PADFormer: Pose-agnostic Anomaly Detection from Sparse View Images

Ruiqi Wang[1,2⋆], Yiming Qian[1], Fenggen Yu[1], Yuxuan Lu[3], Dakuo Wang[3], Hao Zhang[2], and Jing Huang[1]

[1] Amazon
[2] Simon Fraser University
[3] Northeastern University

Abstract. Pose-agnostic Anomaly Detection (PAD) remains challenging as anomalies can appear under arbitrary viewpoints, requiring methods to handle significant pose variations. Existing approaches rely on complex 3D reconstruction, which are computationally expensive and require extensive multi-view data. We propose PADFormer, a novel image-space approach that leverages Vision Transformer (ViT) to directly reconstruct anomaly-free versions of query images while preserving pose information. Our key insight is to adapt cross-view masked reconstruction for anomaly detection through training exclusively on normal data, combined with dynamic patch selection and spatial alignment mechanisms that enable effective learning from sparse reference views under significant pose variations. During inference, we perform multiple forward passes with different masking patterns to generate an ensemble of anomaly-free reconstructions, ensuring comprehensive coverage of the query image. Anomalies are detected by comparing these reconstructions with the query image. PADFormer achieves state-of-the-art results on the PAD benchmark while maintaining comparable performance on classic few-shot anomaly detection (FSAD) tasks, demonstrating superior efficiency and generalization without requiring 3D reconstruction.

1 Introduction

Unsupervised visual anomaly detection has emerged as a critical task in industrial quality control and manufacturing [3, 4, 28, 34, 45, 58], aiming to identify shape or texture anomalies using only normal samples and representations from pre-trained models [6, 39]. With the increasing demand for automated inspection of manufactured products and the growing complexity of object shapes and textures, object-centric anomaly detection has attracted significant research interest. However, traditional anomaly detection methods [6, 13, 26] operate under a restrictive pose-aligned assumption, where objects are inspected from predefined, fixed viewpoints. This limitation severely constrains their practical applicability, as real-world anomalies may manifest from arbitrary viewing angles and potentially occluded perspectives.

⋆ Work carried out during internship at Amazon.

The *pose-agnostic anomaly detection* (PAD) setting [55] was recently introduced to address this limitation, offering enhanced flexibility by enabling anomaly detection when the query image is captured from an arbitrary, unknown viewpoint. A common practice in PAD is to first reconstruct the target object in 3D from dense anomaly-free input views with ground-truth (GT) poses, then perform cross-comparison between the query image and rendered views from the 3D model after estimating the query pose and aligning it with the 3D reconstruction [22, 27, 30, 49]. However, these 3D reconstruction-based methods face critical bottlenecks: they require GT camera poses for hundreds of reference images (often 200+) to build the 3D model, and they incur substantial computational costs due to the 3D reconstruction process. This requirement for dense multi-view data with precise pose annotations poses a significant challenge for real-world deployment, where obtaining comprehensive multi-view coverage with GT poses may be infeasible or prohibitively expensive. In this paper, we introduce PADFormer, a novel approach that extends pose-agnostic anomaly detection to the *sparse-view* input setting, as illustrated in Fig. 1. Our key insight is that by leveraging Vision Transformers (ViT) [8] with a masked reconstruction strategy [15], we can directly reconstruct an *anomaly-free* version of the query *image*, with or without anomalies, while bypassing explicit 3D model acquisition. Anomaly detection and localization can then be accomplished through pixel-wise comparison between the query image and its reconstructed anomaly-free counterpart. Specifically, we apply random masking to the query image, and PADFormer learns to reconstruct the masked patches in an anomaly-free manner by training exclusively on anomaly-free data, naturally preserving the exact pose of the input image. This pipeline significantly reduces the computational complexity associated with dense input view requirements while achieving state-of-the-art performance on PAD benchmarks with sparse views.

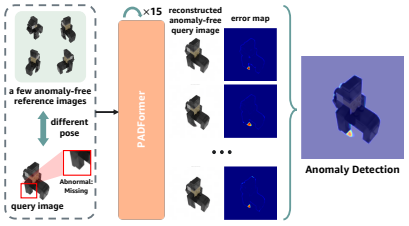

Fig. 1: Method overview – We propose PADFormer, a novel ViT-based method that directly reconstructs anomaly-free query images from sparse-view references for *pose-agnostic anomaly detection* (PAD). PADFormer performs multiple stochastic inference passes to generate diverse reconstructions and error maps, which are averaged to produce final anomaly detection results.

To enable effective learning from sparse views, PADFormer incorporates several technical innovations. First, we introduce a dynamic patch selection mechanism with Spatial Transformation Networks (STN) [20] that efficiently identifies and aligns the most relevant reference patches for each query position, reducing computational overhead while maintaining reconstruction quality. Second, our cross-attention decoder [44] leverages learnable anomaly-agnostic tokens that encode both global and local anomaly-free priors, enabling the model to suppress anomalous patterns during reconstruction. Third, we employ a two-stage train-

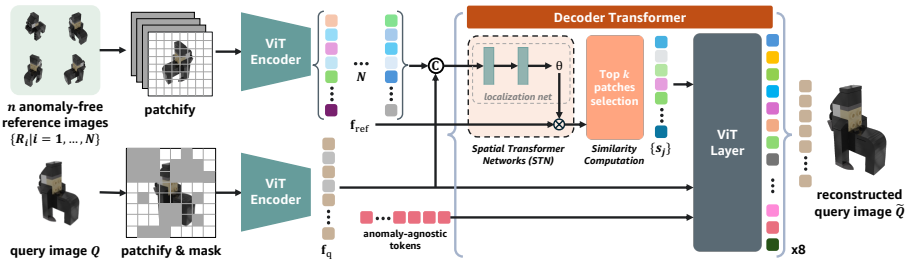

Fig. 2: PADFormer model architecture – PADFormer first patchifies the reference images $\{R_i|i = 1, \ldots, N\}$ into tokens. The query image to be reconstructed Q is randomly masked and also tokenized. Both are processed through two weight-sharing ViT encoders to obtain query feature token $\mathbf{f}_{\text{q}}$ and concatenated reference features token $\mathbf{f}_{\text{ref}}$. A Spatial Transformation Network (STN) takes their concatenation to compute transformation $\boldsymbol{\theta}$, which is applied *only to* $\mathbf{f}_{\text{ref}}$ via bilinear sampling for alignment, producing $\tilde{\mathbf{f}}_{\text{ref}}$. A similarity computation module (SIM) then selects top-k most similar aligned reference patches $\{\mathbf{s}_j\}$ for each query position. In the decoder, query features token $\mathbf{f}_{\text{q}}$ attend to anomaly-agnostic tokens and selected reference patches $\{\mathbf{s}_j\}$ through cross-attention across 8 ViT layers. A linear projection maps the output to pixel values, which are unpatchified to produce the reconstructed query image $\tilde{Q}$.

ing strategy to progressively build our model's capability: an initial *within-view* stage learns geometric invariance of object features through self-augmentation, followed by a *cross-view* training that enables generalization across object categories. During inference, we perform multiple reconstruction passes with different masking patterns to ensure comprehensive coverage, computing anomaly scores in CIELAB color space [42] for robust detection.

Our main contributions are summarized as follows:

- We introduce and formalize the sparse-view PAD setting, addressing a critical gap between research assumptions and real-world deployment constraints where dense multi-view data is impractical or unavailable.
- We propose PADFormer, a novel ViT-based method that directly reconstructs anomaly-free query images in 2D image space, eliminating the need for complex 3D reconstruction while maintaining pose consistency through learned masked reconstruction.
- Our overall design achieves state-of-the-art performance on PAD benchmarks under sparse view conditions with superior computational efficiency compared to 3D-based methods, while demonstrating strong generalization to classical *few-shot anomaly detection* (FSAD) scenarios. Note that all existing FSAD benchmarks assume geometric alignment between training and test images, operating under the constraint of similar or identical viewpoints.

2 Related Works

2.1 Anomaly Detection

Reference-free Image Anomaly Detection. Classical reference-free image anomaly detection framework trains a customized model for each class to detect anomalies without requiring normal reference images during testing. Self-supervised learning methods [10, 38, 43] learn normal patterns through pretext tasks, while hybrid approaches [6, 39] integrate feature matching with distribution modeling. Transformer-based methods [24, 32, 35, 50, 51] leverage self-attention mechanisms to capture long-range dependencies for anomaly localization. However, these methods assume fixed object poses with similar viewpoints. In contrast, our work addresses pose-agnostic anomaly detection where objects exhibit significant pose variations.

Pose-agnostic Anomaly Detection (PAD). Current approaches to PAD predominantly rely on explicit 3D geometric reconstruction pipelines. Zhou et al. [55] introduced this problem alongside the Multi-pose Anomaly Detection (MAD) dataset, which contains 20 object categories with over 200 views each. Their OmniposeAD method leverages Neural Radiance Fields (NeRF) to encode anomaly-free objects from diverse viewpoints, performs coarse-to-fine pose estimation on query images, and detects anomalies by comparing reconstructed normal references with queries. To improve reconstruction efficiency and quality, SplatPose [27] and SplatPose++ [30] adopt 3D Gaussian Splatting for faster training and inference. More recently, PIAD [49] extends the framework to jointly handle pose and illumination variations. Despite their effectiveness, these 3D reconstruction-based approaches require extensive multi-view training data with pose annotations, suffer from reconstruction artifacts, and need test-time pose optimization for each query. In contrast, our method enables efficient PAD by directly reconstructing anomaly-free query images from few reference views, eliminating 3D reconstruction and test-time optimization.

Few-shot Anomaly Detection (FSAD). Current FSAD methods can be categorized into three groups: (1) Memory-Bank and Embedding Methods, such as PatchCore [39] and PaDiM [6], which leverage coreset subsampling to define representative patch-level features from few normal samples; (2) Vision-Language Models (VLMs), which exploit pretrained models through prompt learning [5, 11, 21, 29, 31, 33, 36, 40, 48, 56] or in-context residual learning [57] for robust cross-domain generalization; and (3) Reconstruction and Registration Methods [2, 9, 17, 18], which employ 2D spatial transformers or diffusion models to align query features with support set features. UniVAD [12] represents the current state-of-the-art training-free FSAD method, leveraging contextual component clustering and multi-level matching to achieve superior cross-domain generalization. While these methods excel at data efficiency, they fundamentally lack the capability to handle unconstrained 3D pose variations. Their 2D alignment mechanisms cannot synthesize the novel appearances required by arbitrary

viewpoint changes, limiting their applicability to real-world scenarios where test-time viewpoints differ from reference views.

Our method bridges PAD and FSAD by addressing their complementary limitations: we achieve pose-agnostic anomaly detection without requiring extensive multi-view data, pose annotations, or 3D reconstruction pipelines. By learning to reconstruct anomaly-free query images directly from few-shot references, we simultaneously achieve data efficiency and pose robustness.

2.2 Masked Image Modeling

Masked Image Modeling (MIM) [7,15] has emerged as a powerful self-supervised learning paradigm that trains models to predict masked image regions from visible context. Pioneering works include BEiT [1], which reconstructs discrete visual tokens, and MAE [15], which introduced a high masking ratio (75%) with an asymmetric encoder-decoder design. Subsequent methods [14, 47, 54] advanced the field through simplified architectures, self-distillation, and siamese designs. MAEDAY [41] pioneered MIM for anomaly detection, leveraging the insight that anomalous regions are harder to reconstruct. Recent works [19, 37] have explored MIM for 3D-aware tasks. However, MAEDAY and similar reconstruction-based methods [52, 53] focus on viewpoint-aligned images and cannot handle unconstrained pose variations in PAD scenarios.

Extending MIM to multi-view scenarios, CroCo [46] introduces cross-view completion where a partially masked input image is reconstructed using both its visible content and a second unmasked reference image from a different viewpoint. Unlike single-view MIM, cross-view completion resolves reconstruction ambiguity by reasoning about scene geometry across views. However, CroCo trains on arbitrary scenes for general 3D vision tasks. Our work differs by adapting cross-view MIM for anomaly detection through training exclusively on anomaly-free data. This enables synthesizing anomaly-free versions of query images under novel viewpoints, where anomalies are replaced with plausible normal content rather than faithfully reconstructed.

3 Method

We begin by formulating the problem statement in Sec. 3.1, followed by a detailed description of our transformer-based model architecture in Sec. 3.2. Finally, we elaborate on the training and inference procedures in Sec. 3.3.

3.1 Problem Statement

Given a test query image Q_{test} (normal or abnormal) and a set of N anomaly-free reference images $\{R_i\}_{i=1}^{N}$ from object o_j, the goal is to detect and localize anomalies in Q_{test}. Both Q_{test} and $\{R_i\}_{i=1}^{N}$ are captured under unknown and different camera poses. In addition, the reference set $\{R_i\}_{i=1}^{N}$ is small and sparse, providing limited viewpoint coverage of the object o_j.

Previous Few-Shot Anomaly Detection (FSAD) methods assume the query and reference images share aligned viewpoints, limiting their ability to detect defects from unseen angles. Conversely, early *pose-agnostic anomaly detection* (PAD) approaches require dense reference sets and per-object training, which reduces their practical applicability. This work addresses both limitations by tackling anomaly detection in a pose-agnostic manner with only a sparse set of reference images, eliminating the need for viewpoint alignment or extensive training data.

3.2 Model Architecture

As shown in Fig. 2, PADFormer uses an end-to-end transformer model to learn the anomaly-free query image reconstruction. In this subsection, we first describe the key components of our PADFormer, then provide details of the training and inference for localization of the anomalies.

Encoder. PADFormer starts by tokenizing the input images. Given a query image Q and N reference images $\{R_i|i = 1, \ldots, N\}$, we first patchify them into non-overlapping patches of size $p \times p$ and linearly project each patch to obtain tokens of dimension d. Fixed sinusoidal positional embeddings are added to all tokens. The query and reference images are then processed through the same Vision Transformer (ViT) encoder with shared weights. The key difference lies in the masking strategy: similar to Siamese MAE [14], we adopt an asymmetric masking approach where we randomly mask a ratio (40%) of patches in the query image, keeping only the visible patches as input to the encoder, while all reference images are encoded without any masking. The encoder consists of L standard transformer blocks with multi-head self-attention and feed-forward layers. After encoding, we obtain the masked query features $\mathbf{f}_q \in \mathbb{R}^{M_v \times d}$ where M_v is the number of visible patches, and complete reference features $\{\mathbf{f}_{r_i} \in \mathbb{R}^{M \times d} | i = 1, ..., N\}$ where M is the total number of patches per image. These encoded features serve as the foundation for subsequent patch-level matching and reconstruction. This asymmetric design forces the model to aggregate information from the intact reference images to reconstruct the masked query regions, allowing it to learn anomaly-free reconstruction patterns.

Dynamic Patch Selection. To enable efficient and accurate cross-attention between query and reference images, we employ a dynamic patch selection mechanism that identifies the most relevant reference patches for each query position. This approach serves two purposes: (1) improving reconstruction quality by focusing on the most informative reference patches, and (2) reducing computational cost by limiting the number of patches used in cross-attention. Let $\mathbf{f}_{r_i} \in \mathbb{R}^{M \times d}$ denote the encoded features of the i-th reference image, where M is the number of patches per image. Before measuring patch similarity, we align each reference feature grid to the query feature grid with a Spatial Transformation Network (STN) [20]. The STN is used only to transform reference features; the query features remain fixed and define the target coordinate frame.

The STN operates on patch-level features through the following pipeline. Given visible query features $\mathbf{f}_{\mathrm{q}} \in \mathbb{R}^{M_v \times d}$ after masking, we first fill masked positions with learnable embeddings and restore the original patch order, producing a complete query grid $\hat{\mathbf{f}}_{\mathrm{q}} \in \mathbb{R}^{M \times d}$. For each reference image R_i, $\hat{\mathbf{f}}_{\mathrm{q}}$ and $\mathbf{f}_{r_i}$ are reshaped from token sequences into 2D feature maps of shape $[B, d, h, w]$, where $h \times w = M$. Their channel-wise concatenation is passed to a lightweight localization network with two convolutional layers (kernel sizes 7×7 and 5×5, with 32 and 10 output channels), each followed by 2×2 max pooling, and fully connected layers that predict a 6-dimensional affine transformation matrix $\boldsymbol{\theta}_i$. This affine transform is applied only to the reference feature map: we generate a sampling grid from $\boldsymbol{\theta}_i$ and use bilinear interpolation in feature space to obtain the aligned reference features $\tilde{\mathbf{f}}_{r_i}$. After applying this process to all N references, the aligned features are concatenated into $\tilde{\mathbf{f}}_{\mathrm{ref}} = \{\tilde{\mathbf{f}}_{r_i}\}_{i=1}^{N} \in \mathbb{R}^{(N \cdot M) \times d}$. This patch-level geometric alignment reduces pose-induced mismatch before nearest-patch retrieval.

We then measure the relevance between query and aligned reference patches using a hybrid similarity score that combines feature and spatial cues. The feature similarity between query patch j and aligned reference patch l is computed via normalized cosine similarity:

$$S_{\mathrm{feat}}^{jl} = \langle \frac{\hat{\mathbf{f}}_{\mathrm{q}}^{j}}{\|\hat{\mathbf{f}}_{\mathrm{q}}^{j}\|}, \frac{\tilde{\mathbf{f}}_{\mathrm{ref}}^{l}}{\|\tilde{\mathbf{f}}_{\mathrm{ref}}^{l}\|} \rangle \tag{1}$$

When the query and reference images have matching resolutions, we additionally incorporate spatial proximity:

$$S_{\mathrm{spatial}}^{jl} = \exp\left(-\frac{\|\mathbf{p}_{\mathrm{q}}^{j} - \mathbf{p}_{\mathrm{ref}}^{l}\|_2}{\sigma} \right) \tag{2}$$

where $\mathbf{p}_{\mathrm{q}}^{j}$ and $\mathbf{p}_{\mathrm{ref}}^{l}$ are the 2D grid positions of patches in their original spatial arrangements, and $\sigma = 2.0$ controls spatial locality. The final similarity is:

$$S^{jl} = (1 - w) S_{\mathrm{feat}}^{jl} + w S_{\mathrm{spatial}}^{jl} \tag{3}$$

where $w = 0.3$ balances feature matching and spatial locality. Based on these scores, we select the top-k most similar aligned reference patches for each query position, denoted as $\mathbf{s}_j \in \mathbb{R}^{k \times d}$ (with $k = 10$ by default). Thus, each query token attends only to a compact, pose-aligned set of reference candidates rather than all $N \cdot M$ reference patches.

Cross-Attention Decoder. The decoder reconstructs the masked query patches through 8 ViT layers, where each layer combines query context, learned anomaly-free priors, and retrieved reference evidence. The encoded query features $\mathbf{f}_{\mathrm{q}}$ and aligned reference features $\tilde{\mathbf{f}}_{\mathrm{ref}}$ are first projected to the decoder dimension d' via separate linear layers. The masked query positions are filled with learnable mask embeddings, and the full query sequence is reordered to the original patch

positions using the unmasking indices from the encoder. Decoder positional embeddings are then added to all query tokens.

To provide reconstruction priors that do not depend on a specific anomaly type, we introduce learnable anomaly-agnostic tokens as a trainable memory bank. This memory contains (1) 32 global tokens $\mathbf{T}_{\text{global}} \in \mathbb{R}^{32\times d'}$ shared across all positions to capture common normal appearance patterns, and (2) position-indexed local tokens $\mathbf{T}_{\text{local}} \in \mathbb{R}^{M\times 1\times d'}$, where each patch position has one associated token for location-aware normal priors. We flatten the local tokens into $\mathbf{T}^{\text{flat}}_{\text{local}} \in \mathbb{R}^{M\times d'}$ and expose all local tokens to every query position during cross-attention. This all-to-all access is important under large pose changes: the most useful normal prior for a query patch may come from a different nominal grid location, while the positional embeddings still preserve each token's location identity.

Each decoder layer proceeds in two steps. First, self-attention among the query tokens propagates visible-query context and updates the masked-token representations. Second, for each query position j, the decoder performs cross-attention over a position-specific key/value set formed by concatenating three sources: the global anomaly-agnostic tokens $\mathbf{T}_{\text{global}}$, the flattened local tokens $\mathbf{T}^{\text{flat}}_{\text{local}}$, and the selected aligned reference patches $\mathbf{s}_j$. In this interaction, the query token decides how much to use global normal priors, location-aware priors, and concrete visual evidence from the reference images. After 8 layers and final layer normalization, a linear projection maps the decoded features to pixel values, which are then unpatchified to produce the reconstructed query image $\tilde{Q}$.

3.3 Training & Inference

Training. PADFormer aims to learn to reconstruct anomaly-free query images based on few-shot reference images. We train our model in two stages to facilitate this objective. During training, both reference images $\{R_i|i = 1,\ldots,N\}$ and query image Q are anomaly-free.

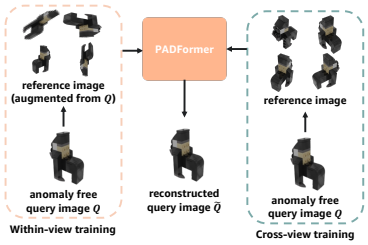

Fig. 3: Within-view pre-training. An augmented version of the query image is used as reference.

– **Within-view training:** For each query image Q, we generate reference images by applying random rotation and deformation transformations to the query image Q itself. The model is then trained to reconstruct the original query image Q from these self-augmented references, as shown in Fig. 3. This strategy enables the model to learn geometric invariance and robust feature representations while focusing on intrinsic object properties, laying a solid foundation for cross-object generalization in the subsequent stage.

– **Cross-view training:** In this stage, we train a unified PADFormer model across all object categories O in the dataset simultaneously. Unlike the

within-view stage where references are derived from the query itself, here the model learns to reconstruct query images using reference images from the same object category but different views. This cross-view training enables the model to generalize beyond instance-specific features and learn category-level representations that are robust to intra-class variations.

For both training stages, following prior works [16,23], we train PADFormer with MSE and perceptual loss [25]:

$$L = \text{MSE}(Q, \tilde{Q}) + \lambda \cdot \text{Perceptual}(Q, \tilde{Q}) \tag{4}$$

where $\tilde{Q}$ denotes the reconstructed query image and λ is the weight for balancing the perceptual loss.

Inference. During inference, given a test query image Q_{test}, we randomly sample $N = 4$ anomaly-free reference images from the training set. We perform $I = 15$ reconstruction passes with different random masking patterns. With mask ratio $r = 0.4$, the probability that any patch is masked at least once is $1 - (1 - r)^I \approx 99.95\%$, providing near-complete spatial coverage. For each reconstruction pass i, we first upsample the reconstructed image $\hat{Q}_i$ to match the original resolution for sharper anomaly localization. We then compute pixel-wise reconstruction errors using perceptual color differences in CIELAB space [42]. Specifically, we apply Gaussian filtering ($\sigma = 1.4$, kernel size 7) to each LAB channel of the query image to reduce high-frequency noise, then compute squared differences with the unfiltered reconstruction:

$$E_i(x, y) = \sum_{c \in \{L,a,b\}} \left(\mathcal{G}_\sigma \left(c(Q_{\text{test}}) \right) - c(\hat{Q}_i) \right)^2 \tag{5}$$

where $\mathcal{G}_\sigma$ denotes Gaussian filtering and c indexes the LAB channels. The I error maps are averaged: $\bar{E}(x, y) = \frac{1}{I} \sum_{i=1}^{I} E_i(x, y)$. To distinguish genuine anomalies from consistent reconstruction artifacts, we compute the variance $\text{Var}(x, y) = \frac{1}{I} \sum_{i=1}^{I} (E_i - \bar{E})^2$ across reconstructions and retain only high-variance regions via a consistency mask $\mathcal{M}(x, y) = \mathbb{1}[\text{Var}(x, y) > P_{50}(\text{Var})]$, where P_{50} is the 50th percentile. This variance-based filtering also provides a natural defense against false positives arising from rare-but-normal patterns such as batch variations or illumination drift. Since such patterns are part of the normal data distribution, the model reconstructs them consistently across different masking passes, resulting in low variance. They are therefore suppressed by the consistency mask, unlike genuine anomalies which the model fails to reconstruct consistently. The final anomaly score map is:

$$\mathcal{E}(x, y) = \bar{E}(x, y) \cdot \mathcal{M}(x, y) \tag{6}$$

The image-level anomaly score is $S = \max(\mathcal{E})$, providing a single scalar score for classification.

Table 1: Anomaly detection; PAD task – Quantitative comparison of image-level AUROC, pixel-level AUROC, and AUPRO on MAD-SIM and PIAD datasets. Best results are in **bold** and second-best are underlined.

Setup	Method	MAD-SIM			PIAD (synt+real)		
		image-AUROC	pixel-AUROC	AUPRO	image-AUROC	pixel-AUROC	AUPRO
2-shot	OmniAD [55]	48.2	80.1	68.2	47.1	77.5	66.7
	SplatPose [27]	48.3	81.4	69.1	48.1	77.2	66.9
	PIAD [49]	49.0	82.2	70.4	48.1	78.5	69.2
	PromptAD [29]	56.4	78.9	67.3	53.2	70.8	60.2
	UniVAD [12]	57.9	80.4	68.8	54.3	71.9	60.7
	PADFormer	**78.8**	**89.2**	**76.5**	**75.6**	**86.3**	**74.1**
4-shot	OmniAD [55]	50.6	84.6	72.9	47.5	78.5	68.4
	SplatPose [27]	52.2	86.7	74.6	48.8	78.9	68.8
	PIAD [49]	52.4	81.1	68.4	54.7	84.4	72.3
	PromptAD [29]	60.3	78.9	67.5	58.8	77.5	66.9
	UniVAD [12]	58.4	80.0	67.9	59.2	79.4	69.8
	PADFormer	**81.3**	**92.9**	**86.7**	**80.7**	**91.2**	**79.4**
10-shot	OmniAD [55]	51.0	86.8	74.9	64.2	83.5	71.1
	SplatPose [27]	56.4	90.2	83.8	67.4	87.3	76.5
	PIAD [49]	58.2	91.8	85.5	78.3	90.5	78.6
	PADFormer	**85.6**	**94.4**	**88.1**	**84.3**	**93.4**	**82.5**
All-shot	OmniAD [55]	91.9	98.4	92.3	79.1	95.9	88.3
	SplatPose [27]	94.9	99.0	94.1	82.6	97.7	93.1
	PIAD [49]	97.4	**99.5**	**95.4**	94.7	**99.0**	**95.2**
	PADFormer	**97.8**	98.9	93.9	**95.4**	98.3	93.7

4 Experiments

In this section, we describe our experimental setup and datasets (Sec. 4.1), introduce our model training details (Sec. 4.2), report evaluation results (Sec. 4.3) and perform an ablation study (Sec. 4.4).

4.1 Datasets

PAD Dataset. We use the MAD-SIM[4] dataset [55] and PIAD dataset [49] to train PADFormer separately. For each anomaly-free image in the training set, we randomly select $N \in \{2, 4, 10\}$ other anomaly-free images from the same training set to serve as reference images. During testing, we evaluate on the test sets of both datasets. For each test image, we randomly select $N \in \{2, 4, 10, \text{all}\}$ anomaly-free images from the training set as references.

FSAD Datasets. We use the MVTec-AD dataset [3] and ViSA dataset [59] to train PADFormer separately. For each anomaly-free image in the training set, we randomly select $N \in \{1, 2, 4\}$ other anomaly-free images from the same training set to serve as reference images. During testing, we evaluate on the test sets of both datasets. For each test image, we randomly select $N \in \{1, 2, 4\}$ anomaly-free images from the training set as references. We follow the standard evaluation

[4] We did not use MAD-Real since OmniAD [55] strongly recommends using MAD-Sim exclusively to explore the PAD problem in their GitHub repository.

Table 2: Anomaly detection; FSAD task – Quantitative comparison of image and pixel-level AUROC on MVTec-AD and VisA dataset.

Setup	Method	MVTecAD		VisA	
		image-level (AUROC)	pixel-level (AUROC)	image-level (AUROC)	pixel-level (AUROC)
1-shot	PatchCore [39]	63.7	83.9	58.9	76.7
	WinCLIP [21]	92.8	92.4	83.1	94.6
	PromptAD [29]	86.3	91.8	80.8	96.3
	UniVAD [12]	**94.8**	**95.7**	**87.1**	**97.2**
	PADFormer	92.8	92.0	83.6	95.0
2-shot	PatchCore [39]	72.4	89.6	60.2	82.4
	WinCLIP [21]	92.7	92.4	83.7	95.1
	PromptAD [29]	89.2	92.2	84.3	96.9
	UniVAD [12]	**96.4**	**96.0**	**89.3**	**97.5**
	PADFormer	92.8	92.0	84.6	95.2
4-shot	PatchCore [39]	74.9	92.6	62.6	85.4
	WinCLIP [21]	94.0	92.9	84.1	95.2
	PromptAD [29]	90.6	92.4	85.7	97.2
	UniVAD [12]	**98.6**	**96.1**	**91.7**	**97.8**
	PADFormer	94.7	93.2	86.6	95.4

protocol established by prior FSAD works [12, 21, 39], which report results exclusively on $\{1, 2, 4\}$-shot settings, as these low-shot scenarios are the primary focus of few-shot anomaly detection research.

4.2 Training Details

We train separate models for each dataset, with all categories trained together in a unified model per dataset. We use a pretrained MAE [15] encoder as our backbone. All input images are resized to 224×224. The training process consists of two stages: we first apply within-view training on the model for 50 epochs, followed by a second stage of cross-view training for 250 epochs. For the within-view stage, we apply random rotations and elastic deformations to generate reference images from each query. We employ a cosine learning rate schedule with a peak learning rate of 4e-4 and a warmup period of 2500 steps. All experiments are conducted on 4 NVIDIA A10 GPUs. We train separate models for each value of $N \in \{2, 4, 10, \text{all}\}$ for PAD datasets and $N \in \{1, 2, 4\}$ for FSAD datasets to evaluate the impact of the number of reference images on performance.

4.3 Comparison to Baselines

Following previous works [3, 55], we evaluate the performance of PADFormer using two standard metrics for anomaly detection: Area Under the Receiver Operating Characteristic curve (AUROC) at image-level and pixel-level. Pixel-AUROC measures the model's ability to localize anomalous regions at the pixel level, while image-AUROC evaluates the accuracy of binary classification for determining whether an entire image contains an anomaly. Both metrics range from 0 to 100, with higher values indicating better performance.

Table 3: Ablation – on model architecture, including Spatial Transformation Network (STN), dynamic patch selection (DPS) and anomaly-agnostic token (AAT).

Modules in PADFormer			MAD (4-shot)	
STN	DPS	AAT	image-AUROC	pixel-AUROC
-	-	-	76.8	80.2
-	✓	-	78.2	85.4
✓	✓	-	79.1	89.2
✓	✓	✓	**81.3**	**92.9**

Table 4: Ablation – on anomaly detection error metric.

Error Metric	MAD (4-shot)		Time (s) ↓
	image-AUROC	pixel-AUROC	
RGB	78.8	86.4	**0.6**
LAB	<u>81.3</u>	**92.9**	<u>0.8</u>
ViT feature	**83.5**	<u>91.3</u>	2.1

PAD Results. We compare PADFormer with OmniAD [55], SplatPose [27], and PIAD [49] on PAD task, as shown in Tab. 1. Our method demonstrates significant advantages in few-shot setting, outperforming the second-best method by over 20% on image-level AUROC and 8% on pixel-level AUROC in the 2-shot setup. Similar trends are observed in 4-shot and 10-shot setup, validating the effectiveness of PADFormer in PAD task with sparse reference images. In the all-shot setting, PADFormer maintains the best image-level AUROC on both benchmarks, while achieving competitive pixel-level results. This marginal pixel-level performance gap is attributed to inherent reconstruction errors. Notably, all compared methods perform per-object training in this setting, wheras PADFormer is trained jointly on all object categories, demonstrating its generalization ability. Beyond detection performance, PADFormer also offers superior computational efficiency for inference, as detailed in the supplementary materials. Qualitative results on both benchmarks under 4-shot setup are shown in Fig. 4. PADFormer successfully detects both fine-grained defects (e.g., stains and burrs) and relative large defects (e.g., missing parts) on MAD-SIM and PIAD, demonstrating robust performance across diverse viewpoints and anomaly types.

FSAD Results. We further compare PADFormer with PatchCore [39], WinCLIP [21], PromptAD [29] and UniVAD [12]. As shown in Tab. 2, PADFormer achieves competitive performance on the general FSAD task. Importantly, PADFormer is designed for PAD, where query images capture arbitrary viewpoints of the object, while general FSAD assumes geometric alignment between query and reference views, which is a fundamentally easier setting that stronger FSAD baselines such as UniVAD are specifically optimized for, leveraging large foundational models and alignment-specific architectures. Despite this task mismatch, our method demonstrates strong generalization without relying on such large foundational models, making it more lightweight and self-contained. Crucially, no existing FSAD method achieves comparable performance on PAD benchmarks, confirming that PADFormer addresses a distinctly harder problem. Fig. 5 shows qualitative results on MVTec-AD [3] and VisA [59] under 4-shot setup. PADFormer effectively handles challenging cases including defects in densely crowded regions (e.g., toothbrush bristles) and subtle anomalies such as tiny stains on PCBs.

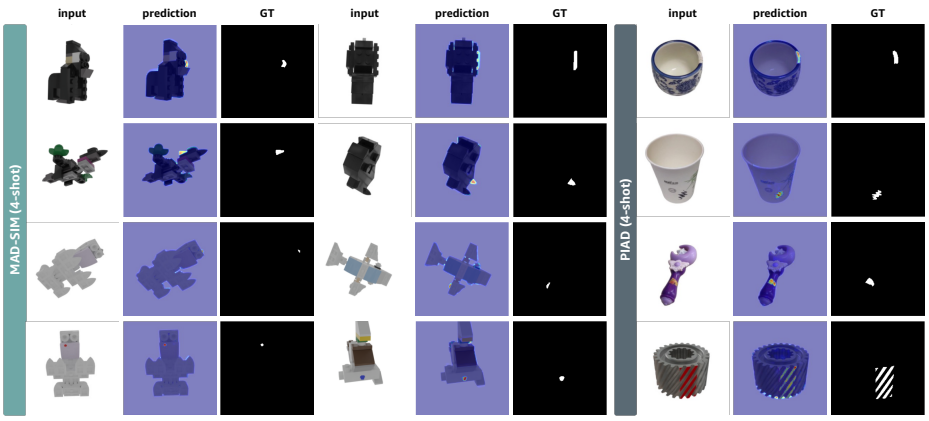

Fig. 4: Qualitative results on PAD benchmarks – On the left we visualize the results on MAD-SIM, with missing parts (row 1, 2), burrs (row 3) and stains (row 4). On the right we visualize results on PIAD. See supplementary materials for more qualitative comparisons.

4.4 Ablation Studies

Model Architecture. As shown in Tab. 3, we evaluate the effectiveness of our model designs. The baseline without any proposed modules uses a standard cross-attention transformer decoder, which is structurally analogous to CroCo [46] adapted for anomaly detection through training on anomaly-free data only. This baseline achieves 76.8% and 80.2% on image-level and pixel-level AUROC, respectively. Adding the Dynamic Patch Selection (DPS) module for top-k reference patches selection in cross-attention improves performance to 78.2% and 85.4%, demonstrating that selecting the most relevant patches is more effective than attending to all reference patches. Incorporating the Spatial Transformation Network (STN) to spatially align reference features before similarity computation further boosts the metrics to 79.1% and 89.2%, as it enables more accurate patch matching across different views. Finally, adding the Anomaly-agnostic Tokens (AAT)—32 global and M position-indexed local learnable tokens—achieves the best performance of 81.3% and 92.9%. Each component contributes meaningfully, with the complete PADFormer achieving 4.5% and 12.7% improvements over the CroCo-style baseline.

Error Evaluation Method. In Tab. 4, we ablate the anomaly scoring metric across three representation spaces: RGB space, LAB space and feature space extracted from intermediate transformer layers. While RGB space offers the fastest computation, it achieves significantly lower performance in both image-AUROC and pixel-AUROC due to its lack of perceptual uniformity. Feature space attains the highest image-level performance, but at approximately 3 times higher computational cost. LAB space achieves the best pixel-level localization while maintaining near-optimal efficiency, only marginally slower than RGB computation. Our LAB-based metric applies Gaussian filtering to the query image channels

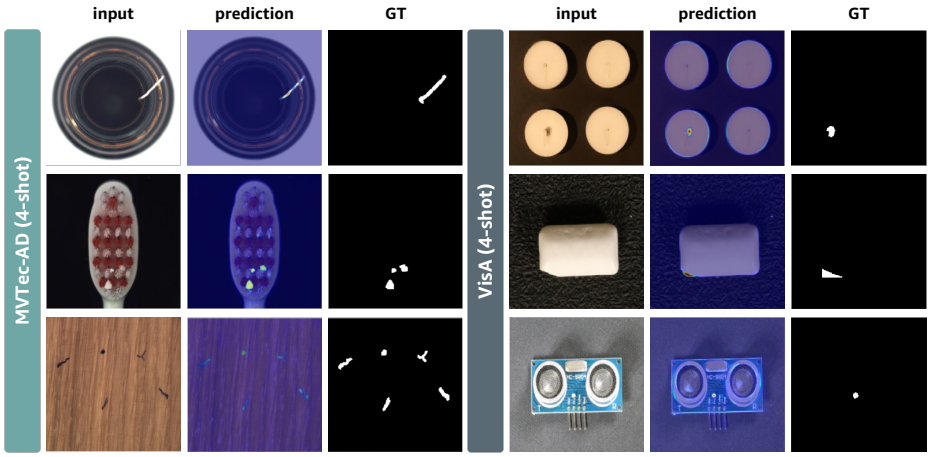

Fig. 5: Qualitative results on FSAD benchmarks: MVTec-AD (L) and VisA (R).

before computing differences, which reduces sensitivity to high-frequency noise and edge artifacts. The variance-based consistency filtering (retaining top 50% variance regions) contributes an additional 2.4% improvement in image-AUROC by suppressing systematic reconstruction errors that appear consistently across all masking patterns.

5 Conclusion

We present PADFormer, a novel Vision Transformer-based approach for *pose-agnostic anomaly detection* (PAD) under sparse-view settings. Our method addresses a critical limitation in existing approaches: the ability to perform few-shot PAD without requiring pose information from either reference or query images, while maintaining low computational cost. By directly reconstructing anomaly-free query images in 2D space through learned masked reconstruction, PADFormer eliminates the computationally expensive 3D reconstruction pipeline employed by prior methods, yet achieves superior pose consistency. Our method achieves superior performance compared to existing PAD approaches across few-shot settings, demonstrating exceptional sample efficiency in handling arbitrary viewpoints. PADFormer exhibits strong robustness and generalization across diverse object categories, while competitive performance on general FSAD benchmarks, where geometric alignment is assumed, further validates the versatility of our learned representations.

Despite these strengths, PADFormer has limitations. The fixed patch size may suboptimally handle large anomalies, and multiple inference passes (15 runs) are needed for comprehensive coverage. Future work could explore adaptive patch mechanisms and intelligent masking strategies to improve efficiency, as well as lightweight pose estimation to enhance performance under extreme viewpoint variations.

Acknowledgements

We sincerely thank all reviewers and Area Chair for carefully reading our paper, supplementary materials, and rebuttal. We appreciate their insightful suggestions and encouraging comments. We also thank Xi Liu and Weiwei Sun from Amazon for helpful discussions and valuable comments; Zhengqing Wang from Simon Fraser University for sharing helpful knowledge and experience with feed-forward transformers; Mingrui Zhao from Simon Fraser University for proofreading.

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT pre-training of image transformers. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=p-BhZSz59o4
2. Beizaee, F., Lodygensky, G.A., Desrosiers, C., Dolz, J.: Correcting deviations from normality: A reformulated diffusion model for multi-class unsupervised anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19088–19097 (2025)
3. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: 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)
4. Bergmann, P., Jin, X., Sattlegger, D., Steger, C.: The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. In: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SCITEPRESS-Science and Technology Publications (2022)
5. Cao, Y., Zhang, J., Frittoli, L., Cheng, Y., Shen, W., Boracchi, G.: Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In: European Conference on Computer Vision. pp. 55–72. Springer (2024)
6. Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: a patch distribution modeling framework for anomaly detection and localization. In: International conference on pattern recognition. pp. 475–489. Springer (2021)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
9. Fang, Z., Wang, X., Li, H., Liu, J., Hu, Q., Xiao, J.: Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17481–17490 (2023)
10. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. Advances in neural information processing systems **31** (2018)
11. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: Anomalygpt: Detecting industrial anomalies using large vision-language models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 1932–1940 (2024)

12. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: Univad: A training-free unified model for few-shot visual anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15194–15203 (2025)
13. Gudovskiy, D., Ishizaka, S., Kozuka, K.: Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 98–107 (2022)
14. Gupta, A., Wu, J., Deng, J., Fei-Fei, L.: Siamese masked autoencoders. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36 (2023)
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
16. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. In: International Conference on Learning Representations. vol. 2024, pp. 50678–50702 (2024)
17. Huang, C., Guan, H., Jiang, A., Zhang, Y., Spratling, M., Wang, X., Wang, Y.: Few-shot anomaly detection via category-agnostic registration learning. IEEE Transactions on Neural Networks and Learning Systems (2024)
18. Huang, C., Guan, H., Jiang, A., Zhang, Y., Spratling, M., Wang, Y.F.: Registration based few-shot anomaly detection. In: European conference on computer vision. pp. 303–319. Springer (2022)
19. Irshad, M.Z., Zakharov, S., Guizilini, V., Gaidon, A., Kira, Z., Ambrus, R.: Nerf-mae: Masked autoencoders for self-supervised 3d representation learning for neural radiance fields. In: European Conference on Computer Vision. pp. 434–453. Springer (2024)
20. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)
21. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19606–19616 (2023)
22. Jiang, B., Xie, Y., Li, J., Li, N., Chen, B., Xia, S.T.: Igspad: Inverting 3d gaussian splatting for pose-agnostic anomaly detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10229–10237 (2024)
23. Jin, H., Jiang, H., Tan, H., Zhang, K., Bi, S., Zhang, T., Luan, F., Snavely, N., Xu, Z.: Lvsm: A large view synthesis model with minimal 3d inductive bias. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=QQBPWtvtcn
24. Jin, J., Xu, Y., He, H., Gao, F., Zeng, W., Wang, W., Guo, B., Xu, Z.: A swin transformer-based hybrid reconstruction discriminative network for image anomaly detection. Scientific Reports **15**(1), 33929 (2025)
25. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
26. Kruse, M., Rosenhahn, B.: Multi-flow: Multi-view-enriched normalizing flows for industrial anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3933–3944 (2025)
27. Kruse, M., Rudolph, M., Woiwode, D., Rosenhahn, B.: Splatpose & detect: Pose-agnostic 3d anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3950–3960 (2024)

28. Li, X., Tan, X., Chen, Z., Zhang, Z., Zhang, R., Guo, R., Jiang, G., Chen, Y., Qu, Y., Ma, L., et al.: One-for-more: Continual diffusion model for anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4766–4775 (2025)
29. Li, Y., Goodge, A., Liu, F., Foo, C.S.: Promptad: Zero-shot anomaly detection using text prompts. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1093–1102 (2024)
30. Liu, Y., Hu, Y.S., Chen, Y., Zelek, J.: Splatpose+: Real-time image-based pose-agnostic 3d anomaly detection. In: European Conference on Computer Vision. pp. 378–391. Springer (2024)
31. Luo, W., Cao, Y., Yao, H., Zhang, X., Lou, J., Cheng, Y., Shen, W., Yu, W.: Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9974–9983 (2025)
32. Luo, W., Yao, H., Cao, Y., Chen, Q., Gao, A., Shen, W., Yu, W.: Inp-former++: Advancing universal anomaly detection via intrinsic normal prototypes and residual learning. arXiv preprint arXiv:2506.03660 (2025)
33. Ma, W., Zhang, X., Yao, Q., Tang, F., Wu, C., Li, Y., Yan, R., Jiang, Z., Zhou, S.K.: Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4744–4754 (2025)
34. Maack, R., Thun, L., Liang, T., Tercan, H., Meisen, T.: Pcad: A real-world dataset for 6d pose industrial anomaly detection. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 1132–1141 (2025)
35. Nafez, M., Koochakian, A., Maleki, A., Habibi, J., Rohban, M.H.: Patchguard: Adversarially robust anomaly detection and localization through vision transformers and pseudo anomalies. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20383–20394 (2025)
36. Qu, Z., Tao, X., Prasad, M., Shen, F., Zhang, Z., Gong, X., Ding, G.: Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In: European Conference on Computer Vision. pp. 301–317. Springer (2024)
37. Rajasegaran, J., Chen, X., Li, R., Feichtenhofer, C., Malik, J., Ginosar, S.: Gaussian masked autoencoders. arXiv preprint arXiv:2501.03229 (2025)
38. Reiss, T., Cohen, N., Bergman, L., Hoshen, Y.: Panda: Adapting pretrained features for anomaly detection and segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2806–2814 (2021)
39. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2022)
40. Sadikaj, Y., Zhou, H., Halilaj, L., Schmid, S., Staab, S., Plant, C.: Multiads: Defect-aware supervision for multi-type anomaly detection and segmentation in zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22978–22988 (2025)
41. Schwartz, E., Arbelle, A., Karlinsky, L., Harary, S., Scheidegger, F., Doveh, S., Giryes, R.: Maeday: Mae for few- and zero-shot anomaly detection. Computer Vision and Image Understanding **241**, 103959 (2024)
42. Sharma, G., Wu, W., Dalal, E.N.: The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of

Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur **30**(1), 21–30 (2005)
43. Sohn, K., Li, C.L., Yoon, J., Jin, M., Pfister, T.: Learning and evaluating representations for deep one-class classification. In: International Conference on Learning Representations (2021)
44. Vaswani, A., et al.: Attention is all you need. NeurIPS (2017)
45. Wang, C., Zhu, W., Gao, B.B., Gan, Z., Zhang, J., Gu, Z., Qian, S., Chen, M., Ma, L.: Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22883–22892 (2024)
46. Weinzaepfel, P., Leroy, V., Lucas, T., Brégier, R., Cabon, Y., Arora, V., Antsfeld, L., Chidlovskii, B., Csurka, G., Revaud, J.: Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
47. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
48. Xu, J., Lo, S.Y., Safaei, B., Patel, V.M., Dwivedi, I.: Towards zero-shot anomaly detection and reasoning with multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20370–20382 (2025)
49. Yang, K., Cao, J., Bai, Z., Su, Z., Tagliasacchi, A.: Piad: Pose and illumination agnostic anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4734–4743 (2025)
50. Yang, Q., Guo, R.: An unsupervised method for industrial image anomaly detection with vision transformer-based autoencoder. Sensors **24**(8), 2440 (2024)
51. You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., Le, X.: A unified model for multi-class anomaly detection. Advances in Neural Information Processing Systems **35**, 4571–4584 (2022)
52. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem–a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8330–8339 (2021)
53. Zavrtanik, V., Kristan, M., Skočaj, D.: Reconstruction by inpainting for visual anomaly detection. Pattern Recognition **112**, 107706 (2021). https://doi.org/10.1016/j.patcog.2020.107706, https://www.sciencedirect.com/science/article/pii/S0031320320305094
54. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image bert pre-training with online tokenizer. In: International Conference on Learning Representations (2022)
55. Zhou, Q., Li, W., Jiang, L., Wang, G., Zhou, G., Zhang, S., Zhao, H.: Pad: A dataset and benchmark for pose-agnostic anomaly detection. Advances in Neural Information Processing Systems **36**, 44558–44571 (2023)
56. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In: The Twelfth International Conference on Learning Representations (2023)
57. Zhu, J., Pang, G.: Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17826–17836 (2024)

58. Zhu, W., Wang, L., Zhou, Z., Wang, C., Pan, Y., Zhang, R., Chen, Z., Cheng, L., Gao, B.B., Zhang, J., et al.: Real-iad d3: A real-world 2d/pseudo-3d/3d dataset for industrial anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15214–15223 (2025)
59. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European conference on computer vision. pp. 392–408. Springer (2022)