically via summation, though alternative reductions such as the maximum can be used to reduce sensitivity to anomaly size. The full detection workflow (based on [7]) is illustrated in Figure 2.

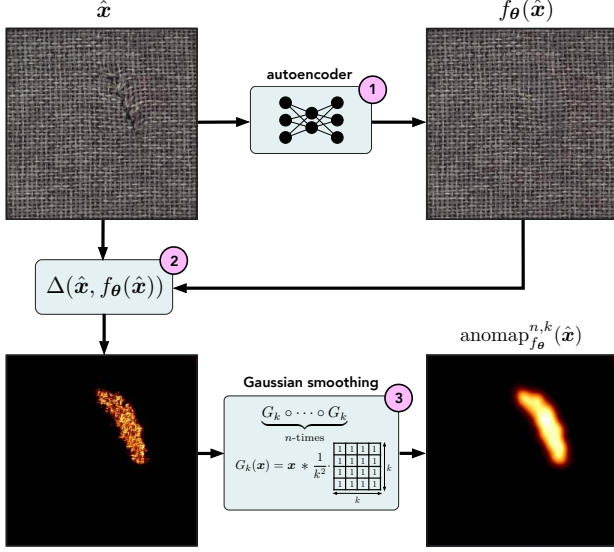


Figure 2. Illustration of our AD process. Given input $\hat{x}$, our model produces an output $f_\theta(\hat{x})$, which replicates normal regions and replaces irregularities with locally consistent patterns. Then we compute a spatial difference map $\Delta(\hat{x}, f_\theta(\hat{x})) \in \mathbb{R}^{h \times w}$. In the last step we apply a series of averaging convolutions G_k to the difference map to produce final anomaly heatmap $\text{anomap}_{f_\theta}^{n,k}(\hat{x})$.

Filtering Autoencoder. To this end, we propose a new type of regularized autoencoder by modifying the objective in (1) to include isotropic Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$:

$$\mathcal{L}_{\text{FAE}}(\hat{x}, x, \epsilon; \theta) = \|f_\theta(\hat{x} + \epsilon) - (x + \epsilon)\|_2^2, \quad (3)$$

where x is a normal sample, $\hat{x}$ its corrupted version, and ϵ Gaussian noise. That is, for each pair $(\hat{x}, x)$, we add small-variance noise to both input and target, encouraging the model output $f_\theta(\hat{x} + \epsilon)$ to match the perturbed normal sample $x + \epsilon$. This seemingly minor modification introduces an effective regularization otherwise absent from (1), as discussed in detail in the next section.

We refer to the model trained with (3) as the *Filtering Autoencoder* (FAE). The name reflects its inductive bias: the model preserves normal regions (identity on the manifold) while filtering out structural anomalies. Theoretically, this behavior can be viewed as a form of conservative projection [7] onto the data manifold, but we use the term *filtering* to emphasize its practical role in anomaly detection.

3. Identity-Anchored Tikhonov Regularization

Self-supervised autoencoders detect structural anomalies by learning to restore artificially corrupted regions while leaving normal regions unchanged. We found that detection performance improves further when the model is additionally trained to preserve small-variance Gaussian noise applied across the entire image after introducing such corruptions.

At first, this may appear counterintuitive, since our strategy resembles a DAE, which is trained to remove rather than preserve noise. However, exposing the model simultaneously to structured corruptions and unstructured Gaussian noise in a complementary manner encourages it to disentangle these effects, thereby enhancing its ability to recognize genuine structural deviations.

A related idea is well established in supervised learning. For image classification, for example, augmenting the data with Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$ leads to the training objective

$$\mathbb{E}_\epsilon [\|f(x + \epsilon) - y\|^2].$$

Bishop [14] showed that, up to second order in σ^2 , this objective is equivalent to augmenting the clean objective with an additional regularization term according to

$$\|f(x) - y\|^2 + \sigma^2 \|D_x f\|_F^2, \quad (4)$$

which is the familiar Tikhonov penalty form. This penalty enforces local insensitivity to small input perturbations, stabilizing the mapping and improving generalization.

In our case, Gaussian noise serves a similar stabilizing role, but with a crucial and *novel* modification: it is added symmetrically to both the input and the target during training. This formulation distinguishes the FAE from conventional denoising or contractive autoencoder variants and yields a clear theoretical interpretation, which we formalize in the following theorem. With a slight abuse of notation, we write $\mathcal{O}(\sigma^4)$ to denote the remainder $R \in \mathcal{O}(\sigma^4)$ of the Taylor expansion.

Theorem 1 (Identity-Anchored Tikhonov Regularization). *Let $(\hat{x}, x) \sim \mathbb{P}_{\hat{x}, x}$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$ be independent. Assume $f_\theta \in C^3$ in a neighborhood of $\hat{x}$ almost surely, and $D^3 f_\theta$ is locally Lipschitz. Let $\mathbf{r} := f_\theta(\hat{x}) - x$, $J_\theta(\hat{x}) := D_{\hat{x}} f_\theta$ denote the Jacobian, and Δf_θ the component-wise Laplacian of f_θ . Define*

$$\mathcal{L}(\sigma; \theta) := \mathbb{E}_{(\hat{x}, x), \epsilon} [\|f_\theta(\hat{x} + \epsilon) - (x + \epsilon)\|_2^2]. \quad (5)$$

Then, as $\sigma \rightarrow 0$, the total loss $\mathcal{L}(\sigma; \theta)$ admits the asymptotic expansion

$$\mathbb{E} [\|\mathbf{r}\|_2^2 + \sigma^2 (\|J_\theta(\hat{x}) - I_d\|_F^2 + \mathbf{r}^\top \Delta f_\theta(\hat{x}))] + \mathcal{O}(\sigma^4), \quad (6)$$

provided that $\mathbb{E} \|\mathbf{r}\|_2^2, \mathbb{E} \|J_\theta(\hat{x})\|_F^2, \mathbb{E} \|D_{\hat{x}}^2 f_\theta\|_F^2 < \infty$. Here, the expectation $\mathbb{E}$ is taken with respect to $(\hat{x}, x)$.

Consider the curvature term $\mathbf{r}^\top \Delta f$ in (6). First note that for locally affine mappings, this term vanishes since the Hessian of f is zero. This includes, for instance, piecewise-linear architectures such as autoencoders with ReLU activations, provided the evaluation point lies within a single linear region. In such cases, the expansion simplifies to

$$\mathbb{E}_{\hat{x}, x} [\|f_\theta(\hat{x}) - x\|_2^2 + \sigma^2 \|J_\theta(\hat{x}) - I_d\|_F^2] + \mathcal{O}(\sigma^4). \quad (7)$$

When a nonlinear output activation is used (e.g., a sigmoid in the final layer), the Hessian no longer vanishes, leaving the term $\mathbf{r}^\top \Delta f$. However, it can be uniformly bounded by a constant multiple of $\|\mathbf{r}\| \cdot \|J(\hat{\mathbf{x}})\|_F^2$, where the constant depends only on the activation. It is therefore proportional to the residual norm $\|\mathbf{r}\|$ and becomes negligible as the training error approaches zero. Under these favourable conditions, Theorem 1 implies that the loss in (5) admits an asymptotic expansion approximating (7). As the remainder $\mathcal{O}(\sigma^4)$ vanishes rapidly for $\sigma \rightarrow 0$, the dominant regularization effect reduces to the Frobenius term $\|J_\theta(\hat{\mathbf{x}}) - I_d\|_F^2$, supporting the interpretation of additive Gaussian noise training as a Tikhonov regularizer that promotes local identity preservation.

3.1. Complementary Effects of the Jacobian Penalty

It is instructive to contrast the objective in (7) with Bishop’s original result in (4), where training with noisy inputs and clean targets yields the regularizer $\|J_\theta(\mathbf{x})\|_F^2$. This penalty pulls the Jacobian toward zero, promoting local flatness and insensitivity to perturbations—an appropriate inductive bias for classification. In our reconstruction setting, however, such a flatness prior is in tension with the need to faithfully reproduce inputs along the data manifold. Here, the identity-anchored penalty $\|J_\theta - I_d\|_F^2$ adds several complementary effects to the objective summarized below.

Reconstruction of normal regions. A standard autoencoder trained only with reconstruction loss can potentially memorize training patterns, yielding spurious mappings that merely interpolate between training examples. The regularizer $\|J_\theta(\hat{\mathbf{x}}) - I_d\|_F^2$ counteracts this by normalizing the gradients of f in tangent directions of the data manifold. This reduces local collapse (vanishing gradients) or overstretching (exploding gradients) and thereby stabilizes reconstructions of unseen normal data. Moreover, if two C^1 -mappings agree on a manifold, their Jacobians must also agree in tangent directions. This penalty encourages this necessary local condition for f to behave as the identity on the manifold. As a result, the reconstruction of fine-grained details in normal regions becomes more accurate, as illustrated in Figure 3.

Idempotency. Geometrically, our autoencoder is designed to project corrupted inputs back onto the manifold of normal samples $\mathcal{M} \subset \mathbb{R}^d$ while acting as the identity on $\mathcal{M}$. Together, these requirements imply that f should be idempotent, i.e. $f(f(\hat{\mathbf{x}})) = f(\hat{\mathbf{x}})$. To analyze how the regularizer promotes idempotency, we expand f at $\mathbf{x} \in \mathcal{M}$ and evaluate at $f(\hat{\mathbf{x}})$ where $\hat{\mathbf{x}}$ is a corrupted version of $\mathbf{x}$. Since the penalty terms pulls the Jacobian toward identity, assuming the extreme case $J_\theta(\mathbf{x}) = I_d$ yields

$$\begin{aligned} f(f(\hat{\mathbf{x}})) &= f(\mathbf{x}) + J_\theta(\mathbf{x})(f(\hat{\mathbf{x}}) - \mathbf{x}) + R_2 \\ &= f(\hat{\mathbf{x}}) + (f(\mathbf{x}) - \mathbf{x}) + R_2, \end{aligned}$$

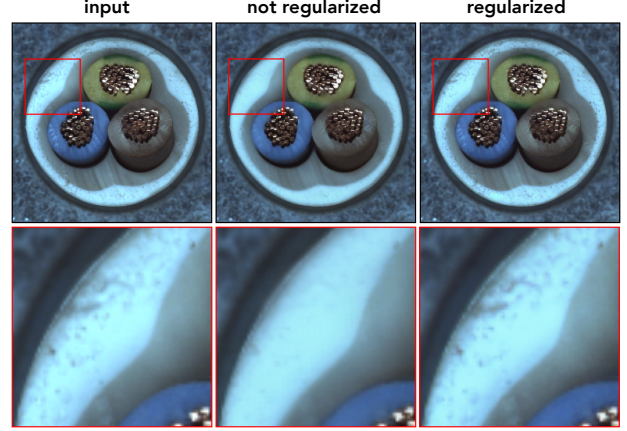


Figure 3. Comparison of reconstruction quality between our models trained with and without Gaussian noise. The first row shows an input image $\mathbf{x}$ and its reconstructions $f(\mathbf{x})$ from both models. The second row provides zoom-ins of the region marked by a red square. Reconstructions in normal regions are visibly **improved** when the model is **regularized** with Gaussian noise.

where R_2 is the second-order Taylor remainder. If the Jacobian is L -Lipschitz along the segment from $\mathbf{x}$ to $f(\hat{\mathbf{x}})$, then $\|R_2\| \leq L\|f(\hat{\mathbf{x}}) - \mathbf{x}\|^2$. Hence,

$$f(f(\hat{\mathbf{x}})) = f(\hat{\mathbf{x}}) + \underbrace{(f(\mathbf{x}) - \mathbf{x})}_{\text{bias on } \mathcal{M}} + \underbrace{\mathcal{O}(\|f(\hat{\mathbf{x}}) - \mathbf{x}\|^2)}_{\text{reconstruction error}}. \quad (8)$$

That is, equation (8) directly links idempotency to the reconstruction accuracy and to how well f approximates the identity on the manifold.

Correction of anomalous regions. The Jacobian penalty enhances the correction of anomalous regions by balancing the influence of the reconstruction and Jacobian terms. While the reconstruction loss dominates in directions orthogonal to the manifold, enforcing projection back to it, the Jacobian term normalizes gradients along tangent directions, preventing *mis-scaling*: if $J_\theta(\hat{\mathbf{x}}) \approx \alpha I_d$ with $\alpha < 1$, tangent variations are overly contracted and corruptions persist in smeared form, whereas for $\alpha > 1$ they are expanded and may be amplified. Only when $\alpha \approx 1$ do tangent directions behave like the identity, preserving normal structure while allowing orthogonal deviations to be corrected by the reconstruction term. This interaction compels the model to separate noise-like tangent perturbations from true structural anomalies, leading it to preserve normal variations while suppressing off-manifold deviations. Consequently, reconstructions become more stable and precise, as illustrated in Figure 4.

3.2. Connection to DAE and RCAE

An interesting question is how our regularization relates to earlier variants of regularized autoencoders that combine reconstruction and Jacobian penalties. A prominent example

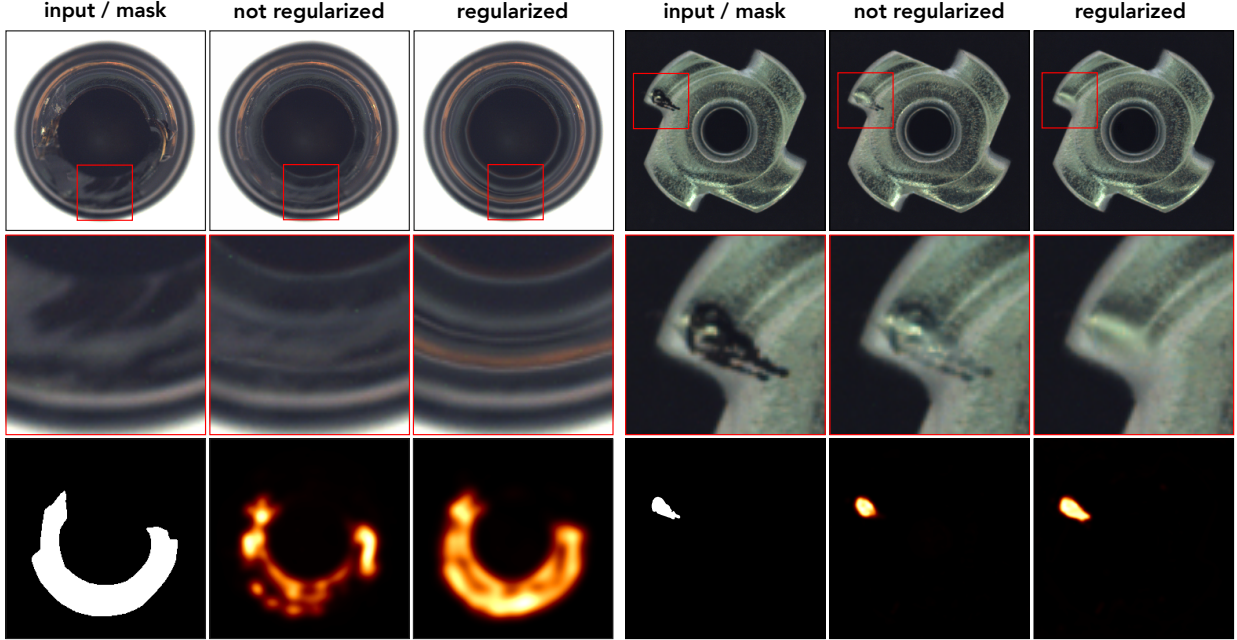


Figure 4. Comparison of reconstruction quality between our models trained with and without Gaussian noise. Each example shows a defective input $\hat{x}$ (top row), its reconstructions $f(\hat{x})$ from both models, and zoom-ins (middle row) of the region marked by a red square for detailed comparison. The bottom row shows the ground-truth anomaly mask and corresponding heatmaps. Reconstructions in corrupted regions and the corresponding anomaly heatmaps are visibly **improved** when the model is **regularized** with Gaussian noise.

is the Reconstruction Contractive Autoencoder (RCAE) [2], defined as

$$\mathcal{L}_{\text{RCAE}}(x) = \|f_{\theta}(x) - x\|^2 + \lambda \|J_{\theta}(x)\|_F^2,$$

which minimizes a reconstruction loss on clean inputs together with a contractive penalty that drives the Jacobian toward zero. This induces an anisotropic effect: gradients remain non-zero along tangent directions for faithful reconstruction, but are suppressed in normal directions to reject off-manifold perturbations. As shown in [2], RCAE connects to the denoising autoencoder (DAE) [59], since training with Gaussian noise implicitly adds a contractive Jacobian term (Bishop’s theorem).

Our approach differs in that it anchors the Jacobian to the identity, $J_{\theta} \approx I_d$, and evaluates it at corrupted inputs $\hat{x}$. Rather than enforcing flatness, this constraint promotes identity-like behavior along the manifold and stabilizes repairs off it. In practice, this reduces the tendency toward high-frequency attenuation induced by strong Jacobian penalties, yielding sharper reconstructions and more reliable detection of structural anomalies. Moreover, while RCAE requires explicit Jacobian computation, our formulation achieves the effect by simply adding Gaussian noise to inputs and outputs during training.

Finally, the asymptotic expansion in (7) shows that a fixed noise variance σ^2 sets the strength of the identity-anchoring term to $\sigma^2 \|J_{\theta}(\hat{x}) - I_d\|_F^2$. In practice, this biases

the model toward the noise level seen during training. Varying σ during training (e.g., $\sigma \sim \text{Uniform}(0, r)$) mitigates this bias and improves robustness to noise-level variation.

4. Corruption Model

Training a standard autoencoder on normal samples alone yields poor anomaly detection performance, as the reconstruction objective is indifferent to whether inputs are normal or anomalous, rewarding only visual fidelity. Consequently, such models often reproduce normal and anomalous structures with comparable accuracy. Reducing capacity through bottleneck compression or sparsity does not alter the objective and thus cannot resolve this limitation.

AD is inherently a binary classification task, yet its unsupervised formulation arises due to absence of labeled anomalous samples. We show that real anomalies are not strictly required: training on artificially corrupted data is sufficient when the corruption model captures essential aspects of structural deviations. Such distortions can be generated in a simple and fully self-supervised manner, enabling scalable and domain-agnostic training, as demonstrated in our experiments.

We distill prior approaches [7, 20, 30, 37, 57, 67] into a compact set of simple design principles enabling models trained on artificial corruptions to generalize to real defects. We argue that artificial anomalies should be *partial* and vary

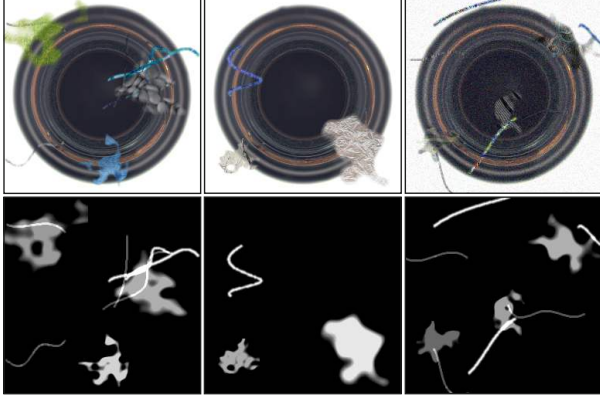


Figure 5. Illustration of a combined realization of our design principles for simulating structural anomalies, covering variations in shape, texture, and occlusion opacity. The first row shows partially corrupted images, and the second row displays the corresponding anomaly masks, where mask intensity reflects the opacity of the occlusions. In the third example, Gaussian noise is applied across the entire image as part of our regularization strategy.

along three dimensions: *shape*, *texture*, and *opacity*. Figure 5 illustrates their joint realization.

4.1. Design Principles: Shape

A crucial requirement is that the artificial occlusions used for training remain *partial*, affecting only localized regions rather than the entire image. In a classical DAE, noise applied to the whole input often produces blurred or averaged reconstructions—even in uncorrupted regions. This occurs because the injected noise perturbs both orthogonal and tangent directions of the data manifold, causing the model to average over multiple plausible reconstructions. Partial distortions create a stronger contrast between normal and corrupted regions, allowing the model to focus on accurately reconstructing intact areas and thereby improving its discriminative ability between normal and anomalous regions.

Remark 1. *Training with partial, localized occlusions of varying sizes—rather than global corruptions—preserves the fidelity of reconstructed normal regions while enhancing the model’s sensitivity to anomalous structures.*

In our implementation, occlusion shapes are generated by applying elastic deformation to ellipsoidal binary masks, following the approach of [7]. While this yields diverse shapes, they exhibit limited local geometric variation, with nearly constant cross-sectional thickness. To address this limitation, we superimpose additional curve-like structures of varying length, generated independently of the base mask. Combining these two extremes—broad, bulky forms and thin, elongated structures—proved sufficient to achieve strong performance. Notably, incorporating thin, curve-like masks substantially improved detection accuracy for

several categories of the MVTec dataset. Figures 7 in the supplement show qualitative improvements in the anomaly heatmaps for several examples.

Remark 2. *Occlusion masks should span the full spectrum of geometric variability, encompassing both diverse outer boundaries and internal shape characteristics—ranging from broad, bulky forms to thin, elongated shapes.*

4.2. Design Principles: Texture Pattern

In general, the range of texture variations used for corruption should be sufficiently diverse to ensure reliable detection of anomalous patterns. At the same time, artificial occlusions risk introducing non-natural artifacts that can lead to overfitting. A natural and effective way to mitigate this risk is to use real textures as anomaly fillers. In our implementation, we sample patches from the publicly available texture background dataset [19] and insert them into training images according to generated shape masks. Optional post-processing steps, such as boundary smoothing or blending, further reduce artifacts while preserving control over the geometry of occlusions.

Remark 3. *Artificial occlusions must balance diversity and naturalness: rich texture variation promotes generalization, whereas adherence to natural image statistics prevents spurious artifacts.*

4.3. Design Principles: Opacity

Fully opaque occlusions maximally promote context-based reconstruction, as the model must infer the missing content solely from the surrounding regions. Reducing opacity, by contrast, partially reveals the underlying signal, encouraging more subtle correction mechanisms and acting as an implicit regularizer. Both effects are complementary: opaque occlusions emphasize global consistency, while transparent ones promote fine-grained fidelity. Therefore, a well-designed corruption model should include both.

Moreover, real-world data support this distinction, as actual anomalies occur in both forms. For example, the *bottle* category of the MVTec dataset contains contaminations that fully occlude the background, whereas the *glue strip* class of the *tile* category represents an example of partially transparent anomalies.

Remark 4. *A corruption model should incorporate both fully opaque and partially transparent occlusions. Opaque occlusions compel the model to reconstruct missing regions from context, whereas transparent ones regularize the learning process by enabling partial signal propagation.*

5. Experiments

In this section, we present results demonstrating the effectiveness of our approach for robust structural anomaly detection. We train a Filtering Autoencoder (FAE) using our

Table 1. Experimental results for **anomaly recognition** measured with **I-AUROC** on the MVTec AD dataset.

Category	AnoGAN	VAE	LSR	RIAD	CutPaste	InTra	DRAEM	SimpleNet	PatchCore	RD++	MSFlow	PNI	RealNet	SDC	GLASS	FAE
	2017 – 2020			2020 – 2022			2022 – 2025			<i>Ours</i>						
carpet	49	78	71	84.2	93.1	98.8	97.0	99.7	98.2	100	100	100	99.8	100	99.8	100
grid	51	73	91	99.6	99.9	100	99.9	99.7	98.3	100	99.8	98.4	100	100	100	100
leather	52	95	96	100	100	100	100	100	100	100	100	100	100	100	100	100
tile	51	80	95	93.4	93.4	98.2	99.6	99.8	98.9	99.7	100	100	100	100	100	100
wood	68	77	96	93.0	98.6	97.5	99.1	100	99.9	99.3	100	99.6	99.2	100	99.9	100
avg. textures	54.2	80.6	89.8	94.0	97.0	98.9	99.1	99.8	99.0	99.8	99.9	99.6	99.8	100	99.9	100
bottle	69	87	99	99.9	98.3	100	99.2	100	100	100	100	100	100	100	100	100
cable	53	90	72	81.9	80.6	70.3	91.8	99.9	99.7	99.2	99.5	99.8	99.2	99.9	99.8	99.9
capsule	58	74	68	88.4	96.2	86.5	98.5	97.7	98.1	99.0	99.2	99.7	99.6	98.8	99.9	99.9
hazelnut	50	98	94	83.3	97.3	95.7	100	100	100	100	100	100	100	98.8	100	100
metal nut	50	94	83	88.5	99.3	96.9	98.7	100	100	100	100	100	99.8	100	100	100
pill	62	83	68	83.8	92.4	90.2	98.9	99.0	97.1	98.4	99.6	96.9	99.1	100	99.4	100
screw	35	97	80	84.5	86.3	95.7	93.9	98.2	99.0	98.9	97.8	99.5	98.8	99.6	100	100
toothbrush	57	94	92	100	98.3	100	100	99.7	98.9	100	100	99.7	99.4	98.9	100	100
transistor	67	93	73	90.9	95.5	95.8	93.1	100	99.7	98.5	100	100	100	100	99.9	100
zipper	59	78	97	98.1	99.4	99.4	100	99.9	99.7	98.6	100	99.9	99.8	100	100	100
avg. objects	56.0	88.8	82.6	89.9	94.4	93.1	97.4	99.5	99.2	99.3	99.6	99.5	99.6	99.7	99.9	99.9
avg. all	55.4	86.1	85.0	91.3	95.2	95.0	98.0	99.6	99.2	99.4	99.7	99.6	99.7	99.8	99.9	99.9

Table 2. Experimental results for **anomaly segmentation** measured with **P-AUROC** on the MVTec AD dataset.

Category	AnoGAN	VAE	LSR	RIAD	CutPaste	InTra	DRAEM	SimpleNet	PatchCore	RD++	MSFlow	PNI	RealNet	SDC	GLASS	FAE
	2017 – 2020			2020 – 2022			2022 – 2025			<i>Ours</i>						
carpet	54	78	94	96.3	98.3	99.2	95.5	98.2	98.7	99.2	99.4	99.4	99.2	99.8	99.6	99.8
grid	58	73	99	98.8	97.5	98.8	99.7	98.8	98.8	99.3	99.4	99.2	99.5	99.8	99.4	99.8
leather	64	95	99	99.4	99.5	99.5	98.6	99.2	99.3	99.4	99.7	99.6	99.8	99.7	99.8	99.8
tile	50	80	88	89.1	90.5	94.4	99.2	97.0	96.3	96.6	98.2	98.4	99.4	99.2	99.7	99.4
wood	62	77	87	85.8	95.5	88.7	96.4	94.5	95.2	95.8	97.1	97.0	98.2	98.4	98.8	98.6
avg. tex.	57.6	80.6	93.4	93.9	96.3	96.1	97.9	97.5	97.7	98.1	98.8	98.7	99.2	99.4	99.5	99.5
bottle	86	87	95	98.4	97.6	97.1	99.1	98.0	98.6	98.8	99.0	98.9	99.3	98.9	99.3	99.3
cable	86	87	95	94.2	90.0	91.0	94.7	97.6	98.7	98.4	98.5	99.1	98.1	98.5	98.7	98.7
capsule	84	74	93	92.8	97.4	97.7	94.3	98.9	99.1	98.8	99.1	99.3	99.3	99.1	99.4	99.3
hazelnut	87	98	95	96.1	97.3	98.3	99.7	97.9	98.8	99.2	98.7	99.4	99.7	99.1	99.4	99.6
metal nut	76	94	91	92.5	93.1	93.3	99.5	98.8	99.0	98.1	99.3	99.3	98.6	98.5	99.4	99.5
pill	87	83	91	95.7	95.7	98.3	97.6	98.6	98.6	98.3	98.8	99.0	99.0	99.3	99.5	99.3
screw	80	97	96	98.8	96.7	99.5	97.6	99.3	99.5	99.7	99.1	99.6	99.5	99.7	99.5	99.7
toothbrush	90	94	97	98.9	98.1	98.9	98.1	98.5	98.9	99.1	98.5	99.1	98.7	99.4	99.3	99.5
transistor	80	93	91	87.7	93.0	96.1	90.9	97.6	97.1	0.0	98.3	98.0	98.0	98.9	97.6	98.9
zipper	78	78	98	97.8	99.3	99.2	98.8	98.9	99.0	94.3	99.2	99.4	99.2	99.6	99.6	99.7
avg. obj.	83.4	88.5	94.6	95.3	95.8	96.9	97.0	98.4	98.7	98.8	98.8	99.1	98.9	99.1	99.2	99.4
avg. all	74.8	85.9	94.2	94.8	96.0	96.7	97.3	98.1	98.4	98.3	98.8	99.0	99.0	99.2	99.3	99.4

corruption model, regularized to preserve Gaussian noise.

Dataset. We evaluate on the widely adopted MVTec AD benchmark [8, 11], which contains 5,354 high-resolution images across 15 categories (10 objects and 5 textures) with substantial variability in defect type and appearance. The training set comprises only defect-free images, while the test set includes both defect-free and anomalous images. There are over 70 different types of anomalies including scratches, dents, contaminations, and structural changes (in total 1,888 manually annotated defect regions).

Training setup. To increase the number of training samples, we apply category-dependent augmentations such as random rotations and flips, holding out 5% of the training data for validation. For the *transistor* class, additional 90° rotations were essential to detect missing (or rotated) components, although this made training more challenging.

We train one model per category using a modified U-Net with dilated convolutions in the bottleneck. Training is performed with the Adam optimizer [34] at a learning rate of 10^{-4} . Texture categories are resized to 512×512 , while object categories are downsampled to 256×256 pixels.

Results. We report image-level (I-AUROC) and pixel-level (P-AUROC) performance in Tables 1 and 2, respectively. Following established convention, the values are rounded to one decimal place for compatibility with prior work. For the detection task, our method achieves a total average score of 99.9. For the segmentation task, we obtain a new state-of-the-art result of 99.4, surpassing a range of top-performing methods. We compare against a selection of established and leading baselines, including AnoGAN [50], VAE [39], LSR [61], RIAD [66], Cut-Paste [37], InTra [44], DRAEM [67], SimpleNet [40],

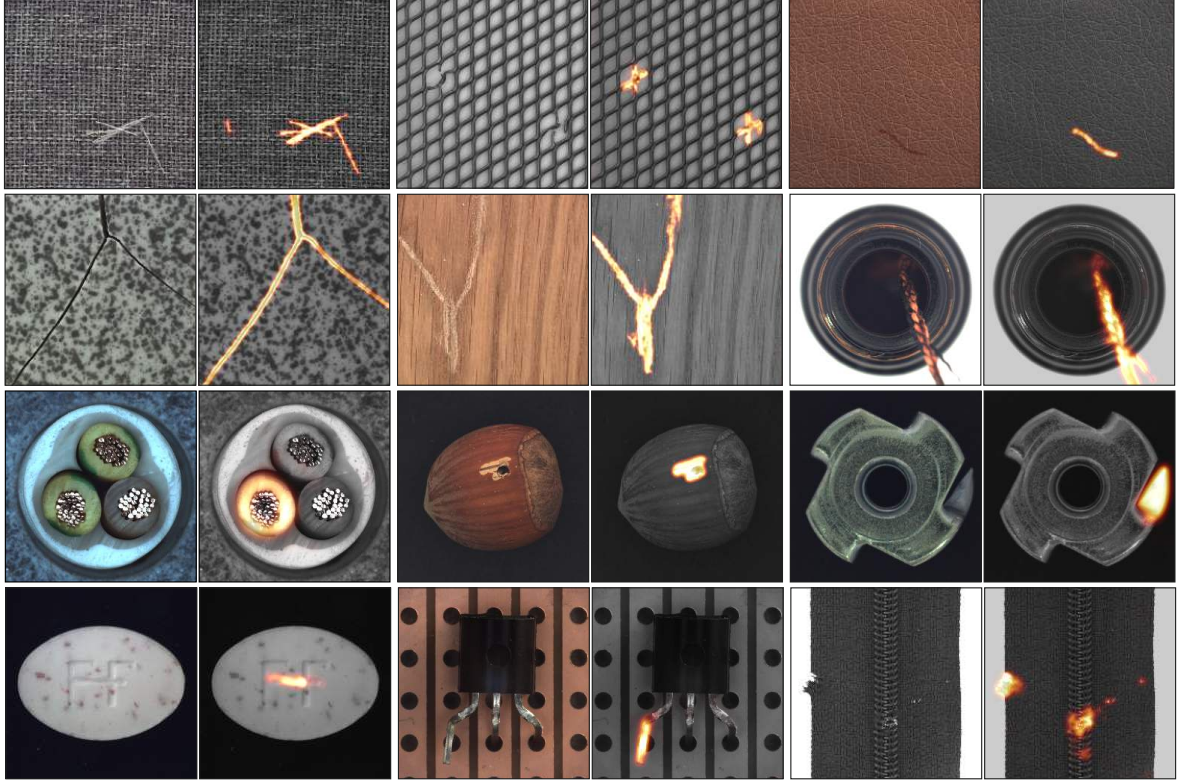


Figure 6. Representative (randomly chosen) examples of anomaly heatmaps on the MVTec AD dataset produced by our model. Each pair shows an input image and its corresponding heatmap overlay highlighting the detected defective regions.

PatchCore [47], RD++ [57] MSFlow [70], PNI [3], Real-Net [68], SDC [7], and GLASS [17]. All comparison scores are taken directly from the respective publications. To illustrate the quality and sharpness of the resulting anomaly heatmaps produced by our model, representative examples are shown in Figure 6. Supplementary material provides additional examples of reconstruction improvements (Figure 8) and resulting anomaly heatmaps (Figures 9 and 10), as well as a direct comparison of heatmap quality with previous methods (Figure 11).

6. Conclusion

We proposed the novel *Filtering Autoencoder* (FAE) for robust structural anomaly detection, trained to correct artificially corrupted samples. Unlike conventional denoising autoencoders, the FAE preserves Gaussian noise, which acts as a Tikhonov regularizer anchoring the Jacobian toward the identity. This stabilizes both reconstruction and anomaly detection, allowing the model to disentangle structured deviations from random fluctuations. Meanwhile, the regularization remains lightweight and method-agnostic, adding negligible computational overhead and integrating seamlessly with other restoration-based approaches.

We further proposed a generic corruption-model tem-

plate that (together with noise regularization) effectively mitigates overfitting to artificial corruption patterns and enhances generalization to real anomalies. On the MVTec AD benchmark, our approach achieves 99.4 P-AUROC, surpassing prior methods. The robustness and accuracy of the FAE make it a strong candidate for industrial inspection.

While effective for structural anomalies, reconstruction-based models alone cannot address logical anomalies involving ambiguous or ill-defined annotations. Future work will explore integrating our approach with constraint optimization [4–6] and reasoning-driven methods [62, 71].

Acknowledgements

This work was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under grant Nr. BIFOLD25B. K.-R.M. was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program, Korea University) and grant funded by the Korea government (MSIT) (No. RS-2024-00457882, AI Research Hub Project).

References

- [1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 481–490. Computer Vision Foundation / IEEE, 2019. [1](#)
- [2] Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *J. Mach. Learn. Res.*, 15(1):3563–3593, 2014. [5](#)
- [3] Jaehyeok Bae, Jae-Han Lee, and Seyun Kim. PNI: industrial anomaly detection using position and neighborhood information. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 6350–6360. IEEE, 2023. [8](#)
- [4] Alexander Bauer, Shinichi Nakajima, and Klaus-Robert Müller. Efficient exact inference with loss augmented objective in structured learning. *IEEE Trans. Neural Netw. Learning Syst.*, 28(11):2566 – 2579, 2017. [8](#)
- [5] Alexander Bauer, Shinichi Nakajima, Nico Görnitz, and Klaus-Robert Müller. Partial optimality of dual decomposition for map inference in pairwise mrfs. In *Proceedings of Machine Learning Research*, pages 1696–1703, 2019.
- [6] Alexander Bauer, Shinichi Nakajima, and Klaus-Robert Müller. Polynomial-time constrained message passing for exact map inference on discrete models with global dependencies. *Mathematics*, 11(12):2628, 2023. [8](#)
- [7] Alexander Bauer, Shinichi Nakajima, and Klaus-Robert Müller. Self-supervised autoencoders for visual anomaly detection. *Mathematics*, 12(24), 2024. [2](#), [3](#), [5](#), [6](#), [8](#)
- [8] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec AD - A comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9592–9600. Computer Vision Foundation / IEEE, 2019. [7](#)
- [9] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2019, Volume 5: VISAPP, Prague, Czech Republic, February 25-27, 2019*, pages 372–380. SciTePress, 2019. [1](#), [2](#)
- [10] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4182–4191. Computer Vision Foundation / IEEE, 2020. [1](#)
- [11] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *Int. J. Comput. Vis.*, 129(4):1038–1059, 2021. [7](#)
- [12] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *Int. J. Comput. Vis.*, 130(4):947–969, 2022. [2](#)
- [13] Anand Bhattad, Jason Rock, and David A. Forsyth. Detecting anomalous faces with ‘no peeking’ autoencoders. *CoRR*, abs/1802.05798, 2018. [2](#)
- [14] Christopher M. Bishop. Training with noise is equivalent to tikhonov regularization. *Neural Comput.*, 7(1):108–116, 1995. [3](#)
- [15] T. Böttger and M. Ulrich. Real-time texture error detection on textured surfaces with compressed sensing. *Pattern Recognit. Image Anal.*, 26:88–94, 2016. [1](#)
- [16] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Robust, deep and inductive anomaly detection. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part I*, pages 36–51. Springer, 2017. [1](#)
- [17] Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXVII*, pages 37–54. Springer, 2024. [8](#)
- [18] Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. *CoRR*, abs/1806.04972, 2018. [1](#)
- [19] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. [6](#)
- [20] Anne-Sophie Collin and Christophe De Vleeschouwer. Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7915–7922, 2021. [5](#)
- [21] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part IV*, pages 475–489. Springer, 2020. [1](#)
- [22] Jun Deng, Zixing Zhang, Erik Marchi, and Björn W. Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013, Geneva, Switzerland, September 2-5, 2013*, pages 511–516. IEEE Computer Society, 2013. [1](#)
- [23] Sarah M. Erfani, Sutharshan Rajasegarar, Shanika Karunasekera, and Christopher Leckie. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognit.*, 58:121–134, 2016. [1](#)
- [24] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*,

- December 3-8, 2018, Montréal, Canada, pages 9781–9791, 2018. 1
- [25] Matthias Haselmann, Dieter P. Gruber, and Paul Tabatabai. Anomaly detection using deep learning based image completion. In *17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, Orlando, FL, USA, December 17-20, 2018*, pages 1237–1242. IEEE, 2018. 1
- [26] Heiko Hoffmann. Kernel PCA for novelty detection. *Pattern Recognit.*, 40(3):863–874, 2007. 1
- [27] János Höner, Shinichi Nakajima, Alexander Bauer, Klaus-Robert Müller, and Nico Görnitz. Minimizing trust leaks for robust sybil detection. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1520–1528. PMLR, 2017. 1
- [28] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933. 1
- [29] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4):107:1–107:14, 2017. 2
- [30] Antanas Kascenas, Nicolas Pugeault, and Alison Q. O’Neil. Denoising autoencoders for unsupervised anomaly detection in brain mri. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, pages 653–664. PMLR, 2022. 5
- [31] Tung Kieu, Bin Yang, Chenjuan Guo, and Christian S. Jensen. Outlier detection for time series with recurrent autoencoder ensembles. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 2725–2732. ijcai.org, 2019. 1
- [32] Ki Hyun Kim, Sangwoo Shim, Yongsu Lim, Jongseob Jeon, Jeongwoo Choi, Byungchan Kim, and Andre S. Yoon. Rapp: Novelty detection with reconstruction along projection pathway. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 1
- [33] Minkyung Kim, Junsik Kim, Jongmin Yu, and Jun Kyun Choi. Active anomaly detection based on deep one-class classification. *Pattern Recognit. Lett.*, 167:18–24, 2023. 1
- [34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 7
- [35] Edwin M. Knorr, Raymond T. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. *VLDB J.*, 8(3-4):237–253, 2000. 1
- [36] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with LRP. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–7. IEEE, 2020. 1
- [37] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9664–9674. Computer Vision Foundation / IEEE, 2021. 1, 2, 5, 7
- [38] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, pages 89–105. Springer, 2018. 2
- [39] WenQian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyang Wu, Bir Bhanu, Richard J. Radke, and Octavia I. Camps. Towards visually explaining variational autoencoders. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8639–8648. IEEE, 2020. 1, 7
- [40] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpNet: A simple network for image anomaly detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 20402–20411. IEEE, 2023. 2, 7
- [41] Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Anomaly detection in nanofibrous materials by cnn-based self-similarity. *Sensors*, 18(1):209, 2018. 1
- [42] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3):106–1076, 1962. 1
- [43] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2536–2544. IEEE Computer Society, 2016. 2
- [44] Jonathan Pirnay and Keng Chai. Inpainting transformer for anomaly detection. In *Image Analysis and Processing - ICIAP 2022 - 21st International Conference, Lecce, Italy, May 23-27, 2022, Proceedings, Part II*, pages 394–406. Springer, 2022. 1, 2, 7
- [45] Emanuele Principi, Fabio Vesperini, Stefano Squartini, and Francesco Piazza. Acoustic novelty detection with adversarial autoencoders. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pages 3324–3330. IEEE, 2017. 1
- [46] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, pages 427–438. ACM, 2000. 1
- [47] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter V. Gehler. Towards total recall in industrial anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14298–14308. IEEE, 2022. 1, 8
- [48] Lukas Ruff, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert A. Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on*

- Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 4390–4399. PMLR, 2018. 1
- [49] Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus Robert Müller. A unifying review of deep and shallow anomaly detection. *Proc. IEEE*, 109(5):756–795, 2021. 1
- [50] 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 - 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*, pages 146–157. Springer, 2017. 1, 7
- [51] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Anal.*, 54:30–44, 2019. 1
- [52] Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, 1998. 1
- [53] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, 2001. 1
- [54] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1
- [55] Jeremy Tan, Benjamin Hou, Thomas Day, John M. Simpson, Daniel Rueckert, and Bernhard Kainz. Detecting outliers with poisson image interpolation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part V*, pages 581–591. Springer, 2021. 1
- [56] David M. J. Tax and Robert P. W. Duin. Support vector data description. *Mach. Learn.*, 54(1):45–66, 2004. 1
- [57] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan T.M. Duong, Chanh D. Tr. Nguyen, and Steven Q. H. Truong. Revisiting reverse distillation for anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24511–24520, 2023. 5, 8
- [58] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. Attention guided anomaly localization in images. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, pages 485–503. Springer, 2020. 1
- [59] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pages 1096–1103. ACM, 2008. 2, 5
- [60] Qian Wan, Liang Gao, Xinyu Li, and Long Wen. Industrial image anomaly localization based on gaussian clustering of pretrained feature. *IEEE Trans. Ind. Electron.*, 69(6):6182–6192, 2022. 1
- [61] Lu Wang, Dongkai Zhang, Jiahao Guo, and Yueying Han. Image anomaly detection using normal data only by latent space resampling. *Applied Sciences*, 10(23), 2020. 1, 2, 7
- [62] Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M. Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 20370–20382. Computer Vision Foundation / IEEE, 2025. 8
- [63] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Trans. Image Process.*, 23(2):684–695, 2014. 2
- [64] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5505–5514. IEEE Computer Society, 2018. 2
- [65] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4470–4479. IEEE, 2019. 2
- [66] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognit.*, 112:107706, 2021. 1, 2, 7
- [67] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Dræm - A discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8310–8319. IEEE, 2021. 1, 2, 5, 7
- [68] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16699–16708, 2024. 8
- [69] Chong Zhou and Randy C. Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 665–674. ACM, 2017. 1
- [70] Yixuan Zhou, Xing Xu, Jingkuan Song, Fumin Shen, and Heng Tao Shen. Msflow: Multi-scale flow-based framework for unsupervised anomaly detection. *CoRR*, abs/2308.15300, 2023. 8
- [71] Jiaqi Zhu, Shaofeng Cai, Fang Deng, Beng Chin Ooi, and Junran Wu. Do llms understand visual anomalies? uncovering llm’s capabilities in zero-shot anomaly detection. In *Proceedings of the 32nd ACM International Conference on*

Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024, pages 48–57. ACM, 2024. [8](#)

- [72] David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus H. Maier-Hein. Unsupervised anomaly localization using variational auto-encoders. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part IV*, pages 289–297. Springer, 2019. [1](#)
- [73] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Dae-ki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [1](#)

Noise & Pattern: Identity-Anchored Tikhonov Regularization for Robust Structural Anomaly Detection

Supplementary Material

7. Proof of Theorem 1

We first prove the following proposition.

Proposition 1. Let $d \in \mathbb{N}$, $\hat{\mathbf{x}}, \mathbf{x} \in \mathbb{R}^d$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$. Assume $f_\theta \in C^4$ in a neighborhood of $\hat{\mathbf{x}}$ and define

$$\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}, \sigma; \theta) := \mathbb{E}_\epsilon \left[\|f_\theta(\hat{\mathbf{x}} + \boldsymbol{\epsilon}) - (\mathbf{x} + \boldsymbol{\epsilon})\|_2^2 \right]. \quad (9)$$

Then, as $\sigma \rightarrow 0$, the loss $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}, \sigma; \theta)$ admits the asymptotic expansion

$$\|\mathbf{r}\|_2^2 + \sigma^2 (\|J_\theta(\hat{\mathbf{x}}) - I_d\|_F^2 + \mathbf{r}^\top \Delta f_\theta(\hat{\mathbf{x}})) + \mathcal{O}(\sigma^4), \quad (10)$$

where $\mathbf{r} = f_\theta(\hat{\mathbf{x}}) - \mathbf{x}$ is the residual, $J_\theta(\hat{\mathbf{x}}) = D_{\hat{\mathbf{x}}} f_\theta$ is the Jacobian, Δf_θ is the component-wise Laplacian of f_θ , and I_d denotes the identity matrix.

Let $\mathbf{r} := f_\theta(\hat{\mathbf{x}}) - \mathbf{x}$, $J := D_{\hat{\mathbf{x}}} f_\theta$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$.

Step 1: Second-order Taylor expansion. Since f_θ is C^3 in a neighborhood of $\hat{\mathbf{x}}$, Taylor's theorem (Fréchet form) gives

$$f_\theta(\hat{\mathbf{x}} + \boldsymbol{\epsilon}) = f_\theta(\hat{\mathbf{x}}) + J\boldsymbol{\epsilon} + \frac{1}{2} D_{\hat{\mathbf{x}}}^2 f_\theta(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) + R(\hat{\mathbf{x}}, \boldsymbol{\epsilon}),$$

where, for some $t \in (0, 1)$,

$$R(\hat{\mathbf{x}}, \boldsymbol{\epsilon}) = \frac{1}{6} D_{\hat{\mathbf{x}}+t\boldsymbol{\epsilon}}^3 f_\theta(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}, \boldsymbol{\epsilon}). \quad (11)$$

Since $D^3 f_\theta$ is continuous in a neighborhood of $\hat{\mathbf{x}}$, it is locally bounded; hence there exist $\rho > 0$ and $C > 0$ such that $\|D_{\mathbf{y}}^3 f_\theta\| \leq C$ for all $\mathbf{y}$ with $\|\mathbf{y} - \hat{\mathbf{x}}\|_2 \leq \rho$. For any $\boldsymbol{\epsilon}$ with $\|\boldsymbol{\epsilon}\|_2 \leq \rho$ we get

$$\|R(\hat{\mathbf{x}}, \boldsymbol{\epsilon})\|_2 \stackrel{(11)}{=} \frac{1}{6} \|D_{\hat{\mathbf{x}}+t\boldsymbol{\epsilon}}^3 f_\theta(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}, \boldsymbol{\epsilon})\|_2 \leq \frac{C}{6} \|\boldsymbol{\epsilon}\|_2^3, \quad (12)$$

where in the last step we used the operator norm of the trilinear map

$$\|D_{\mathbf{y}}^3 f_\theta(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}, \boldsymbol{\epsilon})\|_2 \leq \|D_{\mathbf{y}}^3 f_\theta\| \|\boldsymbol{\epsilon}\|_2^3 \leq C \|\boldsymbol{\epsilon}\|_2^3.$$

Set $M := J - I_d$ and $A := D_{\hat{\mathbf{x}}}^2 f_\theta$. Then

$$g(\boldsymbol{\epsilon}) := f_\theta(\hat{\mathbf{x}} + \boldsymbol{\epsilon}) - (\mathbf{x} + \boldsymbol{\epsilon}) = \mathbf{r} + M\boldsymbol{\epsilon} + \frac{1}{2} A(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) + R(\hat{\mathbf{x}}, \boldsymbol{\epsilon}).$$

Step 2: Expand $\|g(\boldsymbol{\epsilon})\|^2$ and take expectations. Expanding,

$$\begin{aligned} \|g(\boldsymbol{\epsilon})\|_2^2 &= \|\mathbf{r}\|_2^2 + \|M\boldsymbol{\epsilon}\|_2^2 + \frac{1}{4} \|A(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})\|_2^2 + \|R(\hat{\mathbf{x}}, \boldsymbol{\epsilon})\|_2^2 \\ &\quad + 2\mathbf{r}^\top (M\boldsymbol{\epsilon}) + \mathbf{r}^\top A(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) + 2\mathbf{r}^\top R(\hat{\mathbf{x}}, \boldsymbol{\epsilon}) \\ &\quad + \boldsymbol{\epsilon}^\top M^\top A(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) + 2\boldsymbol{\epsilon}^\top M^\top R(\hat{\mathbf{x}}, \boldsymbol{\epsilon}) \\ &\quad + A(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})^\top R(\hat{\mathbf{x}}, \boldsymbol{\epsilon}). \end{aligned}$$

Taking expectations and using $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \sigma^2 I_d$, we obtain

$$\begin{aligned} \mathbb{E}[\mathbf{r}^\top (M\boldsymbol{\epsilon})] &= 0, \\ \mathbb{E}\|M\boldsymbol{\epsilon}\|_2^2 &= \mathbb{E}[(M^\top M\boldsymbol{\epsilon})^\top \boldsymbol{\epsilon}] = \mathbb{E}[\text{tr}((M^\top M\boldsymbol{\epsilon})^\top \boldsymbol{\epsilon})] \\ &= \mathbb{E}[\text{tr}(M^\top M\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top)] = \text{tr}(M^\top M\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top]) \\ &= \sigma^2 \text{tr}(M^\top M) = \sigma^2 \|M\|_F^2 \\ \mathbb{E}[\mathbf{r}^\top A(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})] &= \mathbf{r}^\top \mathbb{E}[A(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})] = \mathbf{r}^\top \mathbb{E}[\text{tr}((H_k \boldsymbol{\epsilon})^\top \boldsymbol{\epsilon})_{k=1}^d] \\ &= \mathbf{r}^\top \mathbb{E}[\text{tr}((H_k \boldsymbol{\epsilon})^\top \boldsymbol{\epsilon})_{k=1}^d] \\ &= \mathbf{r}^\top \mathbb{E}[\text{tr}(H_k \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top)_{k=1}^d] \\ &= \mathbf{r}^\top (\text{tr}(H_k \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top]))_{k=1}^d \\ &= \sigma^2 \mathbf{r}^\top (\text{tr}(H_k))_{k=1}^d = \sigma^2 \mathbf{r}^\top \Delta f_\theta(\hat{\mathbf{x}}), \end{aligned}$$

where H_k is the Hessian of the k -th component of f_θ .

Let $\phi(\boldsymbol{\epsilon}) := \boldsymbol{\epsilon}^\top M^\top A(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})$. Then

$$\mathbb{E}[\boldsymbol{\epsilon}^\top M^\top A(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})] = \mathbb{E}[\phi(\boldsymbol{\epsilon})] = \mathbb{E}[\phi(-\boldsymbol{\epsilon})] = -\mathbb{E}[\phi(\boldsymbol{\epsilon})]$$

This implies:

$$\mathbb{E}[\boldsymbol{\epsilon}^\top M^\top A(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})] = 0. \quad (13)$$

Now recall

$$\begin{aligned} R(\hat{\mathbf{x}}, \boldsymbol{\epsilon}) &= \frac{1}{6} D_{\hat{\mathbf{x}}+\boldsymbol{\epsilon}}^3 f_\theta(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}, \boldsymbol{\epsilon}), \\ R(\hat{\mathbf{x}}, -\boldsymbol{\epsilon}) &= -\frac{1}{6} D_{\hat{\mathbf{x}}-\boldsymbol{\epsilon}}^3 f_\theta(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}, \boldsymbol{\epsilon}), \end{aligned} \quad (14)$$

for some $\theta, \theta' \in (0, 1)$. Hence

$$R(\hat{\mathbf{x}}, \boldsymbol{\epsilon}) + R(\hat{\mathbf{x}}, -\boldsymbol{\epsilon}) = \frac{1}{6} (D_{\hat{\mathbf{x}}+\boldsymbol{\epsilon}}^3 f_\theta - D_{\hat{\mathbf{x}}-\boldsymbol{\epsilon}}^3 f_\theta)(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}, \boldsymbol{\epsilon}).$$

Since $f_\theta \in C^4$ near $\hat{\mathbf{x}}$, $D^4 f_\theta$ is continuous and therefore bounded on a small closed ball B around $\hat{\mathbf{x}}$; set

$$L := \sup_{\mathbf{y} \in B} \|D_{\mathbf{y}}^4 f_\theta\| < \infty.$$

By the mean-value inequality for the map $\mathbf{y} \mapsto D_{\mathbf{y}}^3 f_\theta$,

$$\begin{aligned} \|D_{\hat{\mathbf{x}}+\boldsymbol{\epsilon}}^3 f_\theta - D_{\hat{\mathbf{x}}-\boldsymbol{\epsilon}}^3 f_\theta\| &\leq L \|\hat{\mathbf{x}} + \boldsymbol{\epsilon} - (\hat{\mathbf{x}} - \boldsymbol{\epsilon})\|_2 \\ &= L (\theta + \theta') \|\boldsymbol{\epsilon}\|_2. \end{aligned}$$

Using the operator norm of a trilinear map,

$$\|T(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}, \boldsymbol{\epsilon})\|_2 \leq \|T\| \|\boldsymbol{\epsilon}\|_2^3,$$

for $T = D_{\hat{\mathbf{x}}+\boldsymbol{\epsilon}}^3 f_\theta - D_{\hat{\mathbf{x}}-\boldsymbol{\epsilon}}^3 f_\theta$ we obtain

$$\begin{aligned} \|R(\hat{\mathbf{x}}, \boldsymbol{\epsilon}) + R(\hat{\mathbf{x}}, -\boldsymbol{\epsilon})\|_2 &\leq \frac{1}{6} L (\theta + \theta') \|\boldsymbol{\epsilon}\|_2^4 \\ &\leq \frac{L}{3} \|\boldsymbol{\epsilon}\|_2^4. \end{aligned} \quad (15)$$

By symmetrization,

$$\mathbb{E}[\mathbf{r}^\top R(\hat{\mathbf{x}}, \epsilon)] = \frac{1}{2} \mathbb{E}[\mathbf{r}^\top (R(\hat{\mathbf{x}}, \epsilon) + R(\hat{\mathbf{x}}, -\epsilon))].$$

Using the pointwise Cauchy–Schwarz inequality $\mathbf{r}^\top v \leq \|\mathbf{r}\|_2 \|v\|_2$,

$$\mathbb{E}[\mathbf{r}^\top R(\hat{\mathbf{x}}, \epsilon)] \leq \frac{\|\mathbf{r}\|_2}{2} \mathbb{E}[\|R(\hat{\mathbf{x}}, \epsilon) + R(\hat{\mathbf{x}}, -\epsilon)\|_2].$$

Combining with (15) and Jensen inequality we get

$$\mathbb{E}[\mathbf{r}^\top R(\hat{\mathbf{x}}, \epsilon)] \leq \frac{\|\mathbf{r}\|_2}{2} \cdot \frac{L}{3} \mathbb{E}[\|\epsilon\|_2^4] = \frac{\|\mathbf{r}\|_2 L}{6} (d^2 + 2d) \sigma^4$$

since $\mathbb{E}[\|\epsilon\|_2^4] = (d^2 + 2d)\sigma^4$ for $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$. Therefore,

$$\mathbb{E}[\mathbf{r}^\top R(\hat{\mathbf{x}}, \epsilon)] \in \mathcal{O}(\sigma^4). \quad (16)$$

Remaining terms. Let $A(\epsilon, \epsilon) = (\epsilon^\top H_1 \epsilon, \dots, \epsilon^\top H_d \epsilon)$ with H_k being the Hessian of the k -th component of f_θ . Since $|\epsilon^\top H_k \epsilon| \leq \|H_k\|_F \|\epsilon\|_2^2$,

$$\frac{1}{4} \mathbb{E}[\|A(\epsilon, \epsilon)\|_2^2] \in \mathcal{O}(\sigma^4), \quad \mathbb{E}[\|R(\hat{\mathbf{x}}, \epsilon)\|_2^2] \in \mathcal{O}(\sigma^6),$$

and by Cauchy–Schwarz,

$$\mathbb{E}[\epsilon^\top M^\top R(\hat{\mathbf{x}}, \epsilon)] \in \mathcal{O}(\sigma^4), \quad \mathbb{E}[A(\epsilon, \epsilon)^\top R(\hat{\mathbf{x}}, \epsilon)] \in \mathcal{O}(\sigma^5).$$

Step 3: Collecting terms. Keeping contributions up to order σ^4 , we can write $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}, \sigma; \theta)$ as

$$\|\mathbf{r}\|_2^2 + \sigma^2 \|J - I_d\|_F^2 + \sigma^2 \mathbf{r}^\top \Delta f_\theta(\hat{\mathbf{x}}) + \mathcal{O}(\sigma^4),$$

which is the claimed expansion.

Proof of Theorem 1 . We now prove Theorem 1 by using the result in Proposition 1. By independence of $(\hat{\mathbf{x}}, \mathbf{x})$ and ϵ we can write

$$\mathcal{L}_{\text{FAE}}(\sigma; \theta) = \mathbb{E}_{(\hat{\mathbf{x}}, \mathbf{x})}[\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}, \sigma; \theta)],$$

where

$$\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}, \sigma; \theta) := \mathbb{E}_\epsilon [\|f_\theta(\hat{\mathbf{x}} + \epsilon) - (\mathbf{x} + \epsilon)\|_2^2].$$

For fixed $(\hat{\mathbf{x}}, \mathbf{x})$, Theorem 1 gives the following expansion for $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}, \sigma; \theta)$

$$\|\mathbf{r}\|_2^2 + \sigma^2 (\|J_\theta(\hat{\mathbf{x}}) - I_d\|_F^2 + \mathbf{r}^\top \Delta f_\theta(\hat{\mathbf{x}})) + R_\sigma(\hat{\mathbf{x}}),$$

where $\mathbf{r} = f_\theta(\hat{\mathbf{x}}) - \mathbf{x}$ and the remainder satisfies

$$|R_\sigma(\hat{\mathbf{x}})| \leq C(\hat{\mathbf{x}}) \sigma^4,$$

for some measurable $C(\hat{\mathbf{x}})$. Since $f_\theta \in C^3$ and $D^3 f_\theta$ is locally Lipschitz, such a bound follows from the Taylor expansion with Gaussian averaging. Taking expectations with respect to $(\hat{\mathbf{x}}, \mathbf{x})$ yields

$$\mathbb{E}[\|\mathbf{r}\|_2^2 + \sigma^2 (\|J_\theta(\hat{\mathbf{x}}) - I_d\|_F^2 + \mathbf{r}^\top \Delta f_\theta(\hat{\mathbf{x}}))] + \mathbb{E}[R_\sigma(\hat{\mathbf{x}})].$$

By the integrability assumption

$$\mathbb{E}[\|\mathbf{r}\|_2^2], \mathbb{E}[\|J_\theta(\hat{\mathbf{x}})\|_F^2], \mathbb{E}[\|D_{\hat{\mathbf{x}}}^2 f_\theta\|_F^2] < \infty,$$

we also have $\mathbb{E}[C(\hat{\mathbf{x}})] < \infty$, so $\mathbb{E}[R_\sigma(\hat{\mathbf{x}})] = \mathcal{O}(\sigma^4)$. Thus, the loss $\mathcal{L}_{\text{FAE}}(\sigma; \theta)$ is equal to

$$\mathbb{E}[\|\mathbf{r}\|_2^2 + \sigma^2 (\|J_\theta(\hat{\mathbf{x}}) - I_d\|_F^2 + \mathbf{r}^\top \Delta f_\theta(\hat{\mathbf{x}}))] + \mathcal{O}(\sigma^4),$$

which is the claimed expansion. $\square$

8. Varying Local Thickness of Occluding Shapes Improves Segmentation Accuracy

As discussed in Section 5, incorporating thin, curve-like masks alongside broad occlusions during training significantly improves the correction of the corresponding thin, curve-shaped anomalies in the testing phase. Figure 7 illustrates two examples. Notably, models trained without curve-shaped occlusions consistently fail to completely eliminate such anomalies in the affected regions.

9. Ablation Study for Gaussian Noise Regularization

We conduct an ablation study by evaluating the performance of models trained with and without Gaussian noise regularization. The corresponding results are reported in Table 3. Two key observations emerge. First, regularization with Gaussian noise consistently improves segmentation accuracy. Second, models trained without this regularization exhibit higher performance variance. While the relative improvement is moderate, ranging between 0.01 and 1.45, this is mainly because the corruption model alone is already responsible for near-saturated results. Nevertheless, noise regularization remains important: it improves segmentation accuracy, stabilizes training by reducing the variance of the final results, and its effect is clearly visible in the enhanced reconstruction quality of the model.

In anomaly detection, perfect reconstructions are not always required, since even small pixel-level differences between input and output can suffice to localize anomalies. However, the reconstructions obtained with noise regularization are noticeably more faithful and less artifact-prone, underscoring its relevance even when the quantitative gains appear modest. Figure 8 illustrates this qualitative improvement, while Figure 9 and 10 present further examples and highlights the quality of the resulting heatmaps.

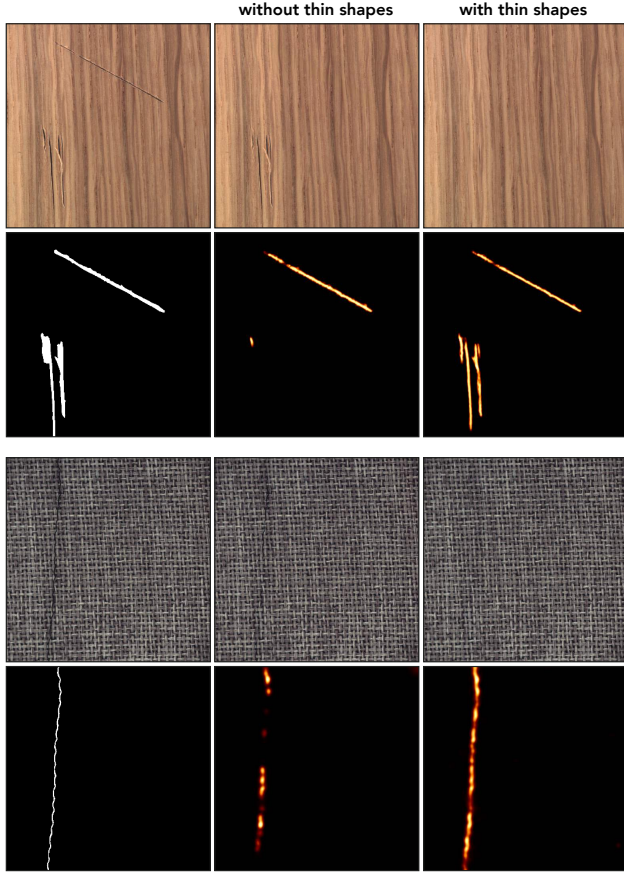


Figure 7. Illustration of the qualitative difference in filtering anomalous patterns during reconstruction for models trained with and without thin shapes. The first row (for each example) shows a defective input image and the corresponding reconstructions from models trained without and with thin shapes, respectively. The second row (for each example) shows the human annotation mask alongside the anomaly heatmaps produced by the two models.

Table 3. Ablation study with and without Gaussian noise regularization measured with **P-AUROC** on MVTec AD.

Category	With noise	W/o noise	Improvement
carpet	99.73 ± 0.02	99.20 ± 0.15	+0.53
grid	99.79 ± 0.01	99.75 ± 0.06	+0.04
leather	99.75 ± 0.00	99.62 ± 0.08	+0.13
tile	98.75 ± 0.38	98.35 ± 0.65	+0.40
wood	98.55 ± 0.03	98.43 ± 0.16	+0.12
bottle	99.24 ± 0.05	98.60 ± 0.24	+0.64
cable	98.62 ± 0.03	98.07 ± 0.33	+0.55
capsule	99.18 ± 0.08	98.09 ± 0.11	+0.09
hazelnut	99.48 ± 0.07	99.03 ± 0.15	+0.45
metal nut	99.00 ± 0.56	97.55 ± 1.13	+1.45
pill	99.52 ± 0.01	99.16 ± 0.09	+0.36
screw	99.38 ± 0.28	99.33 ± 0.33	+0.05
toothbrush	99.45 ± 0.02	99.44 ± 0.03	+0.01
transistor	98.12 ± 0.78	97.88 ± 1.02	+0.24
zipper	99.70 ± 0.00	99.69 ± 0.00	+0.01

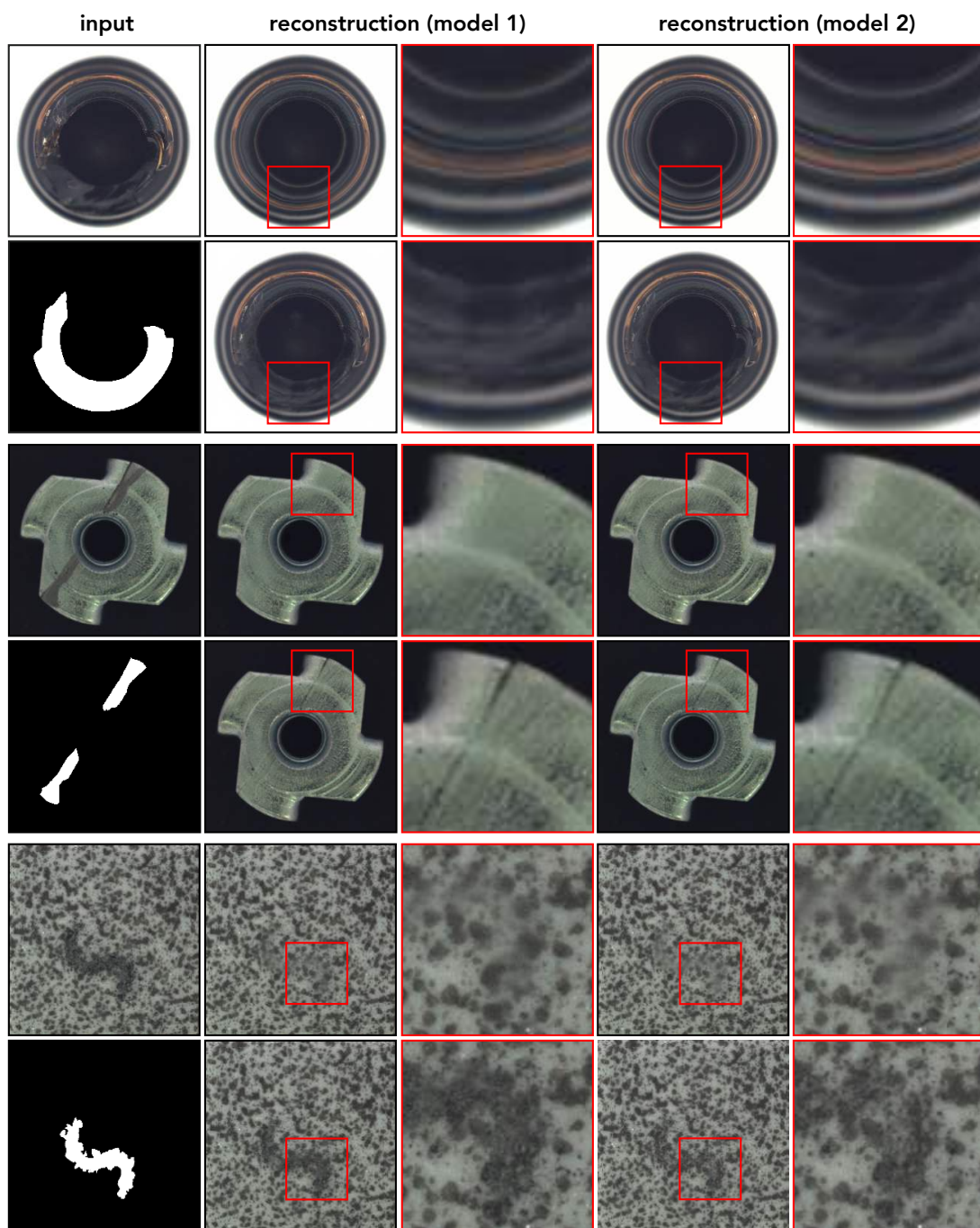
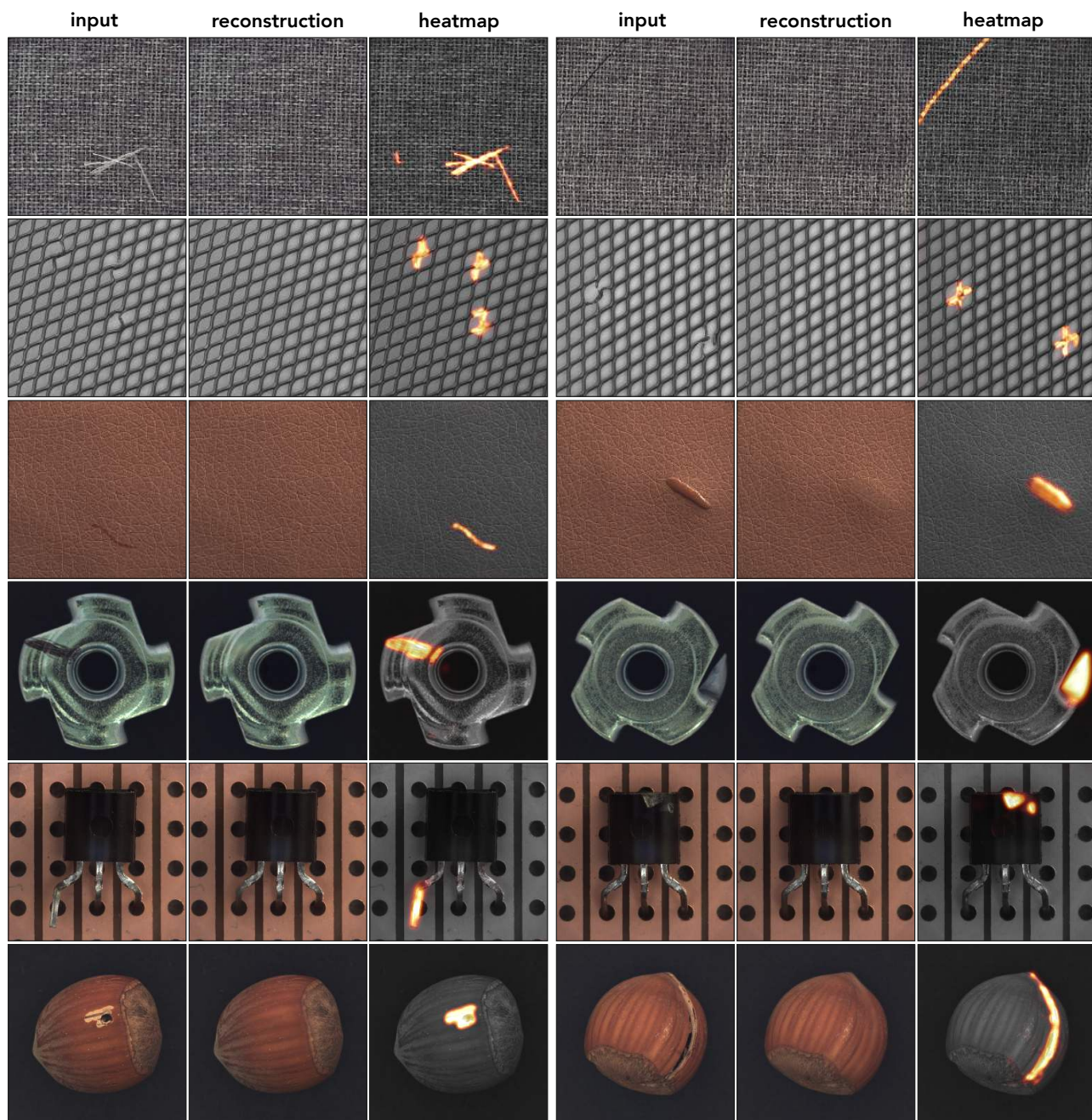


Figure 8. Qualitative comparison of reconstructions from models trained with and without Gaussian noise regularization. Odd rows show a defected input image followed by reconstructions from two regularized models, while even rows show the corresponding human-annotated anomaly mask followed by reconstructions from two unregularized models. For each reconstruction, a zoom-in of the region marked by the red rectangle is provided to enable detailed comparison. Gaussian noise regularization consistently enhances the correction of anomalous regions.



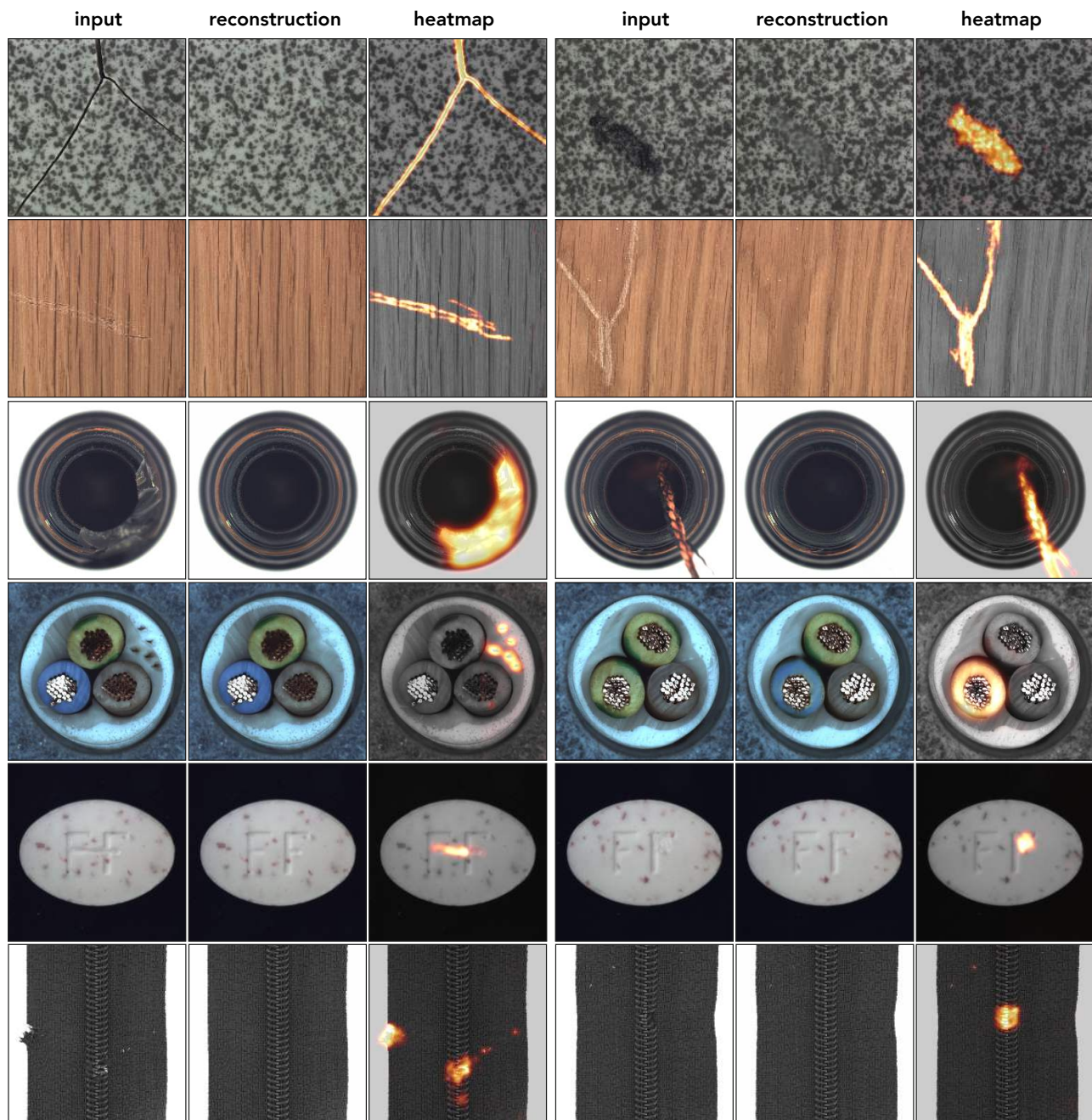


Figure 10. Qualitative results of reconstructions and corresponding anomaly heatmaps produced by our model. Each row shows two examples from one category. For each example we show the defective input image, its reconstruction, and the heatmap overlay.

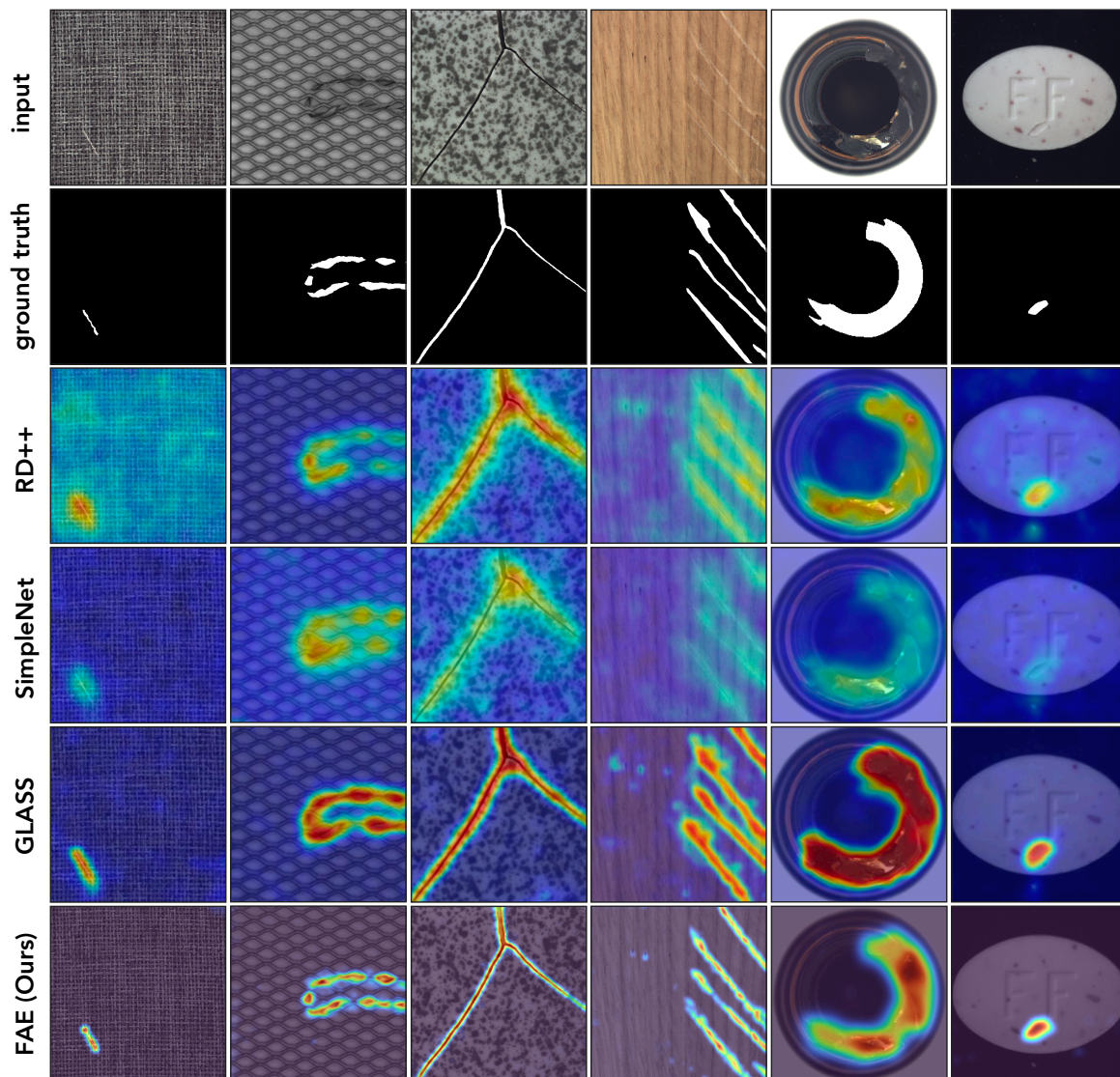


Figure 11. Qualitative comparison of anomaly heatmaps on representative samples from the MVTec AD dataset. Each column displays the input image, the corresponding human annotation mask, and heatmaps generated by multiple existing methods, with the results of our method shown at the bottom for comparison.