VALUE-ORDER DECOMPOSITION FOR GENERALIST ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Industrial anomaly detection suffers from limited data, making cross-domain generalization particularly challenging. Generalist Anomaly Detection (GAD) aims to train a unified model on a source domain that can effectively detect anomalies in unseen target domains. In the initial semantic feature space, strong entanglement between anomalies and object categories or defect types hinders effective generalization across domains. Recent works address this issue by projecting features into a residual space; however, such methods primarily increase cross-domain overlap for normal features, while anomalous features remain specific to object categories, defect types and data domains, leading to poor alignment and generalization. To address this limitation, we propose Value-order Decomposition (VOD), a simple yet effective technique that bridges **three types of generalization gaps** across object categories, defect types (including real and synthetic defects), and data domains. VOD disentangles and suppresses object-category-, defect-type-, and domain-specific information, promoting alignment within normal and abnormal samples while preserving their separability, thereby enabling robust generalization across the three gaps. Leveraging the strong alignment between real and synthetic defects within the same object, we perform anomaly detection using only normal and synthetic-abnormal reference, and effectively generalize to unseen real defect types. Experiments on diverse industrial and medical benchmarks demonstrate that our method, using a simple cut-and-paste anomaly simulation strategy, achieves strong generalization across the three gaps.

1 INTRODUCTION

Visual anomaly detection (VAD), which aims to classify images as normal or abnormal and localize anomalies, plays a crucial role in industrial quality inspection and medical diagnosis (Barusco et al., 2025; Bao et al., 2024). In real-world applications, however, data scarcity and privacy constraints make it impractical to train a dedicated model for each task. Instead, models are typically pretrained on source domains (Yao et al., 2024a; Wang et al., 2025b) and expected to generalize to unseen target domains. However, VAD is inherently a long-tail problem. Objects and their defects vary significantly in category, color, texture, and size across domains. Consequently, both normal and abnormal features exhibit large variations in the initial feature space (Figure 1 (a)), leading to insufficiently representative samples and poor generalization of source-trained anomaly detection (AD) models (Jeong et al., 2023b; Xu et al., 2025; Gu et al., 2025). To alleviate this issue, a line of unified AD methods has been proposed (Yao et al., 2024b; 2023b;a; You et al., 2022), which retrain or fine-tune models on new classes or domains (Yao et al., 2024a). However, these approaches are computationally expensive and require substantial data, limiting their scalability. As a result, achieving generalization across object categories and domains remains a fundamental challenge in the AD community.

Recently, Generalist Anomaly Detection (GAD) has emerged to address this limitation. It aims to train a unified model on source domains and generalize to unseen target domains without retraining or fine-tuning (Yao et al., 2024a; Wang et al., 2025a; Zhu & Pang, 2024). Several popular GAD methods focus on learning residual feature distributions rather than initial feature distributions, which effectively reduces variations of normal features across object categories. Nevertheless, abnormal residual features still exhibit significant variations across object categories, defect types, and domains. We refer to these variations as the "**three generalization gaps**." This entanglement

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

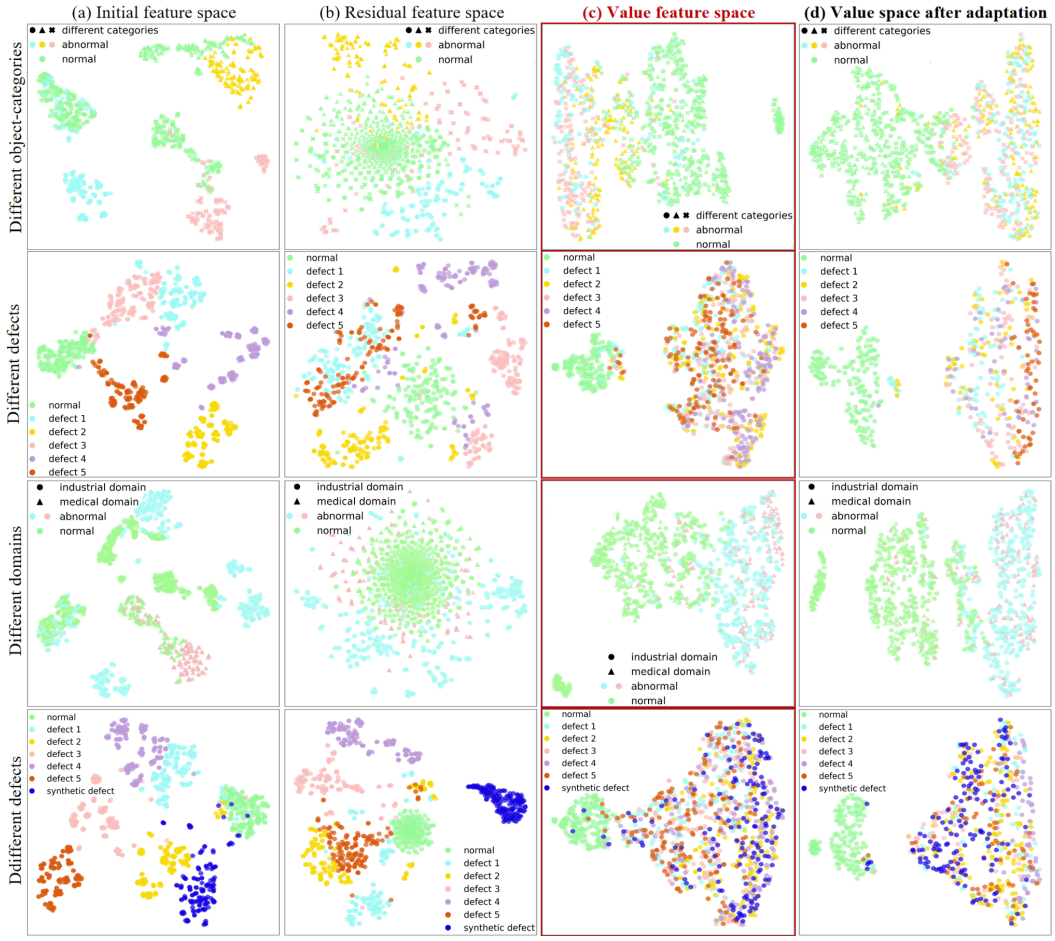


Figure 1: t-SNE visualization of normal and anomalous patches under different representations across the three generalization gaps. (a) The initial feature space exhibits three distinct gaps (object categories, defect types, and data domains) rather than the normal/abnormal distinction, which hinders generalization. (b) Residual-space methods mainly improve the alignment among normal patches. (c) In contrast, the value component of VOD enforces intra-class alignment for both normal and anomalous patches across all gaps. (d) After adaptation, the value component further enhances the separability between normal and anomalous patches. (See Appendix E.4 for more details.)

hinders generalization to unseen scenarios (see Figure 1(b)), which often necessitates elaborate designs (Yao et al., 2024a; Gao et al., 2026). To further mitigate this issue, recent works explore normal–abnormal guidance (Wang et al., 2025b), which leverages anomalous samples of the same object and same defect-type, in addition to normal samples, to better distinguish anomalies across domains. However, such methods rely on predefined defect types, making them infeasible in scenarios where anomaly samples are unavailable, and prone to overfitting to seen defect types, thereby limiting their generalization to novel ones.

To address the three generalization gaps and reduce variations across object categories, defect types, and domains in abnormal features, we focus on the GAD paradigm and propose a novel Value-Order Decomposition (VOD) technique inspired by feature distribution matching (Zhang et al., 2022; Rolland et al., 2000), enabling strong generalization without requiring abnormal references. Specifically, VOD decomposes residual features into two components (see Figure 1 (c) and Figure 3 (a)): (i) a value component that captures general anomaly information and remains invariant across object categories, defect types, and domains; and (ii) an order component that encodes semantic and structural information varying across the three gaps.

Leveraging the value component, we construct an anomaly representation that consistently aligns normal and abnormal samples across diverse object categories, defect types (including real and synthetic ones), and domains. Based on this invariant representation, we further develop a synthetic-anomaly-based detection framework that requires only a few normal samples and generates synthetic anomalies via simple augmentation. During inference, each sample is compared against both normal and synthetic-anomaly references to enable more accurate anomaly detection. Notably, our method relies only on simple cut-and-paste augmentation, rather than complex diffusion-based

synthesis (Zhang et al., 2024), while still achieving state-of-the-art generalization performance. **The main contributions of this work are:** (i) We propose a novel Value-Order Decomposition (VOD) technique, which decomposes residual features into gap-invariant and gap-specific components, where the former bridges generalization gaps across object categories, defect types, and data domains. (ii) By bridging the gap between synthetic and real defects via VOD, we develop a simple yet effective synthetic-anomaly-based framework termed Value-Order-Decomposition-based Anomaly Detection (VODA), which avoids complex designs and generalizes well without relying on real anomaly data. (iii) Extensive experiments demonstrate that our method, despite its simplicity, outperforms existing baselines by a large margin in most cases, achieving strong generalization and even matching or surpassing methods that rely on real anomaly references.

2 Related work

2.1 Generalist Anomaly Detection

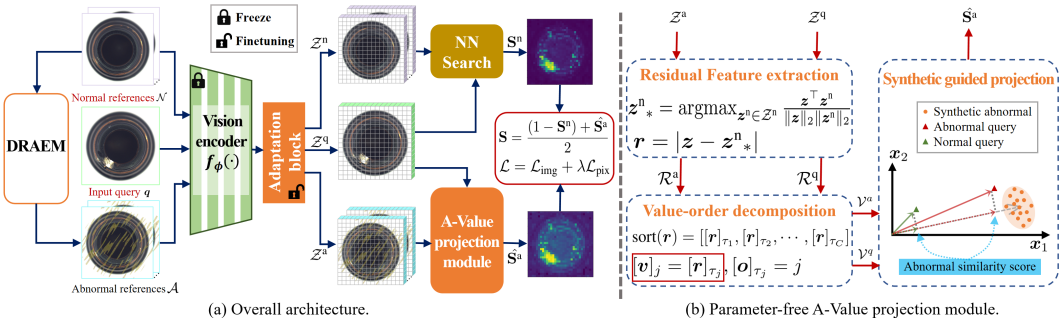
Due to the scarcity and high cost of collecting anomalous data, most prior works focus on unsupervised AD, which can be broadly categorized into three technical streams: (i) embedding distance-based methods, which identify anomalies through statistical deviations in offline features extracted from pretrained models that preserve discriminative information (*e.g.,* SPADE (Cohen & Hoshen, 2020), PaDiM (Defard et al., 2021b), PatchCore (Roth et al., 2022), HGAD (Yao et al., 2024b), PyramidFlow (Lei et al., 2023)); (ii) reconstruction-based methods, which leverage autoencoders (Hou et al., 2021) or GANs (Yan et al., 2021) to detect anomalies through high reconstruction errors (Yao et al., 2023b;a; Xiang et al., 2023); (iii) knowledge distillation-based methods, which detect anomalies by measuring discrepancies between student features and those of a fixed pretrained teacher (Cao et al., 2023; Wang et al., 2021; Tien et al., 2023; Zhao, 2023; Gao, 2024). These methods are typically trained using only normal data and identify anomalies by measuring deviations from learned normal patterns. While effective in category-specific settings, they tend to overfit to seen object categories and often perform poorly on unseen categories and domains. Moreover, their reliance on retraining limits practical applicability when data collection is costly or restricted.

To address these limitations, GAD has recently emerged to study cross-domain anomaly detection (Zhu & Pang, 2024). In this work, we follow the GAD setting and focus on generalization to unseen domains without retraining or fine-tuning. Existing GAD methods typically achieve generalization by modeling residuals with respect to normal references. For example, InCTRL (Zhu & Pang, 2024) establishes a baseline for cross-dataset anomaly classification by capturing contextual residuals between query images and normal references, but it focuses on classification and lacks precise localization. ResAD (Yao et al., 2024a) and AdaptCLIP (Gao et al., 2026) further leverage residual features with explicit normality constraints or comparative learning between query and normal image prompt to enable both detection and localization. However, residual-based methods primarily align normal distributions, while abnormal features still exhibit significant variation across domains. To improve generalization to unseen defect types, some works introduce anomalous references. For instance, NAGL (Wang et al., 2025b) utilizes real anomalous samples to guide detection, but such approaches rely on the availability of anomaly data and are impractical in many real-world scenarios.

To bypass the need for annotated anomalous data, a line of self-supervised methods generates synthetic anomalies with ground-truth annotations and formulates proxy tasks on them (Liu et al., 2023; Chen et al., 2024). For example, CutPaste (Li et al., 2021) creates anomalies by transplanting image patches, while NSA (Pérez et al., 2023) employs Poisson image editing for more seamless synthesis. DRAEM (Zavrtanik et al., 2021) leverages external texture datasets to generate diverse anomalies. RealNet (Zhang et al., 2024) and AnoGen (Gui et al., 2024) utilize diffusion models to synthesize more realistic samples. However, these methods heavily rely on the quality and diversity of synthetic data, are typically limited to in-domain settings, and require regeneration and retraining when applied to new tasks.

2.2 Feature distribution matching

In image style transfer, feature distributions are closely related to image styles and can be used to align target and source domains while preserving content information (Huang & Belongie, 2017). This idea has also been widely used in domain generalization. Several methods study feature



(a) Overall architecture. (b) Parameter-free A-Value projection module.

Figure 2: The framework of Value-Order-Decomposition-based Anomaly Detection (VODA).

distribution alignment from this perspective (Shapira et al., 2013; Coltuc et al., 2006; Hall, 2006; Rolland et al., 2000). Among them, Sort-Matching (Rolland et al., 2000) performs Exact Histogram Matching(EHM)(Coltuc et al., 2006; Hall, 1974), leading to Exact Feature Distribution Matching (EFDM) (Zhang et al., 2022). EFDM has been successfully applied to both style transfer and domain generalization by aligning feature statistics across domains.

Our VOD is inspired by EFDM but differs in two key aspects. (i) We explicitly decompose features into gap-invariant and gap-specific components, enabling the separation of transferable abnormal patterns from variations induced by object categories, defect types, and data domains. In contrast, EFDM focuses on feature mix-up via value-order recombination as a measure-preserving transformation for data augmentation, without explicit feature disentanglement. (ii) Instead of jointly utilizing both components as in EFDM, we identify the value component as the most informative for anomaly detection and discard the order component, resulting in a simpler and more robust formulation.

3 Our Method

Focusing on the GAD setting, our goal is to train a unified anomaly detection model on a source dataset that generalizes to arbitrary unseen target domains. For a given query image, the detection process requires only a few normal samples as references. In the following, we first introduce the proposed VOD technique, highlighting the invariant nature of the value component and the variability captured by the order component. We then demonstrate the effectiveness of VOD in bridging the three generalization gaps. Based on this, we construct a universal anomaly detection framework, termed **Value-Order-Decomposition-based Anomaly Detection (VODA)**, which leverages both normal and synthetic-anomaly references. The overall architecture is illustrated in Figure 2. For a feature vector $\boldsymbol{r} \in \mathbb{R}^C$, we use $[\boldsymbol{r}]_j$ represents its j-th element.

3.1 Value-order decomposition

In this section, we introduce the proposed Value-Order Decomposition (VOD) and investigate the strong generalization capability of the value component. Given a query image $\boldsymbol{q} \in \mathbb{R}^{H\times W\times 3}$ and a corresponding normal reference set $\mathcal{N} = \{\boldsymbol{x}_k\}_{k=1}^K$ consisting of K shots from the same object category, where $\boldsymbol{x}_k \in \mathbb{R}^{H\times W\times 3}$, H and W denote the height and width of the image, respectively. For both the query and reference images, we extract features using a pre-trained backbone $f_{\boldsymbol{\phi}}(\cdot)$ parameterized by $\boldsymbol{\phi}$, following common practice (Wang et al., 2025b; Yao et al., 2024a). We then obtain query patch features $\mathcal{Z}^{\mathrm{q}} = \{\boldsymbol{z}^{\mathrm{q}}{}_i\}_{i=1}^L$ and normal reference patch features $\mathcal{Z}^{\mathrm{n}} = \{\boldsymbol{z}^{\mathrm{n}}{}_i\}_{i=1}^{KL}$, where $\boldsymbol{z}^{\mathrm{q}}{}_i, \boldsymbol{z}^{\mathrm{n}}{}_i \in \mathbb{R}^C$, $L = H \times W$ is the number of patches per image, and C is the channel dimension.

Next, we compute residual features for the query following Yao et al. (2024a) to obtain an initial anomaly localization. Specifically, for each $\boldsymbol{z}^{\mathrm{q}} \in \mathcal{Z}^{\mathrm{q}}$, we perform nearest-neighbor search in $\mathcal{Z}^{\mathrm{n}}$ using cosine similarity to identify the closest normal reference feature[1]: $\boldsymbol{z}^{\mathrm{n}}{}_* = \mathrm{argmax}_{\boldsymbol{z}^{\mathrm{n}}\in\mathcal{Z}^{\mathrm{n}}} \frac{(\boldsymbol{z}^{\mathrm{q}})^\top \boldsymbol{z}^{\mathrm{n}}}{\|\boldsymbol{z}^{\mathrm{q}}\|_2\|\boldsymbol{z}^{\mathrm{n}}\|_2}$, with $\|\cdot\|_2$ denoting the ℓ_2 norm. The residual feature is then computed as:

$$\boldsymbol{r}^{\mathrm{q}} = |\boldsymbol{z}^{\mathrm{q}} - \boldsymbol{z}^{\mathrm{n}}{}_*| \tag{1}$$

where $|\cdot|$ denotes the element-wise absolute value. By subtracting the most similar normal feature, the residual is expected to suppress object-category-related components in the original features, while

[1]Here, we consider Vision Transformer as backbone, where cosine similarity is a more appropriate distance metric.

preserving the discrepancy between normal and anomalous patterns. However, this process primarily aligns the distribution of normal residual features toward an origin-centered region across different object categories, while abnormal residual features remain scattered around the origin with diverse magnitudes and limited separability (Figure 1 (b)). As a result, a complex one-class classification mechanism is often required to further distinguish normal and abnormal samples (Yao et al., 2024a).

We observe that abnormal residual features naturally cluster according to the three generalization gaps (Figure 1 (b)), indicating that residual representations entangle anomaly cues with gap-specific factors and are therefore poorly aligned across gaps. Our motivation is inspired by EFDM, which shows that feature representations contain complementary order and value information: order is closely associated with image content, while value captures high-order feature statistics. Motivated by this observation, we hypothesize that feature ordering mainly encodes gap-specific structural patterns, while feature magnitudes provides a more gap-invariant representation of anomaly cues. Based on this insight, we propose Value-Order Decomposition (VOD), which decomposes a residual features into (i) a gap-invariant value component that captures feature magnitudes and (ii) a gap-specific order component that records the relative ranking of feature response. Specifically, we sort the residual feature $\boldsymbol{r}^{\mathrm{q}}$ in descending order:

$$\mathrm{sort}(\boldsymbol{r}^{\mathrm{q}}) = [[\boldsymbol{r}^{\mathrm{q}}]_{\tau_1}, [\boldsymbol{r}^{\mathrm{q}}]_{\tau_2}, [\boldsymbol{r}^{\mathrm{q}}]_{\tau_3}, \cdots, [\boldsymbol{r}^{\mathrm{q}}]_{\tau_C}] \tag{2}$$

where $\{[\boldsymbol{r}^{\mathrm{q}}]_{\tau_j}\}_{j=1}^{C}$ denotes the sorted values of $\boldsymbol{r}^{\mathrm{q}}$, with $[\boldsymbol{r}^{\mathrm{q}}]_{\tau_1}$ being the largest element. Here, τ_j indicates the index of the j-th sorted element in the original feature. We then define the value component $\boldsymbol{v}^{\mathrm{q}} \in \mathbb{R}^C$, where its j-th element is computed as:

$$[\boldsymbol{v}^{\mathrm{q}}]_j = [\boldsymbol{r}^{\mathrm{q}}]_{\tau_j} \tag{3}$$

Intuitively, the value component $\boldsymbol{v}^{\mathrm{q}}$ corresponds to the sorted version of $\boldsymbol{r}^{\mathrm{q}}$ in descending order, capturing the degree of abnormality (*i.e.,* from the most abnormal dimension to the least abnormal one). The order component $\boldsymbol{o}^{\mathrm{q}} \in \mathbb{R}^C$ is then defined such that its element at index τ_j is given by:

$$[\boldsymbol{o}^{\mathrm{q}}]_{\tau_j} = j \tag{4}$$

Specifically, $[\boldsymbol{o}^{\mathrm{q}}]_{\tau_j}$ represents the rank (i.e., the value order) of the element $[\boldsymbol{r}^{\mathrm{q}}]_{\tau_j}$ in the descending order of $\boldsymbol{r}^{\mathrm{q}}$. For clarity, a toy example is shown at the top of Figure 3(c).

Property 1 (Complementary Properties of Value and Order)**.** *Given a residual feature vector, the value and order components preserve complementary information. The order component preserves the ordinal structure of feature responses, whereas the value component preserves the empirical distribution of feature magnitudes. (See Appendix A for detailed proofs)*

The key observation is that the order component depends only on the relative ordering among feature responses rather than their exact magnitudes. Replacing feature values with equally spaced ranks preserves all pairwise ordering relationships while removing magnitude differences. When visualized as a spatial feature map/vector, locations with stronger responses remain stronger than locations with weaker responses, preserving structural layouts such as edges, textures, and defect morphology. In contrast, the value component sorts all feature responses into a canonical ordered sequence, completely discarding their original ordering. Therefore, any feature vectors sharing the same set of feature values, regardless of their spatial or channel arrangement, produce the same value component. The value component thus preserves magnitude statistics while removing ordering-dependent semantic structures. Under our hypothesis that anomaly cues are primarily reflected by abnormal residual magnitudes, the canonical value component suppresses gap-specific ordering while aligning anomalies across heterogeneous object categories, defect types, and domains. Toy examples in 2/1D (Figure 3(c)) provide an intuitive illustration of this complementary property: different residuals preserve distinct structural layouts in the order component, whereas their value components become highly aligned after sorting.

Significance of the Value Component. Figure 1(c) and 3(a) empirically validates Property 1. Specifically, (i) value representations become substantially better aligned across object categories, defect types, and domains; (ii) normal and anomalous samples become approximately linearly separable in the value space; and (iii) synthetic anomalies effectively capture the mixed distribution of multiple real defect types, enabling a decision boundary that generalizes well to unseen real anomalies. Thus we adopt the value component as the primary representation for anomaly detection.

The synthetic defects are generated using a simple cut-and-paste anomaly simulation strategy adopted from DRAEM (Zavrtanik et al., 2021). This approach leverages a Perlin noise generator (Perlin, 1985)

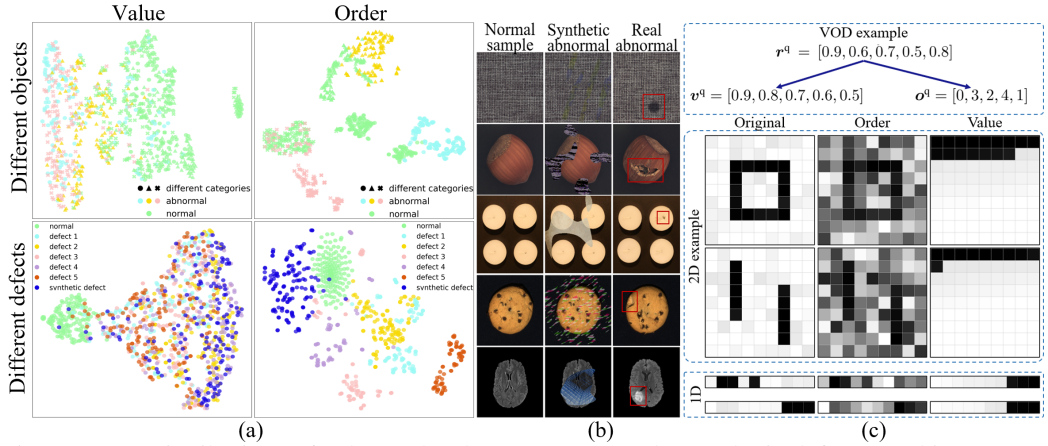


Figure 3: (a) Distributions of value and order components. (b) Synthetic defects used in our work differ from real anomalies. (c) Toy examples on the complementarity of value and order.

and an external texture dataset (Cimpoi et al., 2014) to produce diverse defects. A few generated defects are shown in Figure 3(b). Although the synthetic defects look very different from real ones—the textures are often unrealistic, and the defect regions may spread across both the object and background—our goal is not to faithfully replicate real anomaly appearances in the target domain. Instead, we aim to generate diverse out-of-distribution abnormal patterns that improve generalization.

3.2 Synthetic-anomaly-based framework for anomaly detection

Motivated by VOD's favorable property for synthetic defects, we propose to leverage it for anomaly detection, and **refer to the overall framework as Value-Order-Decomposition-based Anomaly Detection(VODA)**. The overall pipeline is shown in Figure 2 and described as follows.

Synthetic abnormal reference generation. Given K-shot normal references, we adopt the synthetic anomaly generation strategy from DRAEM (Zavrtanik et al., 2021). Specifically, for each normal reference, we generate a random mask $\mathbf{M}$ that captures diverse anomaly shapes. For each mask, only 20% of its values are kept non-zero to define the anomalous regions. We then randomly sample a texture image from DTD (Cimpoi et al., 2014) as the anomaly source. The normal reference is combined with the sampled texture via $\mathbf{M}$, following DRAEM, to produce K-shot synthetic abnormal references. We then construct a tuple $\{\mathcal{N}, \mathcal{A}, \boldsymbol{q}\}$ consisting of normal references, synthetic abnormal references, and the query image. Anomaly detection is then performed as follows.

We first extract features for each image in the tuple using a pretrained backbone (*e.g.,* DINOv2 (Oquab et al., 2024)) with an appended adaptation block (*i.e.,* a Transformer encoder layer with 4 attention heads), which provides additional flexibility to enhance the separability between normal and abnormal features. This results in the feature tuple $\{\mathcal{Z}^{\mathrm{n}}, \mathcal{Z}^{\mathrm{a}}, \mathcal{Z}^{\mathrm{q}}\}$. Based on the assumption that normal patches are mutually similar while abnormal patches exhibit diverse patterns, we compute the anomaly score for each query patch by comparing it with normal references in the original feature space and with abnormal references in the value space of VOD.

Normal similarity score. Given a query patch feature $\boldsymbol{z}^{\mathrm{q}}{}_i \in \mathcal{Z}^{\mathrm{q}}$, we perform nearest-neighbor search in $\mathcal{Z}^{\mathrm{n}}$ to identify the closest normal feature $\boldsymbol{z}^{\mathrm{n}}{}_* = \mathrm{argmax}_{\boldsymbol{z}^{\mathrm{n}} \in \mathcal{Z}^{\mathrm{n}}} \frac{(\boldsymbol{z}^{\mathrm{q}}{}_i)^\top \boldsymbol{z}^{\mathrm{n}}}{\|\boldsymbol{z}^{\mathrm{q}}{}_i\|_2 \|\boldsymbol{z}^{\mathrm{n}}\|_2}$. Let $\mathbf{S}^{\mathrm{n}} \in \mathbb{R}^L$ denote the normal similarity scores for all query patches. The score for patch $\boldsymbol{z}^{\mathrm{q}}{}_i$ is defined as the cosine similarity to its nearest normal neighbor (rescaled to interval $[0, 1]$),

$$[\mathbf{S}^{\mathrm{n}}]_i = \left(\frac{(\boldsymbol{z}^{\mathrm{q}}{}_i)^\top \boldsymbol{z}^{\mathrm{n}}{}_*}{\|\boldsymbol{z}^{\mathrm{q}}{}_i\|_2 \|\boldsymbol{z}^{\mathrm{n}}{}_*\|_2} + 1\right)/2 \tag{5}$$

A-Value projection module for abnormal similarity score. To achieve strong generalization, we compute abnormal similarity in the value space. We refer to this process as A-Value projection. Specifically, we first compute residual features for both abnormal references and the query following Equation 1, and then extract the value components via VOD (Equation 3), resulting in $\mathcal{V}^{\mathrm{a}} = \{\boldsymbol{v}^{\mathrm{a}}{}_i\}_{i=1}^{KL}$ and $\mathcal{V}^{\mathrm{q}} = \{\boldsymbol{v}^{\mathrm{q}}{}_i\}_{i=1}^{L}$. Since real and synthetic abnormal features are well aligned in the value space while remaining clearly separable from normal features, the abnormality of a query feature can be naturally measured by its projection onto abnormal directions (Figure 2(b)). The abnormal similarity score $\mathbf{S}^{\mathrm{a}} \in \mathbb{R}^L$ is defined as:

$$[\mathbf{S}^{\mathrm{a}}]_i = \max_{\boldsymbol{v}^{\mathrm{a}} \in \mathcal{V}^{\mathrm{a}}} \frac{(\boldsymbol{v}^{\mathrm{q}}{}_i)^\top \boldsymbol{v}^{\mathrm{a}}}{\|\boldsymbol{v}^{\mathrm{a}}\|_2} \tag{6}$$

Table 1: AUROC (%) for anomaly detection and localization on 6 real-world datasets under few-shot settings. "·/·" denotes image-/pixel-level results, "$\mathcal{N}$/$\mathcal{A}$" denotes using normal/abnormal references. Values are averaged over 3 runs; best and second-best are bolded and underlined.

Shots	Datasets	$\mathcal{N}$								$\mathcal{N}\&\mathcal{A}$	$\mathcal{N}$
		SPADE	PaDiM	Patchcore	RegAD	WinCLIP	InCTRL	ResAD	AdaptCLIP	NAGL	VODA
2	VisA	71.7/65.4	68.7/91.5	65.0/80.4	70.6/93.3	81.9/94.9	85.8/-	84.5/95.1	92.3/97.1	89.9/97.6	**92.4/98.4**
	MVTecAD	74.6/64.0	79.5/93.8	74.7/85.2	80.4/93.3	93.1/93.8	94.0/-	94.4/95.6	95.7/94.5	**96.8**/96.8	95.7/**97.4**
	BTAD	80.7/65.4	88.9/95.2	80.9/83.1	87.2/93.9	85.5/95.8	92.3/-	91.1/96.4	93.4/**96.7**	93.0/76.1	**95.4**/96.0
	MVTec3D	62.5/78.6	59.6/94.3	58.8/83.4	59.5/96.4	74.1/96.8	68.9/-	78.5/97.5	82.9/97.8	82.8/94.8	**84.2/98.0**
	MVTecLOCO	59.3/65.7	53.2/66.4	57.5/68.0	55.0/66.2	62.6/64.7	58.7/-	65.6/68.0	**67.4**/69.1	64.6/66.3	66.7/**75.5**
	BraTS	58.0/92.8	49.4/90.2	58.2/93.5	54.6/81.4	55.6/91.5	74.6/-	67.9/94.3	85.7/97.1	82.1/96.8	**88.5/97.6**
4	VisA	75.0/65.4	75.3/93.3	71.7/87.1	78.0/93.5	84.1/95.2	87.7/-	90.8/97.5	93.1/97.3	91.2/97.8	**93.3/98.5**
	MVTecAD	75.5/64.0	82.5/94.9	80.6/90.2	84.8/94.5	94.6/94.2	94.5/-	94.2/96.9	96.6/94.8	**97.1**/97.0	96.8/**97.5**
	BTAD	81.7/65.5	89.9/95.8	84.0/89.4	90.8/94.9	87.2/95.8	91.7/-	91.5/**96.8**	93.3/**96.8**	93.1/96.1	**95.0/96.8**
	MVTec3D	62.3/78.6	62.8/94.5	61.5/87.1	62.3/96.7	76.0/97.0	69.1/-	82.4/97.9	84.2/98.0	**86.9**/95.4	85.0/**98.7**
	MVTecLOCO	64.0/66.8	54.7/67.9	61.7/69.4	56.6/66.1	65.5/65.2	60.8/-	70.0/69.0	69.6/69.3	**71.5**/66.9	70.5/**76.9**
	BraTS	66.3/94.8	60.6/94.5	71.2/95.9	60.0/87.3	67.3/93.2	76.9/-	84.6/96.1	86.8/97.3	84.9/97.1	**90.5/97.9**
8	VisA	73.7/94.2	77.9/95.0	85.4/93.9	80.0/95.3	85.4/95.4	88.7/-	92.3/97.8	**94.0**/97.5	93.4/97.7	93.0/**98.6**
	MVTecAD	78.9/95.7	85.0/95.6	90.2/94.8	88.2/95.9	94.8/94.5	95.3/-	97.7/96.7	97.1/95.0	**98.1**/97.1	97.8/**97.6**
	BTAD	84.2/96.3	93.2/97.2	91.4/96.0	91.6/**97.3**	90.2/96.0	89.0/-	91.6/96.8	93.5/96.9	**95.3**/96.4	94.6/97.0
	MVTec3D	63.8/95.6	63.5/96.3	68.3/94.2	67.4/96.9	75.8/97.1	71.4/-	83.0/97.9	86.0/98.1	86.0/**98.9**	**89.0**/98.6
	MVTecLOCO	64.3/67.3	57.7/69.1	65.7/69.7	62.2/69.6	68.0/65.5	62.9/-	69.5/69.3	71.2/69.7	69.7/68.1	**71.7/76.7**
	BraTS	72.6/95.4	71.2/96.0	76.4/96.4	66.6/86.1	68.9/93.8	79.3/-	85.9/96.3	88.7/97.6	88.5/98.0	**90.1/98.1**

Here, $\boldsymbol{v}^{\mathrm{a}} \in \mathcal{V}^{\mathrm{a}}$ is normalized to retain only directional information. The anomaly score of $\boldsymbol{v}^{\mathrm{q}}{}_i \in \mathcal{V}^{\mathrm{q}}$ is computed as the maximum projection onto these abnormal directions. The score is further rescaled to $[0, 1]$ using the mean norm of synthetic abnormal features: $[\hat{\mathbf{S}}^{\mathrm{a}}]_i = \min([\mathbf{S}^{\mathrm{a}}]_i / (\frac{1}{|\mathcal{V}^{\mathrm{a}}|} \sum_{\boldsymbol{v}^{\mathrm{a}} \in \mathcal{V}^{\mathrm{a}}} \|\boldsymbol{v}^{\mathrm{a}}\|_2), 1)$, where values larger than 1 are clipped to emphasize more challenging anomalies, $|\mathcal{V}^{\mathrm{a}}|$ denotes the number of elements in $\mathcal{V}^{\mathrm{a}}$. Finally, the overall anomaly score $\mathbf{S} \in \mathbb{R}^L$ is computed by combining the normal similarity and abnormal similarity:

$$[\mathbf{S}]_i = ((1 - [\mathbf{S}^{\mathrm{n}}]_i) + [\hat{\mathbf{S}}^{\mathrm{a}}]_i)/2 \tag{7}$$

Training and inference. We train VODA in a supervised manner on a source domain and try to generalize to unseen target domains. Based on the anomaly score, we adopt the loss function from Wang et al. (2025b) to jointly optimize image-level and pixel-level predictions,

$$\mathcal{L} = \mathcal{L}_{\text{img}} + \lambda \mathcal{L}_{\text{pix}} \quad where\ \mathcal{L}_{\text{img}} = \text{BCE}(s, y),\ \mathcal{L}_{\text{pix}} = \text{Focal}(\mathbf{S}, \mathbf{M}) + \text{Dice}(\mathbf{S}, \mathbf{M}) \tag{8}$$

Here, $\text{BCE}(\cdot,\cdot)$, $\text{Focal}(\cdot,\cdot)$, $\text{Dice}(\cdot,\cdot)$ denote the binary cross-entropy loss, Focal loss (Lin et al., 2017), and Dice loss (Li et al., 2020), respectively. λ is used to balance image- and pixel-level losses, which is set to 1 for best performance (see Appendix D.1). The image-level anomaly score s is defined as the average of the top 1% highest values in $\mathbf{S}$. y and $\mathbf{M}$ denote the ground-truth image label and pixel-level anomaly mask, respectively.

During inference, given a query image and its normal references, we first generate synthetic abnormal references, and then compute both image-level and pixel-level anomaly scores as described above to obtain the final detection results.

4 Experimental results

We evaluate the effectiveness of the proposed method from three perspectives: (i) its ability to generalize across the three generalization gaps: defect types, object categories, and data domains; (ii) its plug-and-play compatibility with different backbones, as well as its competitive performance compared to settings that use real abnormal references; (iii) an ablation study to analyze the contribution of each technical component. More experimental results are provided in Appendix E.

4.1 Experimental Setup

Datasets. To demonstrate the broad applicability of our VODA model, we conduct extensive experiments on 6 datasets spanning industrial and medical domains. The industrial domain includes several representative benchmarks, namely MVTecAD (Bergmann et al., 2019), VisA (Zou et al., 2022), BTAD (Mishra et al., 2021), MVTec3D (Bergmann et al., 2021), and MVTecLOCO (Bergmann et al., 2022). In addition, the medical domain is represented by the BraTS (Menze et al., 2015).

Implementation and evaluation metrics. We use a DINOv2-S (Oquab et al., 2024) pre-trained Vision Transformer (ViT) as the feature extractor, with all its parameters frozen while only training

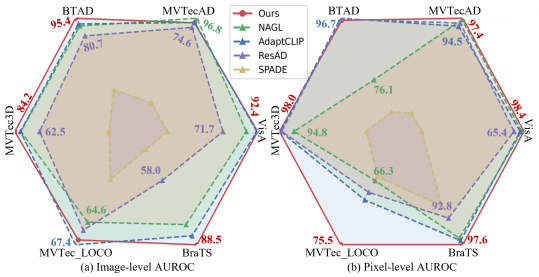


Figure 4: 2-shot performance comparison across different datasets. Results are normalized per dataset for better visualization.

Table 2: VODA with more complex backbones.

Dataset	DINOv2-S	DINOv2-B	DINOv2-L
VISA	**92.4/98.4**	**92.4**/97.7	92.2/97.9
MVTecAD	95.7/**97.4**	**96.0**/97.1	95.8/96.3
BTAD	95.4/96.0	95.1/97.0	**95.9/97.3**
MVTecLOCO	66.7/**75.5**	68.3/75.0	**71.8**/74.5
BraTS	88.5/97.6	89.2/98.2	**90.7/98.8**

Table 3: VODA with synthetic/real $\mathcal{A}$.

VODA		VisA	MVTec AD	BTAD	MVTec 3D	MVTec LOCO	BraTS
Ima.	real	**93.3**	96.5	**95.0**	**85.3**	**70.6**	90.4
	syn.	**93.3**	**96.8**	**95.0**	85.0	70.5	**90.5**
Pix.	real	**98.5**	**97.5**	96.7	98.3	**77.0**	**98.0**
	syn.	**98.5**	**97.5**	**96.8**	**98.7**	76.9	97.9

the introduced adaptation block. Following prior work (Yao et al., 2024a; 2025), we train on one dataset and evaluate on the others. The number of few-shot normal reference samples is set to $K \in \{2, 4, 8\}$. All experiments are conducted on a single NVIDIA A40 GPU. See Appendix C for more implementation details. Following Yao et al. (2024a), we evaluate anomaly detection at both image and pixel levels using AUROC (See Appendix E.3 for more evaluation metrics).

Comparison Methods. We compare VODA with typical few-shot approaches grouped by the type of reference images used during inference. Full-shot AD methods (PaDiM (Defard et al., 2021a), PatchCore (Roth et al., 2022), and SPADE (Cohen & Hoshen, 2020)) are adapted to the few-shot setting using a few normal samples for distance-based scoring (Yao et al., 2024a). Methods including RegAD(Huang et al., 2022), WinCLIP(Jeong et al., 2023a), InCTRL(Zhu & Pang, 2024), ResAD[2](Yao et al., 2024a), and AdaptCLIP(Gao et al., 2026) rely only on normal reference images, while NAGL uses both normal and abnormal references. Among them, AdaptCLIP, ResAD, and NAGL serve as strong baselines due to their superior cross-domain generalization ability.

4.2 Generalization Results Across Three Gaps

To demonstrate the strong domain generalization ability under limited source data (*i.e.,* few-shot normal references), we follow the one-to-many setting (Wang et al., 2025b), training on MVTecAD and evaluating on other datasets, and vice versa for MVTecAD using VisA as the training set.

Quantitative comparison. The quantitative results of our VODA across different datasets under different few-shot settings are summarized in Table 1. Although trained using synthetic defects, our method generalizes effectively to real defect detection and consistently outperforms methods that either do not use reference images or rely solely on normal references across all shot settings (except for a few cases where it is surpassed by ResAD, RegAD or AdaptCLIP). Notably, our method is even comparable to, and often surpasses, NAGL, which relies on real abnormal references. When extended to the medical dataset BraTS, our method outperforms other methods by a large margin, further verifying its effectiveness in mitigating domain gaps. Figure 4 presents a radar plot comparing selected competing methods. In terms of image-level AUROC, our method performs comparably to AdaptCLIP and outperforms other methods by a large margin on most datasets. In terms of pixel-level AUROC, our method consistently achieves the best performance across most datasets, with particularly significant improvements on MVTecLOCO. Although NAGL demonstrates competitive performance in image-level AUROC, it yields the second-lowest pixel-level AUROC in most cases. Since industrial datasets span diverse object categories and defect types, the consistent performance across datasets demonstrates that VOD effectively mitigates both category and defect-type gaps. See Appendix E.5 for controlled evaluations of each gap.

Qualitative comparison. Several straightforward visualization results from industrial and medical datasets under the 4-shot setting are shown in Figure 5. ResAD exhibits false alarms in anomaly localization when generalizing. NAGL produces noisy predictions in the background, *e.g.,* assigning relatively high responses to background regions in the 1st, 2nd, and last cases. In comparison, our method achieves more accurate anomaly localization and is less affected by background noise.

4.3 Effect of Backbone Complexity and Real Abnormal References

We further analyze performance from two perspectives: (i) increasing backbone capacity under the 2-shot setting (*i.e.,* using DINOv2-Base or DINOv2-Large instead of DINOv2-Small), and (ii) replacing synthetic abnormal references with real ones under the 4-shot setting. Results are reported in Tables 2

[2]We adopt the ImageBind-based variant for improved performance.

Table 4: Ablation studies. ✗ indicates that the corresponding component is removed.

	Adaptation	VOD	$\mathbf{S}^{n}$	$\hat{\mathbf{S}}^{a}$	Datasets VisA	MVTecAD	BTAD	MVTec3D	MVTecLOCO	BraTS
i	✗				87.5/96.7	95.2/96.6	89.5/96.2	80.7/97.3	68.6/76.4	**97.2/97.9**
ii		✗			90.1/97.2	94.7/96.8	93.4/93.4	81.2/97.6	69.4/**77.1**	81.9/96.2
iii			✗		**93.4/98.5**	96.6/**97.6**	91.5/93.4	84.7/92.2	**70.8**/76.9	89.9/97.8
iv				✗	90.5/97.6	96.7/97.4	94.6/96.6	83.5/97.8	69.6/76.8	85.8/97.7
v					93.3/**98.5**	**97.8**/97.5	**95.1/96.8**	**85.0/98.7**	70.5/76.9	90.5/**97.9**

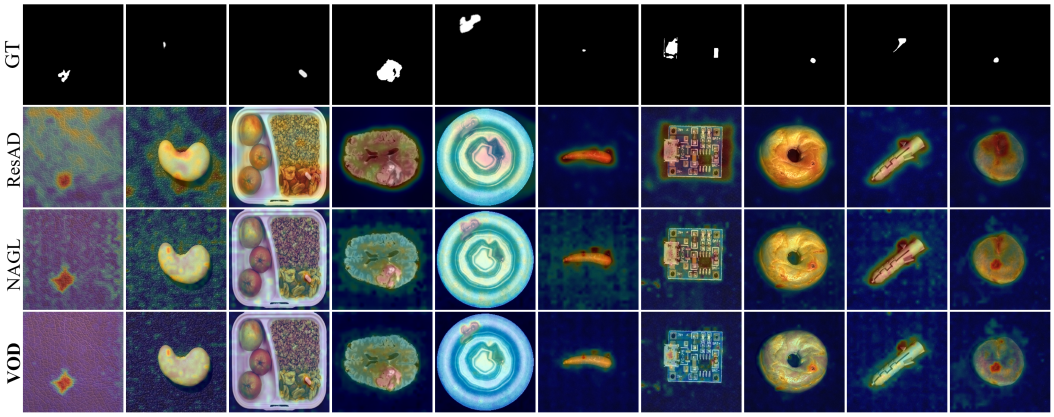


Figure 5: Qualitative Comparison of anomaly detection results across different datasets.

and 3. We observe that: (i) as the backbone complexity increases, the performance improves steadily in most cases, with DINOv2-Large achieving the best results at the cost of additional computational overhead; (ii) incorporating real abnormal references does not lead to noticeable performance gains across all test datasets, which further supports the effectiveness of our proposed value space in enabling synthetic abnormalities to serve as reliable proxies for real ones.

4.4 Ablation Study

To evaluate the individual contribution of each proposed component to the overall performance, we remove each component in turn and report the results in Table 4 (see Appendix E.1 for more results). (i) Without adaptation (row i), both feature representation and separability are entirely determined by the pre-trained model, leading to the lowest performance among all variants. An exception is observed on BraTS, where the model without adaptation unexpectedly outperforms its adapted counterparts. We attribute this to the domain bias introduced by adapting on MVTecAD, which shifts the model toward the industrial domain and consequently degrades its generalization to the medical domain. (ii) When adaptation is enabled (row ii), VOD contributes more significantly to performance improvements than $\mathbf{S}^{n}$ and $\hat{\mathbf{S}}^{a}$ (as observed on VisA, MVTecAD, MVTec3D, and BraTS). This result highlights the importance of aligning both normal and abnormal distributions across the three generalization gaps via value-order decomposition. (iii) Comparing row iii and iv, $\hat{\mathbf{S}}^{a}$ has a larger impact than $\mathbf{S}^{n}$ on the final performance (see ima. AUROC on VisA, MVTec3D, MVTecLOCO, and BraTS), demonstrating the importance of incorporating synthetic abnormal samples.

5 Conclusions

With the goal of GAD, we propose VODA, a simple yet effective method trained on a single source dataset and generalizes to unseen target domains. We identify three generalization gaps across object categories, defect types, and data domains, and leverage VOD technique to disentangle representations into two components: value and order. The value component serves as a gap-invariant indicator of abnormality, particularly robust across real and synthetic anomalies, while the order component captures gap-specific variations. This disentanglement may provide insights for more robust and generalizable anomaly detection. Based on the value component, we build VODA framework using only a few-shot set of normal references. Experimental results show consistent improvements over existing methods in cross-domain generalization. In addition, the identification and localization of logical anomalies remain a frontier problem. Our method provides a potential direction for addressing this problem, and extending it to logical anomaly detection by incorporating global structure modeling will be explored in future work. Beyond the above discussions, further analysis of the limitations and applicability of VOD is provided in Appendix F.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript to improve readability and language clarity, which falls under the recommended disclosure category. We have reviewed all AI-assisted work and verified that the resulting text accurately reflects the authors' intended content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *CVPR*, pp. 4042–4053, 2024.

Manuel Barusco, Francesco Borsatti, Davide Dalle Pezze, Francesco Paissan, Elisabetta Farella, and Gian Antonio Susto. Paste: Improving the efficiency of visual anomaly detection at the edge. In *CVPR*, pp. 4026–4035, 2025.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. *arXiv preprint arXiv:2112.09045*, 2021.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4):947–969, 2022.

Tri Cao, Jiawen Zhu, and Guansong Pang. Anomaly detection under distribution shift. In *ICCV*, pp. 6511–6523, 2023.

Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In *ECCV*, pp. 37–54. Springer, 2024.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pp. 3606–3613, 2014.

Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

Dinu Coltuc, Philippe Bolon, and J-M Chassery. Exact histogram specification. *IEEE Transactions on Image processing*, 15(5):1143–1152, 2006.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (eds.), *Pattern Recognition. ICPR International Workshops and Challenges*, pp. 475–489, Cham, 2021a. Springer International Publishing. ISBN 978-3-030-68799-1.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pp. 475–489. Springer, 2021b.

Bin-Bin Gao. Learning to detect multi-class anomalies with just one normal image prompt. In *ECCV*, pp. 454–470. Springer, 2024.

Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. Adaptclip: Adapting clip for universal visual anomaly detection. In *AAAI*, volume 40, pp. 4095–4103, 2026.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Univad: A training-free unified model for few-shot visual anomaly detection. In *CVPR*, pp. 15194–15203, 2025.

Guan Gui, Bin-Bin Gao, Jun Liu, Chengjie Wang, and Yunsheng Wu. Few-shot anomaly-driven generation for anomaly classification and segmentation. In *ECCV*, pp. 210–226. Springer, 2024.

Ernest L Hall. Almost uniform distributions for computer image enhancement. *IEEE Transactions on Computers*, 100(2):207–208, 1974.

Ernest L Hall. Almost uniform distributions for computer image enhancement. *IEEE Transactions on Computers*, 100(2):207–208, 2006.

Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *ICCV*, pp. 8791–8800, 2021.

Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 303–319, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20053-3.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pp. 1501–1510, 2017.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19606–19616, June 2023a.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, pp. 19606–19616, 2023b.

Jiarui Lei, Xiaobo Hu, Yue Wang, and Dong Liu. Pyramidflow: High-resolution defect contrastive localization using pyramid normalizing flow. In *CVPR*, pp. 14143–14152, 2023.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, pp. 9664–9674, 2021.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. In *ACL*, pp. 465–476, 2020.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pp. 2980–2988, 2017.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *CVPR*, pp. 20402–20411, 2023.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pp. 01–06, 2021. doi: 10.1109/ISIE45552.2021.9576231.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 577–582. Association for Computing Machinery, 2023.

Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985.

Jannick P Rolland, V Vo, B Bloss, and Craig K Abbey. Fast algorithms for histogram matching: Application to texture synthesis. *Journal of Electronic Imaging*, 9(1):39–45, 2000.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, pp. 14318–14328, 2022.

Dori Shapira, Shai Avidan, and Yacov Hel-Or. Multiple histogram matching. In *2013 IEEE international conference on image processing*, pp. 2269–2273. IEEE, 2013.

Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *CVPR*, pp. 24511–24520, 2023.

Shenzhi Wang, Liwei Wu, Lei Cui, and Yujun Shen. Glancing at the patch: Anomaly localization with global and local feature comparison. In *CVPR*, pp. 254–263, 2021.

Xu Wang, Jingyuan Zhuo, Zhiyuan You, Zhiyu Tan, Yikuan Yu, Siyu Wang, and Xinyi Le. ReinAD: Towards real-world industrial anomaly detection with a comprehensive contrastive dataset. In *NeurIPS*, 2025a.

Yuexin Wang, Xiaolei Wang, Yizheng Gong, and Jimin XIAO. Normal-abnormal guided generalist anomaly detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Tiange Xiang, Yixiao Zhang, Yongyi Lu, Alan L Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. Squid: Deep feature in-painting for unsupervised anomaly detection. In *CVPR*, pp. 23890–23901, 2023.

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 *CVPR*, pp. 20370–20382, 2025.

Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning semantic context from normal samples for unsupervised anomaly detection. In *AAAI*, volume 35, pp. 3110–3118, 2021.

Xincheng Yao, Ruoqi Li, Zefeng Qian, Yan Luo, and Chongyang Zhang. Focus the discrepancy: Intra-and inter-correlation learning for image anomaly detection. In *ICCV*, pp. 6803–6813, 2023a.

Xincheng Yao, Chongyang Zhang, Ruoqi Li, Jun Sun, and Zhenyu Liu. One-for-all: Proposal masked cross-class anomaly detection. In *AAAI*, volume 37, pp. 4792–4800, 2023b.

Xincheng Yao, Zixin Chen, Chao Gao, Guangtao Zhai, and Chongyang Zhang. Resad: A simple framework for class generalizable anomaly detection. *NeurIPS*, 37:125287–125311, 2024a.

Xincheng Yao, Ruoqi Li, Zefeng Qian, Lu Wang, and Chongyang Zhang. Hierarchical gaussian mixture normalizing flow modeling for unified anomaly detection. In *ECCV*, pp. 92–108. Springer, 2024b.

Xincheng Yao, Chao Shi, Muming Zhao, Guangtao Zhai, and Chongyang Zhang. Resad++: Towards class agnostic anomaly detection via residual feature learning. *arXiv preprint arXiv:2509.23741*, 2025.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *NeurIPS*, 35:4571–4584, 2022.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, pp. 8330–8339, 2021.

Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *CVPR*, pp. 16699–16708, 2024.

Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *CVPR*, pp. 8035–8045, 2022.

Ying Zhao. Omnial: A unified cnn framework for unsupervised anomaly localization. In *CVPR*, pp. 3924–3933, 2023.

Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *CVPR*, pp. 17826–17836, 2024.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pp. 392–408. Springer, 2022.

APPENDIX

A THEORETICAL PROOF OF VOD

The effectiveness of VOD is motivated by the observation that different aspects of visual representations can be encoded by complementary feature statistics. EFDM (Zhang et al., 2022) shows that feature ordering is closely associated with image content, whereas feature values capture higher-order feature statistics. Motivated by this observation, we investigate whether residual features exhibit similar complementary properties in anomaly detection. We hypothesize that the order component preserves the ordinal structure of feature responses, which is closely related to gap-specific semantic patterns such as object categories, defect types, and domains. In contrast, the value component preserves the empirical distribution of residual magnitudes while removing ordering-dependent semantics, providing a canonical representation that is less sensitive to cross-gap variations. This complementary property forms the theoretical motivation for VOD and explains why the value component is better suited for cross-gap anomaly alignment.

To further formalize and justify Property 1 presented in the main paper, we introduce the following propositions and their proofs.

Proposition 2 (Order Preserves Ordinal Structure)**.** *Let $o(\boldsymbol{r})$ denote the order representation obtained by ranking the elements of $\boldsymbol{r}$. For any strictly increasing function $g(\cdot)$, the order representation is invariant: $o(\boldsymbol{r}) = o(g(\boldsymbol{r}))$.*

Proof. A strictly increasing transformation preserves the ordering between every pair of feature responses. Therefore, for any two elements $[\boldsymbol{r}]_i$ and $[\boldsymbol{r}]_j$, $[\boldsymbol{r}]_i$ less than $[\boldsymbol{r}]_j$ if and only if $g([\boldsymbol{r}]_i)$ less than $g([\boldsymbol{r}]_j)$. Since the ranking indices depend only on these pairwise ordering relationships, the resulting order representation remains unchanged. □

This proposition shows that the order representation discards absolute magnitudes while preserving the ordinal structure of feature responses. Consequently, replacing feature values with equally spaced ranks preserves the structural layout encoded by feature rankings.

Proposition 3 (Value is Permutation Invariant)**.** *Let $v(\boldsymbol{r})$ denote the value representation obtained by sorting $\boldsymbol{r}$ in descending order. For any permutation matrix P, $v(P\boldsymbol{r}) = v(\boldsymbol{r})$.*

Proof. A permutation only changes the positions of feature values but not the multiset of values themselves. Sorting produces the unique descending ordering of this multiset, so the sorted vector is identical before and after permutation. □

The value representation maps all permutations of the same feature vector to a unique canonical representation. Therefore, ordering-dependent semantic structures are removed, while the empirical distribution of feature magnitudes is preserved.

Corollary 4 (Canonical Property of the Value Component)**.** *All feature vectors sharing the same set of feature values but different ordering are mapped to the same value component. Therefore, the value component serves as a canonical representation of feature magnitudes.*

This corollary follows directly from Proposition 3, which shows that the value component is invariant to any permutation of feature ordering. As a result, multiple ordering-dependent feature vectors collapse into a unique value component while preserving the complete set of feature magnitudes. This canonical property reduces variability caused by gap-specific ordering and makes the value component naturally suited for aligning anomalies across heterogeneous object categories, defect types, and domains.

Assumption and Scope. Our motivation relies on the assumption that anomaly cues are primarily reflected by abnormal residual magnitudes, whereas gap-specific semantics are mainly encoded by feature ordering. This assumption is consistent with residual features extracted from frozen vision foundation models, where anomalous regions typically induce unusually strong responses regardless of object identity. However, if anomalies are expressed purely through global geometric transformations without significant magnitude changes, the value representation may lose discriminative structural cues.

Further experimental validations are provided as follows.

We first validate the hypothesis through a comprehensive visualization of normal and abnormal patch distributions under different representation spaces across distinct types of generalization gaps (including both real and synthetic defects), as shown in Figure 6. The detailed implementation of this visualization is provided in Appendix E.4. Figure 6 (c) shows that the value component can capture transferable gap-invariant characteristics, aligning normal and abnormal patches within each category across all gaps. We also investigate the order component, and the results in Figure 6 (e) demonstrate that its distribution is more strongly associated with object categories, defect types, and data domains, while still preserving gap-sensitive information.

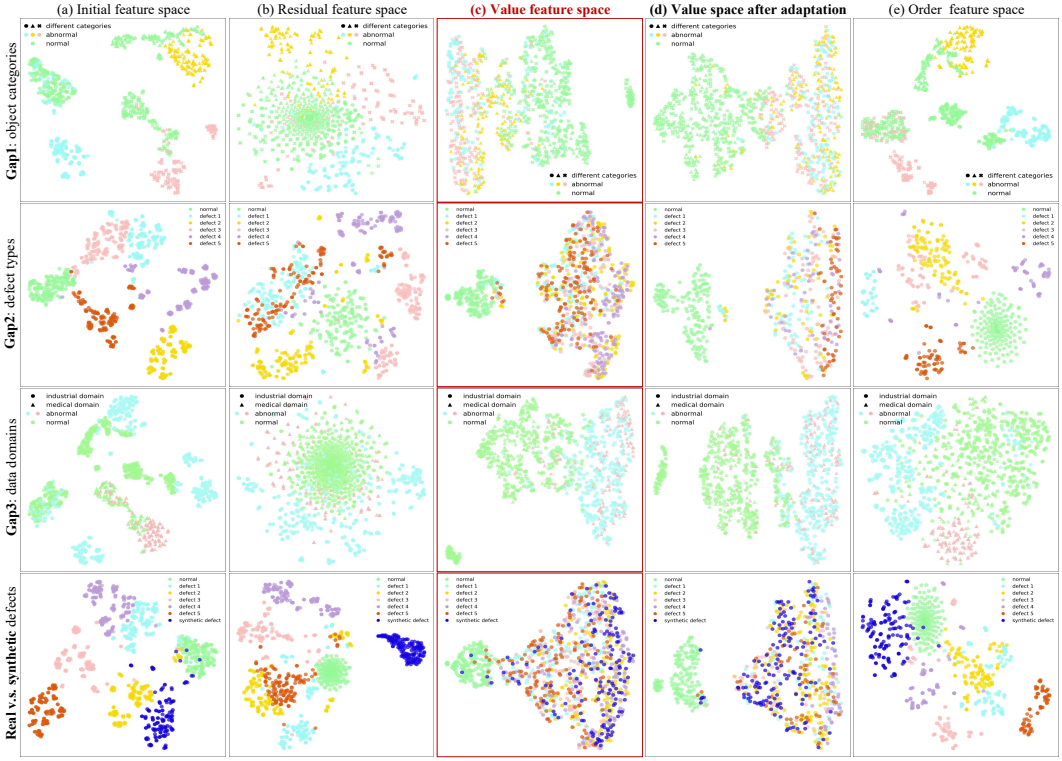


Figure 6: t-SNE visualization of normal and anomalous patches under different representations. (a) The initial feature space exhibits three distinct gaps, which hinder generalization and negatively affect anomaly detection performance. (b) Residual-space methods mainly improve the alignment among normal samples. (c) In contrast, the value component of VOD enforces intra-class alignment for both normal and anomalous samples. (d) After adaptation, the value component further enhances the separability between normal and anomalous samples. (e) The order component of VOD shows a similar variation across the three generalization gaps.

To provide quantitative evidence beyond visualization, we conduct additional linear probing experiments. Specifically, we perform linear probing on the value and order components separately, using them to predict object categories, defect types, domains, and normal/abnormal status. The results in Table 5 show that the value component contains substantially less gap-specific information while preserving strong anomaly discrimination capability, whereas the order component retains substantially more gap-specific information and shows weaker anomaly discrimination capability. These results support our claim that VOD helps mitigate the three generalization gaps by isolating transferable residual characteristics.

In fact, the nearest-neighbor residual representation already reduces category-related appearance variations by measuring deviations from the closest normal reference. Based on this residual representation, VOD decomposes each residual vector into value and order components. The value component preserves the complete set of residual magnitudes while removing explicit channel correspondence, making it invariant to channel permutations and primarily characterizing the magnitude distribution

Table 5: classification accuracy (%) for different tasks.

Task	obj. cate.	defect type	domain	normality
Value	41.47	19.13	35.96	**92.37**
Order	**99.18**	**94.67**	**96.44**	54.08

of residual responses. In contrast, the order component preserves the relative ranking of channel responses, which empirically exhibits stronger correlations with object categories, defect types, and domain variations. Therefore, by retaining the value component and discarding the order component, VOD introduces an inductive bias that suppresses gap-sensitive ordering information while preserving transferable anomaly-related residual characteristics.

Finally, we would like to emphasize that the disentanglement capability of VOD does not arise from the sorting operation alone. Instead, VOD should be viewed as a canonical decomposition of residual representations into value and order components, and the original residual representation can be exactly reconstructed from the joint value and order components. To the best of our knowledge, this is the first work to explicitly decompose residual representations into value and order components for reference-based anomaly detection, providing a new perspective for mitigating reference-query discrepancies.

B DATASETS

The details of the datasets used in this work are summerized in Table 6

Table 6: Statistics of the datasets used in this work.

Data name	# Object categories	# Defect types	# Normal test samples	# Abnormal test samples
MVTecAD	15	73	467	1258
VisA	12	1	962	1200
BTAD	3	1	451	290
MVTec3D	10	40	249	948
MVTecLOCO	5	10	575	993
BraTS	1	1	154	1097

MVTecAD (Bergmann et al., 2019). MVTecAD is a widely used benchmark for unsupervised anomaly detection. It contains 5,354 high-resolution RGB images from 15 categories, including both textures (5 categories) and objects (10 categories). The training set consists of 3,629 normal images, while the test set includes 1,725 samples with both normal and anomalous instances. The anomalies span more than 70 defect types, such as scratches, crack, contamination, and structural irregularities.

VisA (Zou et al., 2022). VisA is another popular benchmark dataset for anomaly detection, comprising 10,821 high-resolution color images. It includes 9,621 normal samples and 1,200 anomalous samples across 12 object categories in three different domains. The dataset is particularly challenging due to its diverse and complex visual conditions, including industrial PCBs with intricate layouts, multiple objects with significant variations in pose and position, and loosely aligned instances with subtle appearance differences. The anomalies involve a wide range of defects, covering both surface-level and structural variations.

BTAD (Mishra et al., 2021). BTAD is designed for multiple vision tasks, including instance segmentation, semantic segmentation, and object detection. It consists of real-world industrial images containing both surface and structural defects.

MVTec3D (Bergmann et al., 2021). MVTec3D is a dataset for unsupervised anomaly detection and localization in 3D industrial scenarios, where certain defects manifest as geometric inconsistencies.

It provides 4,147 point cloud scans along with corresponding 2D RGB images across 10 categories. As our method operates on 2D inputs, we only utilize the RGB images.

MVTecLOCO (Bergmann et al., 2022). MVTecLOCO focuses on unsupervised anomaly detection involving both structural and logical anomalies in industrial inspection tasks. The dataset includes 1,772 training images, 304 validation images, and 1,568 test images across five categories, each defined by specific logical rules. For example, a breakfast box is required to contain exactly two tangerines and one nectarine arranged on the left side, while a screw bag must include two washers, two nuts, one long screw, and one short screw. Any deviation from these rules is considered a logical anomaly.

BraTS (Menze et al., 2015). BraTS is a well-established multimodal MRI dataset for brain tumor segmentation, released annually as part of the MICCAI challenges since 2012. In this work, we adopt the 2021 BraTS dataset, which includes 1,251 labeled cases, comprising 1,097 tumor cases and 154 normal cases without tumors.

C Experimental setup

Given the few-shot normal references, we adopt the synthetic anomaly generation strategy from DRAEM (Zavrtanik et al., 2021). Specifically, for each normal reference, we first generate a noise image using a Perlin noise generator to model diverse anomaly shapes, and then binarize it with a threshold such that only 20% of the values remain non-zero, forming a mask M. We randomly sample a texture image from DTD (Cimpoi et al., 2014) as the anomaly source. To further enhance the distinctiveness of the anomalous regions, we apply random augmentation by sampling three operations from the set {posterize, sharpness, solarize, equalize, brightness change, color change, auto-contrast}.

Implementation Details. We adopt a DINOv2 (Oquab et al., 2024) pre-trained Vision Transformer (ViT) as the feature extractor and freeze all its parameters during training. For the adaptation block, we adopt a Transformer encoder layer with 4 attention heads to improve the discriminability between normal and abnormal features. The block consists of a feedforward network with an expansion dimension of 1536, uses GELU activation to enhance representational capacity, and applies a dropout rate of 0.1 for regularization. The backbones used by the compared methods are summarized in Table 7.

Our method is trained for 20 epochs, with each epoch consisting of 500 sampled episodes and a batch size of 8. We use the AdamW optimizer with a weight decay of 1×10^{-5}. The initial learning rate is set to 1×10^{-5} and decayed by a factor of 0.1 at epochs 10 and 15. All training and testing images are resized to 448×448, and all experiments are conducted on a single NVIDIA A40 GPU. Following prior work (Yao et al., 2024a; 2025), we train the model on one dataset and evaluate it on the others. The number of few-shot normal reference samples is set to $K \in \{2, 4, 8\}$. Our implementation is based on the publicly released codebase of NAGL (Wang et al., 2025b), and we follow the same model selection protocol, where the validation set is constructed from VisA. We thank the authors for their contribution to the community.

Table 7: Backbone and parameter size of each compared method.

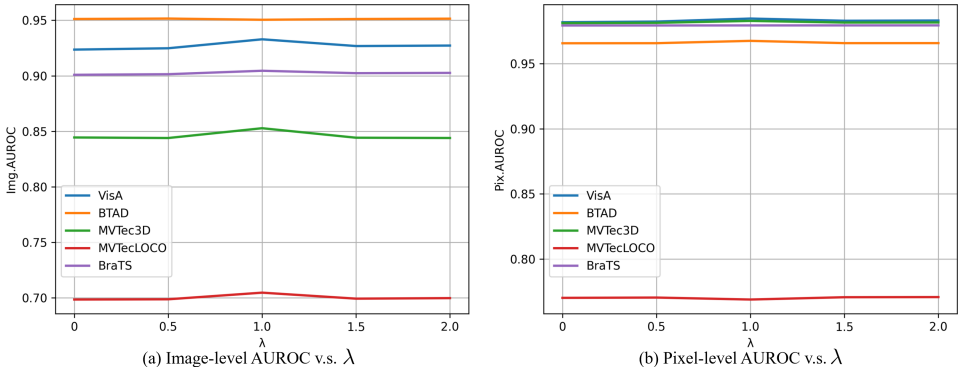
Methods	SPADE	PaDiM	Patchcore	RegAD	WinCLIP	InCTRL	ResAD	AdaptCLIP	NAGL	**VODA**
Backbones	WideResNet50-2			ResNet-18	ViT-B/16+		ImageBind-Huge	ViT-L/14@336px	DINOv2-Small	
Paramters	68.9 M			11.7 M	86.6 M		2.3 B	303.9 M	21.7 M	

D Hyperparameter sensitivity analysis

D.1 The effect of the hyperparameter λ

The hyperparameter λ is used to balance the image-level and pixel-level prediction losses. We analyze its effect on the final performance by varying λ from 0 to 2 with a step size of 0.5, where $\lambda = 0$ corresponds to using only the image-level prediction loss. The results, in terms of image-level and pixel-level AUROC across five test datasets, are shown in Figure 7. We observe that our chosen setting

$\lambda = 1$ consistently achieves the best performance on both metrics across all datasets, highlighting the importance of jointly considering image-level and pixel-level supervision.



(a) Image-level AUROC v.s. λ (b) Pixel-level AUROC v.s. λ

Figure 7: The effect of λ on our VODA.

D.2 Selection of the number of synthentic abnormal references

We investigate the effect of varying the number of synthetic abnormal references generated for each normal image during both training and inference. As shown in Table 8, increasing the number of synthetic abnormal references yields only marginal performance gains, while each additional reference incurs extra similarity computations and leads to an approximately linear increase in inference cost. Therefore, we use a single synthetic abnormal reference as a favorable trade-off between detection performance and computational efficiency.

Table 8: Performance comparison under different numbers of synthentic abnormal references

Dataset	Synth. Refs/Normal		
	1	5	10
VisA	92.4/98.4	92.5/98.4	92.4/98.5
BraTS	88.5/97.6	88.6/97.5	88.6/97.7

E Additional experimental results

E.1 Contributions of the similarity scores $\mathbf{S}^{\mathrm{N}}$ and $\hat{\mathbf{S}}^{\mathrm{A}}$

To better understand the contributions of the similarity scores $\mathbf{S}^{\mathrm{n}}$ and $\hat{\mathbf{S}}^{\mathrm{a}}$, we further visualize them, along with their combined anomaly score $\mathbf{S}$, in Figure 8. In most cases, $\mathbf{S}^{\mathrm{n}}$ is able to localize the most salient abnormal regions but tends to miss less prominent ones, whereas $\hat{\mathbf{S}}^{\mathrm{a}}$ provides more precise localization of fine-grained abnormal details. By combining the two, the final anomaly segmentation results are more accurately aligned with the ground truth.

E.2 Efficiency analysis

All experiments are conducted on a single NVIDIA A40 GPU. We compare the efficiency of our method with NAGL, AdaptCLIP, and ResAD in terms of the total number of parameters, training time, peak GPU memory and inference speed.

As shown in Table 9, the parameter size has a significant impact on training time and inference speed. For example, ResAD and AdaptCLIP have substantially larger parameter sizes than NAGL and VODA, and thus exhibit lower efficiency. Our method introduces slightly fewer parameters than NAGL, while achieving slightly lower inference speed and modestly higher peak GPU memory. This is due to the additional process of generating synthetic abnormal references. Overall, VODA achieves a favorable balance between efficiency and performance.

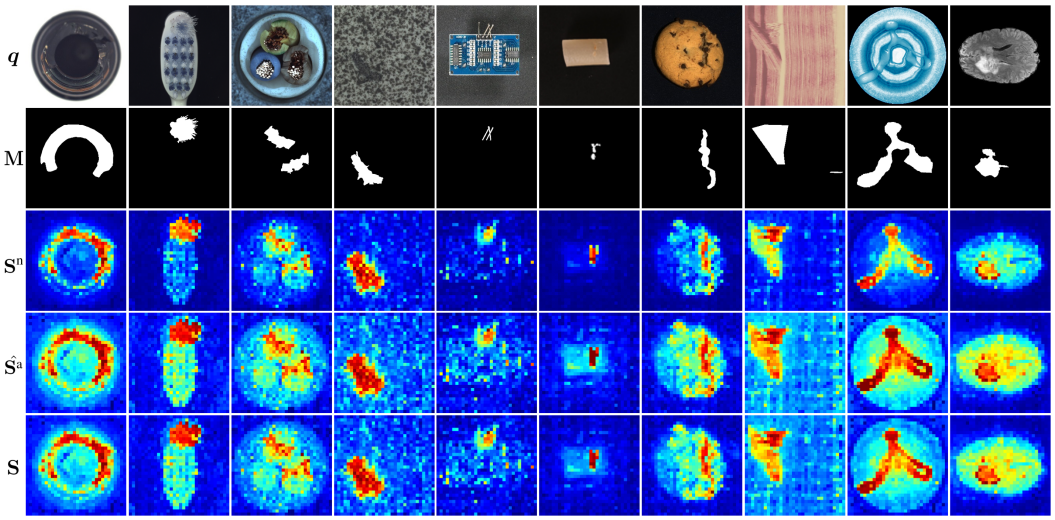


Figure 8: Visualize the contribution of the similarity scores $\mathbf{S}^{\mathrm{n}}$ and $\hat{\mathbf{S}}^{\mathrm{a}}$.

Table 9: The efficiency comparison under the 4-shots setting.

Method	Total Parameters (M)	Training Time (H)	GPU memory(GB)	Inference Speed (FPS)
ResAD	59.2	20.6	2.1	7.8
AdaptCLIP	429.7	1.5	30.0	5.9
NAGL	24.4	0.3	0.2	17.1
VODA	23.8	0.6	0.3	14.3

E.3 Results with more metrics

To provide a more comprehensive evaluation of VODA, we further adopt additional metrics following Wang et al. (2025b); Gao et al. (2026). For image-level anomaly detection, we report the Area Under the Receiver Operating Characteristic Curve (AUROC), the maximum F1-score at the optimal threshold (F1-max), and Average Precision (AP). For pixel-level anomaly localization, we additionally employ the per-region overlap (PRO) metric together with AUROC and F1-max. Each metric is explained as follows.

- **AUROC.** AUROC measures the area under the Receiver Operating Characteristic (ROC) curve and evaluates the overall ranking ability of anomaly scores across all thresholds. A higher AUROC indicates better separability between normal and anomalous samples. For pixel-level evaluation, it may be biased toward large anomalous regions since larger defects contribute more pixels.
- **F1-max.** F1-max is the maximum F1-score achieved over all possible decision thresholds, reflecting the best achievable detection performance under optimal threshold selection.
- **AP.** Average Precision (AP) is the area under the Precision-Recall (PR) curve, which evaluates the overall ability of the model to detect anomalous samples by balancing precision and recall across all thresholds. It is particularly suitable for anomaly detection tasks with severe class imbalance.
- **PRO.** Per-Region Overlap (PRO) measures localization quality at the region level by computing the average correctly classified pixel ratio for each connected anomalous region under different false positive rates (typically within $[0, 0.3]$). The PRO score is defined as the normalized area under this curve. In this work, we report the PRO score with false positive rates as 0.3. A higher PRO indicates better localization performance for both large and small anomalies.

We compare VODA with existing methods on additional image-level (AUROC, AP, and F1-max) and pixel-level (AUROC, PRO, and F1-max) metrics. The results on MVTec and VisA under 2-shot and 4-shot settings are summarized in Table 10. Overall, VODA achieves the best performance on the

majority of metrics across both datasets and shot settings. Notably, even compared with NAGL, which leverages real anomaly samples as references, VODA outperforms it in most cases. Additionally, as the number of reference shots increases, VODA consistently achieves improved performance across all metrics, highlighting its strong scalability.

Table 10: Comparison with existing methods on MVTec and VisA under 2-shot and 4-shot few shot settings.

Shots	Methods	MVTecAD						VisA					
		Ima.			Pix.			Ima.			Pix.		
		AUROC	AP	F1-max	AUROC	PRO	F1-max	AUROC	AP	F1-max	AUROC	PRO	F1-max
2	SPADE	82.9	91.7	91.1	92.0	85.7	44.5	80.7	82.3	81.7	96.2	85.7	40.5
	PatchCore	86.3	93.8	92.0	93.3	82.3	53.0	81.6	84.8	82.5	96.1	82.6	41.0
	WinCLIP	94.4	97.0	94.4	96.0	88.4	58.4	84.6	85.8	83.0	96.8	86.2	43.5
	ResAD	87.2	93.9	92.2	94.8	85.5	50.2	86.6	88.3	84.1	96.5	82.3	40.0
	AdaptCLIP	95.7	97.9	95.4	94.5	90.5	54.9	92.3	**94.4**	**88.8**	97.1	91.8	45.6
	NAGL	**96.8**	97.9	96.3	96.8	93.2	**59.8**	89.8	90.6	86.8	97.6	91.4	43.3
	VODA	95.7	**98.1**	**96.7**	**97.4**	**93.7**	59.5	**92.4**	92.5	88.4	**98.4**	**92.1**	**46.0**
4	SPADE	84.8	92.5	91.5	92.7	87.0	46.2	81.7	83.4	82.1	96.6	87.3	43.6
	PatchCore	88.8	94.5	92.6	94.3	84.3	55.0	85.3	87.5	84.3	96.8	84.9	43.9
	WinCLIP	95.2	97.3	94.7	96.2	89.0	59.5	87.3	88.8	84.2	97.2	87.6	**47.0**
	ResAD	90.7	95.7	93.9	95.8	88.7	53.0	89.3	90.7	86.5	96.8	84.1	41.6
	AdaptCLIP	96.6	98.3	96.0	94.8	90.8	56.1	93.1	**94.5**	88.8	97.3	**92.2**	46.7
	NAGL	**97.1**	98.0	96.4	97.0	93.5	60.1	91.2	91.7	87.8	97.8	91.5	44.0
	VODA	96.8	**98.7**	**96.9**	**97.5**	**94.0**	**60.6**	**93.3**	93.5	**89.2**	**98.5**	**92.2**	46.1

Futhermore, Table 11 presents a comparison of these metrics on two challenging datasets under 2-shot setting. Our method achieves the best performance on most localization metrics, demonstrating robust generalization when transferred to more complex industrial scenarios as well as the medical domain.

Table 11: Comparison with several methods on MVTecLOCO and BraTS under 2-shot setting.

Methods	MVTecLOCO						BraTS					
	Ima.			Pix.			Ima.			Pix.		
	AUROC	AP	F1-max	AUROC	PRO	F1-max	AUROC	AP	F1-max	AUROC	PRO	F1-max
AdaptCLIP	**67.4**	77.2	77.7	69.2	66.0	20.8	84.4	97.2	93.6	97.1	81.3	57.4
NAGL	64.6	77.4	78.0	66.3	**69.3**	18.5	82.1	96.6	93.9	96.8	79.9	51.8
VODA	66.7	**77.6**	**78.6**	**75.5**	66.1	**25.2**	**88.5**	**98.0**	**94.5**	**97.6**	**81.8**	**57.5**

We additionally report the results of VODA under 2/4/8-shot settings for all datasets in Table 12, 13, and 14 for completeness.

Table 12: VODA results with additional metrics under the 2-shot setting.

Metrics		Datasets					
		VisA	MVTecAD	BTAD	MVTec3D	MVTecLOCO	BraTS
Ima.	AUROC	92.4	95.7	95.4	82.4	66.7	88.5
	AP	92.5	98.1	97.5	94.8	77.6	98.0
	F1	88.4	96.7	93.6	91.4	78.6	94.5
Pix.	PRO	92.1	93.7	76.6	93.7	66.1	81.8
	AUROC	98.4	97.4	96.0	98.0	75.5	97.6
	F1	46.0	59.5	56.4	50.1	25.2	57.5

Table 13: VODA results with additional metrics under the 4-shot setting.

Metrics		Datasets					
		VisA	MVTecAD	BTAD	MVTec3D	MVTecLOCO	BraTS
Ima.	AUROC	93.3	96.8	95.0	85.0	70.5	90.5
	AP	93.5	98.7	97.5	95.5	80.1	98.3
	F1	89.2	96.9	93.6	92.6	79.1	94.8
Pix.	PRO	92.2	94.0	74.5	95.9	66.0	82.9
	AUROC	98.5	97.5	96.8	98.7	76.9	97.9
	F1	46.1	60.6	58.4	52.2	27.0	60.8

Table 14: VODA results with additional metrics under the 8-shot setting.

Metrics		Datasets					
		VisA	MVTecAD	BTAD	MVTec3D	MVTecLOCO	BraTS
Ima.	AUROC	93.0	97.8	94.6	89.0	71.1	90.1
	AP	93.0	98.9	97.3	96.9	81.3	98.1
	F1	89.0	97.3	92.7	93.0	79.4	95.0
Pix.	PRO	92.7	94.2	75.9	95.2	66.2	83.5
	AUROC	98.6	97.6	97.0	98.6	76.7	98.1
	F1	45.7	97.8	60.4	54.1	26.7	62.9

E.4 Full visualization of different representations across the three generalization gaps

A comprehensive visualization of normal and abnormal patch distributions under different representation spaces across the three generalization gaps is shown in Figure 6. Each row is constructed as follows:

- For **Gap1**, we select three objects (*i.e.,* bottle, carpet, and hazelnut) from the MVTecAD dataset. For each object, 150 image patches are randomly sampled from normal images, and another 150 patches are sampled from anomalous regions in abnormal images for visualization.
- For **Gap2**, we select one object (*i.e.,* carpet) with five defect types from the MVTecAD dataset. Specifically, 150 patches are randomly sampled from normal images, and for each defect type, 150 patches are randomly sampled from the anomaly regions based on the ground-truth masks.
- For **Gap3**, we select an industrial dataset (MVTecAD) and a medical dataset (BraTS) for demonstration. From MVTecAD, we choose three object categories (*i.e.,* bottle, carpet, and hazelnut), and for each category, 150 patches are randomly sampled from normal images and 150 from anomaly regions. For BraTS, which contains a single category (*i.e.,* brain), 150 patches are randomly sampled from normal images and 150 from lesion regions based on the provided masks.
- For **real vs. synthetic** defects, we generate synthetic anomalies on carpet images from the MVTecAD dataset. For each real defect type and synthetic anomaly type, 150 patches are randomly sampled from the corresponding anomaly regions. In addition, 150 patches are sampled from normal images.

E.5 Controlled evaluation of individual gaps

To more directly substantiate the three-gap formulation, we conduct controlled experiments that isolate each gap. Specifically, we consider the following four experimental settings and compare the performance with and without the proposed VOD transformation: (a) **Object-related gap1:** We use MVTecAD and randomly select 8 object categories as the training set, with the remaining

categories used as the testing set. (b) **Object-related gap2:** We select 6 object categories containing color-related defects from MVTecAD. 4 categories are randomly chosen as the training set, while the remaining 2 categories are used as the query/testing set. (c) **Defect-type-related gap:** We use only the pill category from MVTec AD, which contains 7 defect types. 4 defect types are randomly selected for training, while the remaining are used for testing. This setting isolates defect-type variation while keeping the object category unchanged. (d) **Domain-related gap:** We use MVTecAD for training and BraTS for testing to evaluate cross-domain generalization. As shown in Table 15, VOD consistently improves performance across all four settings, demonstrating the robustness of the proposed transformation against each individual gap.

Table 15: Performance comparison with and without VOD under 4 experimental settings.

	our method without VOD	our method with VOD
(a)	93.7/96.9	**95.1/97.5**
(b)	99.1/98.0	**99.3/98.2**
(c)	95.4/98.8	**95.7/99.3**
(d)	81.9/96.2	**90.5/97.9**

E.6 Discussion on the stability of the VODA performance

We report AUROC comparison results with standard deviations over three random seeds across six test datasets for ResAD, NAGL, and VODA in Table 16. (i) For image-level AUROC, NAGL exhibits relatively large variability, while both ResAD and our VODA show significantly lower deviations, with ResAD achieving the smallest variance. (ii) For pixel-level AUROC, both ResAD and NAGL demonstrate large deviations on certain datasets, whereas our VODA maintains consistently low variability across all datasets. These results demonstrate the superior stability and robustness of our method with respect to different random seeds.

Table 16: AUROC (%) with mean and standard deviation over 3 random seeds.

		Datasets					
		VisA	MVTecAD	BTAD	MVTec3D	MVTecLOCO	BraTS
Ima.	ResAD	86.5±**0.31**	90.7±**0.21**	95.4±**0.12**	70.5±**0.29**	-	-
	NAGL	91.2±1.1	97.1±0.6	94.4±1.2	86.9±1.8	71.5±3.1	84.9±1.7
	VODA	93.3±0.53	96.8±0.43	95.0±0.21	85.0±0.97	70.5±**0.70**	90.5±**0.13**
Pix.	ResAD	97.5±0.08	95.9±0.12	97.6±0.05	97.2±0.09	-	-
	NAGL	97.8±0.0	97.0±**0.0**	76.4±0.9	95.4±**0.0**	66.9±0.6	97.1±0.4
	VODA	98.5±**0.004**	97.5±0.07	96.8±**0.06**	98.7±0.03	76.9±**0.24**	97.9±**0.03**

E.7 Multi-layer feature concatenation

In the current work, VOD is applied solely to the feature representations extracted from the final layer of DINOv2. We further investigate whether applying multi-layer features yeilds the improvement of performance.As summarized in the Table 17, concatenating more intermediate layers gradually leads to performance degradation, suggesting that shallow-layer features may contain less discriminative information or introduce irrelevant variations that interfere with anomaly detection. These results indicate that effective multi-layer feature aggregation requires further investigation.

F Limitations

Our method is particularly suitable for reference-based anomaly detection scenarios where multiple normal references are available and the target is to identify whether a query deviates from normal

Table 17: Performance comparison of VODA with different multi-layer feature aggregation strategies.

	Last layer	Last two layers	Last three layers
VisA	**92.4/98.4**	91.8/97.6	88.7/97.2
BraTS	**88.5/97.6**	87.2/96.7	87.7/96.8

patterns despite object-category, defect-type, or domain discrepancies. Our primary application domain is industrial visual inspection, where the goal is to improve robustness across diverse products and domains rather than identify semantic defect categories. However, the proposed method has several limitations. First, its performance depends on the diversity and quality of the available normal references, which is an inherent limitation of reference-based anomaly detection. Second, the current framework focuses on detecting whether a sample is anomalous rather than explaining the underlying defect type, resulting in limited semantic interpretability. Third, extending VODA to more complex scenarios, such as logical anomaly detection that requires modeling global object structures and relationships, is beyond the scope of this work.