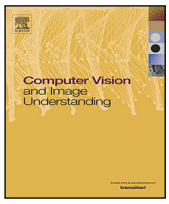
Computer Vision and Image Understanding 271 (2026) 104901

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Pyramidal anomaly detection with state-space models ☆

Nasar Iqbal *, Niki Martinel

University of Udine, Department of Mathematics, Computer Science and Physics, Via Delle Scienze 206, Udine, 33100, UD, Italy

ARTICLE INFO

Keywords:
State space models (SSMs)
Mamba
Pyramidal scanning approach
Small anomalies
Linear complexity

ABSTRACT

Recent advancements in CNNs and transformer-based methods have significantly improved multiclass anomaly detection; however, accurately localizing small anomalies in industrial images remains challenging. Although promising, CNNs suffer in capturing long-range dependencies and the effectiveness of the transformers is hindered by their quadratic computational costs. This paper proposes a novel State Space Model (SSMs)-based multi-class anomaly detection method, termed Pyramidal Anomaly Detection with State-Space Models (PAD-SSM), which offers state-of-the-art performance with linear complexity. The main contribution of our work is the Pyramidal Scanning Approach (PSA), which performs the scanning hierarchically. This enables more localized and scale-aware analysis. PSA operates in a pyramid-like style, recursively partitions the image into equal non-overlapping patches across multiple scales. The SSM is then applied independently to each patch. This approach enables detailed, multi-scale, and high-resolution analysis, which is crucial for detecting small anomalies in industrial images. We integrate the PSA block with a pre-trained encoder for multi-scale feature extraction, a feature adapter to transform the encoded features to a target domain, and a synthetic anomaly generator that introduces noise at the feature level, further enhancing detection capabilities. Rigorous evaluations on the MVTec-AD and VisA datasets demonstrate that our method achieves state-of-the-art mean AU-ROC of 98.7, AP of 99.6, and $F1_{max}$ of 98.1 on MVTec-AD benchmark while using the lowest number of FLOPS (e.g., 10.3% fewer FLOPs than the next best method).

1. Introduction

Anomaly detection (AD) in industrial images aims at identifying and accurately localizing anomalous regions within an image that deviate from the normal appearance (Huang et al., 2022). This problem is of paramount importance in modern manufacturing and inspection processes, where early and accurate detection of anomalies can improve safety, reduce material waste, and ensure compliance with industry standards.

AD models must operate under constrained hardware to ensure cost-effective industrial automation applicability. However, a vast majority of AD approaches follow the *single-class* framework, thus requiring a separate model for each object class (Liu et al., 2023; Zhang et al., 2023; Rudolph et al., 2022; Roth et al., 2022). *Multi-class* AD methods introduce a single model to detect anomalies on multiple object classes within a unified framework (Zhao, 2023; Iqbal and Martinel, 2025; Lu et al., 2023; He et al., 2024a). This approach conserves substantial memory and training time.

Transformer-based models (You et al., 2022; Lu et al., 2023) were introduced to follow such a multiclass setting, thus improving training and inference efficiency. However, despite their strong performance, these methods suffer from quadratic computational complexity, making them less practical to be adopted in real-world industrial settings. This highlights the need to develop a method that is not only multiclass but also lightweight (i.e., balancing accuracy with lower computational cost).

Recent developments in State Space Models (SSMs) (Gu and Dao, 2023) have shown their capability to model long-range dependencies while maintaining linear computational complexity. Several research works (Liu et al., 2024; Zhu et al., 2024; Shi et al., 2024; Ruan and Xiang, 2024; Wang et al., 2024; Huang et al., 2024) have applied Mamba (Gu and Dao, 2023) to computer vision tasks and achieved promising results. MambaAD (He et al., 2024b) was the first to utilize such a framework for AD, offering computational efficiency. However, it mainly relies on a single-scale SSM, which limits its ability to detect small and subtle anomalies that are often localized within specific regions — and may be missed by a single-scale approach.

In our earlier work (Iqbal and Martinel, 2025), we introduced Pyramid-based Mamba to address the limitations of single-scale SSM by proposing a multi-scale approach that improves detection of small and localized anomalies. However, this comes with increased computational

☆ This article is part of a Special issue entitled: 'ACVR 2024' published in Computer Vision and Image Understanding.

* Corresponding author.
E-mail addresses: iqbal.nasar@spes.uniud.it (N. Iqbal), niki.martinel@uniud.it (N. Martinel).

https://doi.org/10.1016/j.cviu.2026.104901
Received 17 May 2025; Received in revised form 28 July 2026; Accepted 5 August 2026
Available online 7 August 2026

cost due to multiple pyramid levels. To reduce this overhead, in this work, we streamline the (Iqbal and Martinel, 2025) by limiting the number of pyramid levels while maintaining comparable detection performance.

Iqbal and Martinel (2025) also incorporates pyramid-mamba with several convolutional blocks to enhance spatial feature extraction, but this further adds to the complexity. Our current work addresses this by replacing the heavy convolutional blocks with lightweight Depthwise Squeeze and Excitation Block (DSEB), achieving efficient performance with lower computational complexity. Thus, the present study is designed to model multi-class industrial anomalies with minimal complexity. The experiments on multiple industrial benchmarks show that the approach maintains competitive results without unnecessary computational overhead on the AD task.

The key contributions of this work are as follows:

- We introduce a novel lightweight PSA which employs a new scanning strategy that utilizes multi-scale image pyramids, and performs localized scanning with each pyramid, enhancing the localization of small, subtle, and large anomalies with significantly lower computational cost.
- We integrate a lightweight Depthwise Squeeze and Excitation Block (DSEB) to capture fine-grained local features, which complement the global modeling capacity of PSA, thus improving the overall anomaly detection accuracy.
- Comprehensive experiments on public benchmarks demonstrate that PAD-SSM achieves state-of-the-art results with a single efficient model capable of handling the multi-class anomaly detection problem.

2. Related work

2.1. Unsupervised anomaly detection

Image-based industrial AD has been the focus of extensive study. Due to the unpredictable nature of defects, there is an increasing interest in leveraging unsupervised learning for the task. Typical unsupervised AD methods can be classified into Reconstruction-based, Synthesis-based, and embedded-based methods.

Reconstruction-based methods such as auto-encoders (Gong et al., 2019) and generative adversarial models (Yan et al., 2021)focus on learning the features of normal samples by reconstructing anomaly-free images during training. At inference time, the anomalies are identified by comparative analysis. The difference between the reconstructed and the input images generates the anomaly maps. Other approaches (Haselmann et al., 2018; Ristea et al., 2022; Zavrtanik et al., 2021) randomly mask patches from input images and predict the masked regions. These methods struggle when defective and normal regions share common patterns (Zavrtanik et al., 2021). Our approach also belongs to this group, but differs by utilizing a pyramid-based scanning, which enables more robust feature reconstruction across multiple scales, addressing the limitations of traditional reconstruction-based methods.

Synthesizing-based methods incorporate synthetic anomalies within anomaly-free samples to help the model learn a decision boundary. This is normally done by masking some pixels or regions of anomaly-free images and training the model to reconstruct them, or by generating synthetic anomalies and training a classifier or discriminator to distinguish them from real, anomaly-free images. Currently, such methods generate pseudo-anomalies by adding Perlin noise computed from a set of predefined texture images (Zavrtanik et al., 2021) or copy&paste patches from one sample to random locations in other samples (Li et al., 2021). In Liu et al. (2023), synthetic noise is added in the feature space, then a binary discriminator is trained to distinguish the noisy features from normal features. Zhang et al. (2024) proposed a model that combines artificial anomaly generation with discriminative feature selection. It adopts a diffusion-based method to produce artificial anomalies with varying strengths, which closely mimics real-world anomalies. It utilizes a diffusion model to progressively reconstruct images. These methods often suffer from the discrepancy between artificial and real-world anomalies. We address these issues and enhance robustness by introducing noise at the feature level during training. Noise added to the latent space can be precisely controlled without distorting the raw image pixels, which allows the model to learn robust internal representations without compromising the quality of input.

Embedding-based methods generally employ ImageNet pre-trained networks (Deng et al., 2009) to obtain low-dimensional embeddings of anomaly-free samples. Anomalies are then detected in the embedded space by calculating deviations from this learned distribution. Patch-based feature embedding through multivariate Gaussian distributions (Defard et al., 2021), and memory banks with Mahalanobis distance (Roth et al., 2022) were utilized to measure the similarity between anomalous and normal features.

These methods face challenges due to the domain gap between industrial images and the ImageNet dataset — composed of natural images. Features obtained from ImageNet-pretrained models may not properly capture industrial images' characteristics. Our PAD-SSM addresses this issue by incorporating a feature adapter to transfer the pre-trained extracted features to the target domain, which effectively mitigates the domain bias.

Multi-class Anomaly Detection Traditional single-class unsupervised AD requires training a separate model for each class. These methods suffer from limitations such as excessive time-consuming training and high storage space. Current works focus on how to utilize a single model to conduct unsupervised multi-class anomaly detection. UniAD (You et al., 2022) was among the first works to introduce a single anomaly detection framework using a vision transformer. OmniAL (Zhao, 2023) utilizes a unified CNN-based model to handle multiclass AD without the need for separate models per class. Further works have utilized more diverse methods for this task; for example, DiAD (He et al., 2024a) adopts diffusion models, while HVQ-Trans (Lu et al., 2023) employs a probabilistic model.

Though these multiclass AD models significantly conserve training time and storage space as compared to single-class settings, nevertheless, they do not inherently guarantee lightweight deployment. In the context of model lightweighting and reduction of model complexity, recent studies (Liu et al., 2023) have started to focus on the significance of computational efficiency in single-class anomaly detection settings. However, in the case of multiclass settings, the vast majority of the existing literature (Zhao, 2023; Iqbal and Martinel, 2025; Lu et al., 2023; He et al., 2024a) focused on achieving high AD performance while neglecting the impacts of model complexity. We introduce a novel, efficient SSM-based approach that considers the scale of model complexity to achieve state-of-the-art multi-class AD performance.

2.2. State space models

State Space Models (SSMs) (Gu and Dao, 2023; Gu et al., 2021; Smith et al., 2022; Mehta et al., 2022; Fu et al., 2022) have demonstrated strong performance in long-sequence modeling by efficiently capturing long-range dependencies. The foundational State-Space Sequence (S4) architecture (Gu et al., 2021) proposed a novel diagonal parameterization structure that enables linear computational complexity for sequence modeling tasks. Leveraging this foundation, models like S5 (Smith et al., 2022), H3 (Gong et al., 2019), and Mamba (Gu and Dao, 2023) have further enhanced the capabilities of SSMs'. Mamba has inspired numerous computer vision applications (Ruan and Xiang, 2024; Wang et al., 2024) as it is based on a data-driven selection method within S4, and maintains linear computational complexity for long-sequence modeling.

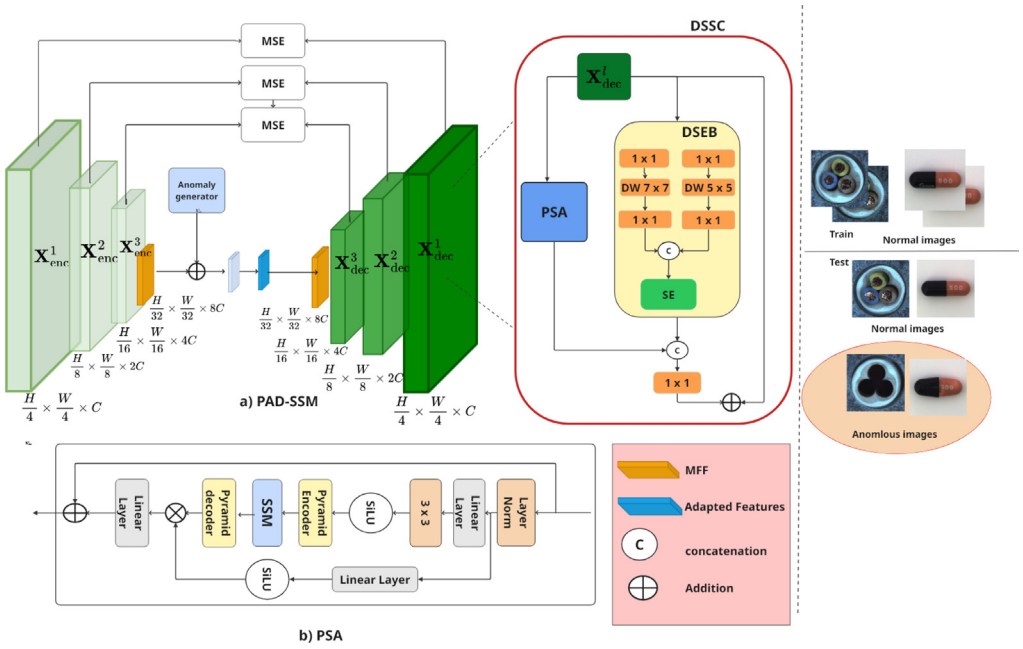


Fig. 1. Overview of the proposed PAD-SSM approach consisting of (a) teacher-student network to reconstruct multi-scale synthetic anomalous features by pre-trained encoders and Pyramid-Mamba-based decoders. (b) Each decoder consists of Pyramidal Scanning Approach (PSA) blocks capturing local and global interaction, and DSEB blocks to capture local information.

Mamba was originally designed for efficient 1D sequence modeling tasks. Extending it to the visual domains involves several key challenges due to the structural differences between ordered 1D sequences and non-causal 2D images. This challenge was first addressed in Liu et al. (2024) through a cross-scan module, while an omni-selective scanning method was introduced in Shi et al. (2024) to overcome the limitations of unidirectional SSMs. Recently, (Huang et al., 2024) exploited Mamba for AD by introducing a hybrid scanning approach that encodes feature maps through five scanning methods across eight directions to better model spatial dependencies. This method employs a single SSM model on the entire image, which limits its capability to localize small anomalies at different levels of detail. Our preliminary work in Iqbal and Martinel (2025) addresses this limitation by introducing a pyramid-based scanning mechanism, with the side effect of being more computationally demanding. We extend such preliminary work via depthwise convolutions and feature compression mechanisms that improve its efficiency while yielding state-of-the-art results.

3. Method

The proposed lightweight PAD-SSM architecture is shown in Fig. 1. The input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ is processed by the pre-trained encoder to extract features from different hierarchies, denoted as $\mathbf{X}^l_{\text{enc}} \in \mathbf{R}^{H^l\times W^l\times C^l}$ for each hierarchy level l. These multi-level features are aligned using the Multi-Scale Feature Fusion (MFF) block. This block effectively fuses semantic details from different depths of the network, capturing both high-level semantics and low-level details in a unified representation. To enable element-wise fusion, we first apply spatial and channel alignment to transform each $\mathbf{X}^l_{\text{enc}}$ into a common shape and then aggregate the transformed features via element-wise addition to get $\mathbf{X}_{\text{MFF}}$.

The fused features are processed by a feature adapter to transfer the features to the target domain, effectively mitigating domain bias. These target domain-adapted features are then perturbed with synthetic noise to get $\tilde{\mathbf{X}}_{\text{MFF}}$. Each decoder performs one level of pyramid-based scanning. Within each of these decoders, we introduce a Dynamic State Space Context (DSSC) module that integrates a Pyramidal Scanning Approach (PSA) block. This block performs multi-directional scanning across different levels of the pyramid. This ensures that both local and global information is effectively captured. The trainable decoder minimizes the reconstruction loss between the pre-trained encoded features $\mathbf{X}^l_{\text{enc}}$ and the reconstructed features from corresponding decoders $\mathbf{X}^l_{\text{dec}} \in \mathbb{R}^{H^l\times W^l\times C^l}$. We employ the same method in Deng and Li (2022) to generate a 2D anomaly map with multi-scale averaging.

3.1. State space models background

Leveraging principles from control systems, SSMs can be regarded as linear time-invariant systems that denotes a sequence models' class that maps a single-dimensional input stimulation $x(t) \in \mathbb{R}$ to response $y(t) \in \mathbb{R}$ through hidden space $h(t) \in \mathbb{R}^N$. SSMs can be formulated using linear ordinary differential equations as:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \tag{1}$$

$$y(t) = \mathbf{C}h(t), \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, and $\mathbf{C} \in \mathbb{R}^{1\times N}$ are state transition matrix.

The continuous-time SSMs must be discretized before being practically implemented using zero-order hold (ZOH), which is expressed as:

$$\bar{\mathbf{A}} = \exp(\Delta\mathbf{A}), \tag{3}$$

$$\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - I)\cdot\Delta\mathbf{B}. \tag{4}$$

After discretization, SSM-based models can be expressed as:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \tag{5}$$

$$y_t = \mathbf{C}h_t. \tag{6}$$

From a global convolution perspective, the output can be expressed as:

$$\mathbf{K} = (\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \ldots, \mathbf{C}\bar{\mathbf{A}}^{L-1}\bar{\mathbf{B}}), \tag{7}$$

$$y = \mathbf{x} * \mathbf{K}. \tag{8}$$

where $*$ represents convolution operation, $\mathbf{K} \in \mathbb{R}^L$ is an SSM kernel, and L is the sequence length of input $\mathbf{x} = [x(0), \ldots, x(L)]$.

3.2. Dynamic state space context (DSSC)

Effective anomaly detection needs to capture both fine-grained local details and long-range dependencies. Convolutional neural networks, while effective in capturing local dependencies, cannot model long-range dependencies because of their restricted receptive fields. Though transformers are capable of global feature extraction through self-attention, they nonetheless suffer from quadratic computational complexity, especially with high-resolution features, and are less effective at capturing small-scale local details. To address these challenges, we propose the DSSC module that leverages the strengths of both local and global modeling through the novel lightweight Pyramidal Scanning Approach (PSA) and the Depthwise Squeeze and Excitation Block (DSEB).

The PSA block utilizes multi-scale scanning to capture both global and local context. For an input feature map, $\tilde{\mathbf{X}}^{l}_{\text{enc}}$, the PSA block operates at multiple levels of the pyramid, which yields global features

$$\mathbf{X}_{\text{g}} = \texttt{PSA}_{M}(\cdots \texttt{PSA}_{1}(\texttt{PSA}_{0}(\tilde{\mathbf{X}}^{l}_{enc}))\cdots) \tag{9}$$

With M denoting the number of sequential PSA blocks.

Depthwise Squeeze and Excitation Block (DSEB) enhances the global feature extraction by capturing local features using lightweight depth-wise convolutions and Squeeze-and-Excite Convolution Block. Local features are acquired through two parallel blocks as:

$$\mathbf{X}_{\text{c},7} = \texttt{Conv}_{1\times1}(\texttt{DWConv}_{7\times7}(\texttt{Conv}_{1\times1}(\tilde{\mathbf{X}}^{l}_{enc}))) \tag{10}$$

$$\mathbf{X}_{\text{c},5} = \texttt{Conv}_{1\times1}(\texttt{DWConv}_{5\times5}(\texttt{Conv}_{1\times1}(\tilde{\mathbf{X}}^{l}_{enc}))) \tag{11}$$

$\mathbf{X}_{\text{c},5}$ and $\mathbf{X}_{\text{c},7}$ are concatenated along the channel dimension that produces the combined feature map. The concatenated output is then passed through a Squeeze-and-Excitation (SE) block to improve the feature representation:

$$\mathbf{X}_{\text{c}} = \text{SE}(\texttt{concat}(\mathbf{X}_{\text{c},7}, \mathbf{X}_{\text{c},5}))$$

Where SE$(\cdot)$ denotes the Squeeze-and-Excitation block on the concatenated feature map. PSA and DSEB features are fused as:

$$\mathbf{X}^{\text{DSSC}} = \texttt{Conv}_{1\times1}(\texttt{concat}(\mathbf{X}_{\text{g}}, \mathbf{X}_{\text{c}})) + \tilde{\mathbf{X}}^{l}_{enc} \tag{12}$$

where the residual connection enhances stability and convergence by improving gradient flow.

3.3. Pyramidal scanning approach (PSA)

The PSA block is proposed to address the shortcomings of standard convolutional networks in capturing both global and local spatial features effectively. The PSA block recursively applies an SSM across multiple scales and constructs a spatial pyramid representation of $\tilde{\mathbf{X}}^{l}_{enc}$, that enables the extraction of both fine-grained and coarse features.

Let $\mathbf{X}^{p}_{pyr} \in \mathbb{R}^{H^{p,i}\times W^{p,i}\times C^{p,i}}$ be the input at pyramid level $p \in \{0,\ldots,P\}$, where P is the total number of pyramid levels For such an input, the pth layer computes

$$\begin{aligned}\hat{\mathbf{X}}^{p}_{pyr} &= \texttt{PS}(\mathbf{X}^{p}_{pyr}) \\ &= \texttt{cat}\left(\left\{\texttt{SSM}\left(\mathbf{X}^{p,i}_{pyr}\right) | \mathbf{X}^{p,i}_{pyr} \in \texttt{split}\left(\mathbf{X}^{p}_{pyr}, p\right)\right\}\right)\end{aligned} \tag{13}$$

where

$$\texttt{split}\left(\mathbf{X}^{p}_{pyr}, p\right) \rightarrow \left\{\mathbf{X}^{p,i}_{pyr},\right\}_{i=1}^{2^{p}\times2^{p}} \tag{14}$$

yields $\mathbf{X}^{p,i}_{pyr}$ representing the ith input sub-region of size $H/2^{p} \times W/2^{p} \times C$ at level p.

To capture local dependencies, each sub-region of the input $\mathbf{x}_{i}$ is independently processed using an SSM$(\cdot)$ layer, following the SSM strategy proposed in Liu et al. (2024). cat$(\cdot)$ rearranges the processed feature maps back into their original grid format, keeping spatial dimensions intact. Due to such a rearrangement, the shapes of $\hat{\mathbf{X}}^{p}_{pyr}$, and $\mathbf{X}^{p}_{pyr}$ are same.

As shown in Fig. 2, we recursively apply the split$(\cdot)$ and SSM$(\cdot)$ operators to every $\mathbf{X}^{p,i}_{pyr}$ to ensure that both local details and broader contextual information are captured across the pyramid levels. Finally, the output is obtained by taking the average of outputs from level 0 to level p, across all sub-regions as:

$$\mathbf{Y}_{output} = \frac{1}{p+1}\left(\sum_{l=1}^{p}\hat{\mathbf{X}}^{l,i}_{pyr}\right) \tag{15}$$

The pyramid-based approach in the PSA block allows for multi-scale feature extraction, with each pyramid level capturing features at different scales. The four-way SS2D scanning strategy ensures that pixel-wise relationships are preserved both within and across patches, which is crucial for reconstructing spatial coherence after patch-wise operations.

3.4. Multi-scale feature fusion (MFF)

In the student-teacher network, the teacher model $\mathbf{X}^{l}_{\text{enc}}$ is more powerful than the student model $\mathbf{X}^{l}_{\text{dec}}$ (Deng and Li, 2022). Since $\mathbf{X}^{l}_{\text{dec}}$ tries to learn and restore the representations (features) from $\mathbf{X}^{l}_{\text{enc}}$, the simple approach is to feed the output of the last encoding layer of the backbone to $\mathbf{X}^{1}_{\text{dec}}$. But this direct feeding has two shortfalls:

(1) The final output ($\mathbf{X}^{3}_{\text{enc}}$) captures high-level features, but to reconstruct the normal image perfectly, the student model $\mathbf{X}^{l}_{\text{dec}}$ also needs to reconstruct low-level features.

A Multi-Scale Feature Fusion (MFF) block takes three feature maps from shallow, middle, and deep layers as input, each having different channel dimensions and spatial resolutions. To ensure dimensional compatibility, the shallow feature ($\mathbf{X}^{1}_{\text{enc}}$) is first downsampled using a 3×3 convolution with stride 2. This is progressively fused with the mid-level ($\mathbf{X}^{2}_{\text{enc}}$) and deep features ($\mathbf{X}^{3}_{\text{enc}}$) after aligning them using 1×1 convolutions to match channel dimensions.

(2) Since the teacher model $\mathbf{X}^{l}_{\text{enc}}$ is very powerful, it learns a lot of details, but not all of the details are important. These redundant and unnecessary details make it harder for the student model $\mathbf{X}^{l}_{\text{dec}}$ to focus on the most useful information. To fix this issue, the aggregated features from the MFF block are fed to the bottleneck layer. This layer compresses the size of the high-dimensional teacher's output into a more compact and smaller form, thus removing the unnecessary details that make learning easier for the student model $\mathbf{X}^{l}_{\text{dec}}$. Formally, the fusion process can be written as:

$$f_{1} = \text{Conv}_{1\times1}(\mathbf{X}^{2}_{\text{enc}}) + \text{Conv}_{3\times3}(\mathbf{X}^{1}_{\text{enc}}) \tag{16}$$

$$f_{2} = \text{Conv}_{3\times3}(f_{1}) + \text{Conv}_{1\times1}(\mathbf{X}^{3}_{\text{enc}}) \tag{17}$$

$$\mathbf{X}_{\text{MFF}} = \text{ResBlock}(\text{Conv}_{1\times1}(f_{2})) \tag{18}$$

3.5. Feature adapter

Generally, there is a distribution mismatch between the Industrial image dataset and the dataset used pre-trained model, therefore, the proposed model adopts a Feature adapter to transfer the training features to the target domain. To keep the feature adapter as light as possible while still being able to achieve competitive results, we employ a 1×1 convolution on the aggregated features.

3.6. Synthetic anomalies

The reconstruction-based techniques for anomaly localization rely on the assumption that the reconstruction difference will be high when unseen anomalous images are fed to the network during inference. The direct use of normal images in the training phase increases the chance of reconstructing the exact copy of any original image. To limit the

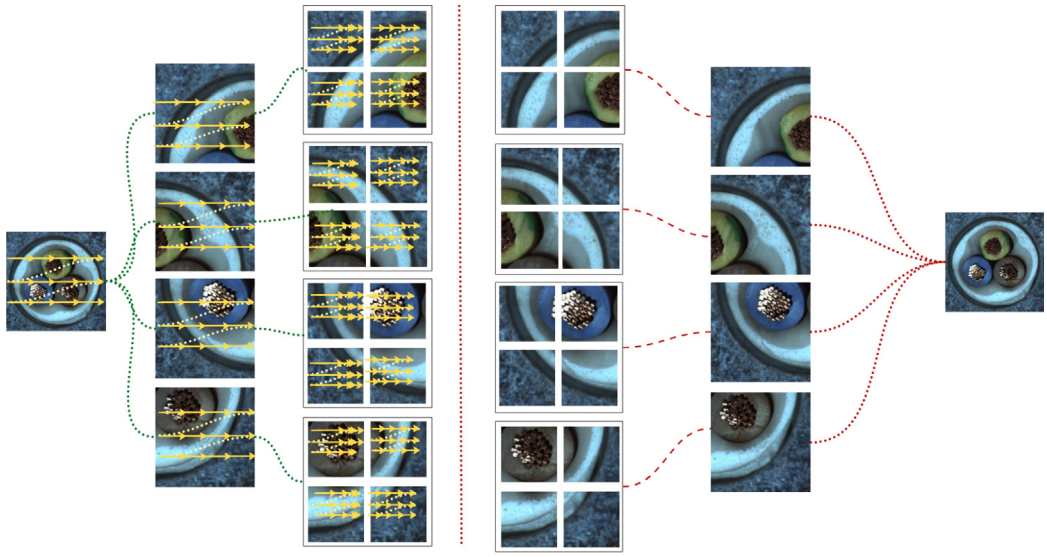


Fig. 2. Proposed pyramid-based scanning: (a) The Selective Scan Module (SSM) is first applied to the entire image. The entire image is then divided into four non-overlapping patches, and each patch is separately processed by SSM. Each patch can be further divided into sub-patches, and each sub-patch is independently processed by SSM (b) The processed feature maps are arranged back to their original grid structure.

learning of identical shortcuts, we add random noise to the normal features only in the training process, as follows:

$$\tilde{\mathbf{X}}_{\text{MFF}} = \mathbf{X}_{\text{MFF}} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \tag{19}$$

where $\boldsymbol{\epsilon}$ is zero-mean Gaussian noise with variance σ^2. where $\boldsymbol{\epsilon}$ is zero-mean Gaussian noise with variance σ^2.

This technique allows our model to simulate the characteristics of synthetic anomalies, which encourages the model to learn more robust features during training and improves the model's ability to generalize and recognize abnormal patterns, ultimately improving the Model's AD performance.

3.7. Training objective

Building upon the multi-scale feature representations defined in Section 3.4, let $\mathbf{X}^l_{\text{enc}}$ and $\mathbf{X}^l_{\text{dec}}$ denote the frozen teacher (encoder) and PAD-SSM based student (decoder) feature maps at layer l, respectively. To align student representations with the teacher across multiple semantic levels, we optimize our model using

$$\mathcal{L}_{\text{train}} = \sum_{l=1}^{L} \left\| \mathbf{X}^l_{\text{enc}} - \mathbf{X}^l_{\text{dec}} \right\|_2^2 \tag{20}$$

This objective encourages the student network to approximate the feature distribution of the teacher across all hierarchical levels, improving sensitivity to anomalous deviations.

Anomaly map inference. During inference, anomaly localization is performed by measuring discrepancies between teacher and student feature representations.

For each layer l and pixel location, we compute the cosine distance along the channel dimension between normalized feature embeddings. This yields the pixel-wise anomaly map as:

$$\mathbf{A}^l = 1 - \cos\left(\mathbf{X}^l_{\text{enc}}, \mathbf{X}^l_{\text{dec}}\right) \tag{21}$$

measuring directional inconsistencies between teacher and student feature representations.

Since different layers produce feature maps at different spatial resolutions, the pixel-wise anomaly maps $\{\mathbf{A}^l\}_{l=0}^3$ have dimensional inconsistencies. To deal with this, we uniformly average the bilinearly upsampled layer-wise anomaly maps as

$$\mathbf{A} = \frac{1}{L} \sum_{l=1}^{L} \texttt{Upsample}(\mathbf{A}^l, H^l, W^l) \tag{22}$$

To obtain the final anomaly map while improving spatial consistency and suppressing high-frequency noise, we perform Gaussian smoothing yielding

$$\tilde{\mathbf{A}} = \mathcal{G}_{\sigma_{blur}} * \mathbf{A} \tag{23}$$

where $\mathcal{G}_{\sigma_{blur}}$ is a 2D Gaussian kernel, with standard deviation $\sigma_{blur} = 4$.

4. Experiments

4.1. Setups: Datasets, metrics, and implementation

Datasets. The **MVTec-AD** (Bergmann et al., 2019) benchmark includes a total of 5354 images that are split into 3629 training anomaly-free and 1725 test images (with and without anomalies). The **VisA** (Zou et al., 2022) dataset has 10,821 images (9621 normal and 1200 anomalous samples) distributed across 12 categories. These public benchmark datasets provide different types of anomalous samples: single instances, multiple instances, and complex structures that are fundamental to evaluating AD methods.

Metrics. We report on the Area Under the Receiver Operating Characteristic Curve (AU-ROC), Average Precision (Zavrtanik et al., 2021) (AP), F1-score-max (Zou et al., 2022) ($F1_{\text{max}}$), and Area Under the Per-Region-Overlap (Bergmann et al., 2020) (AU-PRO) for anomaly detection and localization.

Implementation Details. The experiments are conducted on a 256 × 256 input image without any further augmentation. ResNet34 is adopted as a pre-trained feature encoder. The decoder has $[3, 4, 6, 3]$ DSSC modules for the corresponding encoder layers. We used $P = 1$ pyramid levels. The loss function during training is computed as the sum of Mean Squared Error at multiple scales. The Adam optimizer is used to optimize the model with a decay rate of 1e-4 and a learning rate of 5e-4 for 500 epochs.

Table 1
Comparison with SoTA methods on MVTec-AD dataset for multi-class anomaly detection with AU-ROC/AP/$F1_{max}$. Bold indicates the best performance, and red indicates the lowest FLOPs.

Class	SimpleNet	DeSTeg	DiAD	MambaAD	Pyramid-Mamba	PAD-SSM(Ours)
Bottle	**100.0/100.0/100.0**	98.7/99.6/96.8	99.7/96.5/91.8	**100.0/100.0/100.0**	**100.0/100.0/100.0**	**100.0/100.0/100.0**
Cable	07.5/98.5/94.7	89.5/94.6/85.9	94.8/98.8/95.2	98.8/99.2/95.7	99.2/99.5/96.8	98.9/99.3/96.2
Capsule	90.7/97.9/93.5	82.8/95.9/92.6	94.4/98.7/94.9	94.1/96.9/96.9	95.1/99.0/94.8	**96.3/99.2/95.9**
Hazelnut	99.9/99.9/99.3	98.8/99.2/98.6	99.5/99.7/97.3	**100.0/100.0/100.0**	**100.0/100.0/100.0**	**100.0/100.0/100.0**
Metal Nut	96.9/99.3/96.1	92.9/98.4/92.2	99.1/96.0/91.6	99.9/100.0/99.5	99.7/99.9/99.5	**100.0/100.0/100.0**
Pill	88.2/97.7/92.5	77.1/94.4/91.7	95.7/98.5/94.5	97.0/99.5/96.2	96.9/99.5/96.6	**98.2/99.7/97.9**
Screw	76.7/90.6/87.7	69.9/88.4/85.4	90.7/**99.7**/97.9	94.7/97.9/94.0	96.9/99.5/**96.6**	93.8/97.5/94.2
Toothbrush	89.7/95.7/92.3	71.7/89.3/84.5	**99.7/99.9/99.2**	98.3/99.3/98.4	94.6/97.9/95.9	**100.0/100.0/100.0**
Transistor	99.2/98.7/97.6	78.2/79.5/68.8	99.8/99.6/97.4	**100.0/100.0/100.0**	98.3/99.3/96.8	**100.0/100.0/100.0**
Zipper	99.0/99.7/98.3	88.4/96.3/93.1	95.1/99.1/94.4	99.3/99.8/97.5	**100.0/99.9/99.9**	99.6/99.7/98.3
Carpet	95.7/98.7/93.2	95.9/98.8/94.9	99.4/99.9/98.3	99.8/99.9/99.4	99.1/99.8/97.9	**99.9/100/99.4**
Grid	97.6/99.2/96.4	97.9/99.2/96.6	98.5/99.8/97.7	**100.0/100.0/100.0**	**100.0/100.0/100.0**	**100.0/100.0/100.0**
Leather	**100.0/100.0/100.0**	99.2/99.8/98.9	99.8/99.7/97.6	**100.0/100.0/100.0**	**100.0/100.0/100.0**	**100.0/100.0/100.0**
Tile	99.3/99.8/98.8	97.0/98.9/95.3	96.8/99.9/98.4	98.2/99.3/95.4	**98.6/99.5/96.5**	**99.0/99.6/97.7**
Wood	98.4/99.5/96.7	99.9/100.0/99.2	**99.7/100.0/100.0**	98.8/99.6/96.6	99.4/99.8/98.3	99.1/99.7/96.8
Mean	95.3/98.4/95.8	89.2/95.5/91.6	97.2/99.0/96.5	98.6/**99.6**/97.8	98.6/99.5/97.9	**98.7/99.6/98.1**
FLOPs (G)	16.1	35.2	451.5	8.7	9.7	8.7

Table 2
Comparison with SoTA methods on MVTec-AD dataset for multi-class anomaly localization with AU-ROC/AP/$F1_{max}$/AU-PRO.

Class	SimpleNet	DeSTeg	DiAD	MambaAD	Pyramid-Mamba	PAD-SSM(Ours)
Bottle	97.8/68.2/67.6/94.0	97.2/53.8/62.4/89.0	98.4/52.2/54.8/86.6	98.8/79.7/76.7/95.2	**98.9/81.0**/77.4/96.1	**98.9**/79.9/76.8/**96.1**
Cable	85.1/26.3/33.6/75.1	96.7/42.4/51.2/85.4	96.8/50.1/57.8/80.5	95.8/42.2/48.1/**90.3**	**96.8**/42.9/48.9/90.1	97.9/**52.1/58.1/91.8**
Capsule	98.8/43.4/50.0/94.8	98.5/5.4/44.3/84.5	97.1/42.0/45.3/87.2	98.4/43.9/47.7/92.6	**98.6/45.2/48.3/94.0**	**98.6**/43.0/46.6/**94.2**
Hazelnut	97.9/36.2/51.6/92.7	98.4/44.6/51.4/87.4	98.3/79.2/80.4/91.5	99.0/63.6/64.4/95.7	**99.2/70.2/68.8/95.6**	99.1/68.5/66.5/**95.9**
Metal Nut	94.8/55.5/66.4/91.9	98.0/83.1/79.4/85.2	97.3/30.0/38.3/90.6	96.7/74.5/79.1/93.7	**97.1/77.3/80.2/93.9**	96.9/74.9/79.6/93.0
Pill	97.5/63.4/65.2/95.8	96.5/72.4/67.7/81.9	95.7/46.0/51.4/89.0	97.4/64.0/66.5/95.7	97.6/65.8/67.4/97.1	**98.3/73.1/71.9**/96.6
Screw	99.4/40.2/44.6/96.8	96.5/15.9/23.2/84.0	97.9/60.6/59.6/95.0	99.5/49.8/50.9/97.1	**99.5/51.1/51.0/97.6**	**99.5**/48.7/49.5/97.2
Toothbrush	99.0/53.6/58.8/92.0	98.4/46.9/52.5/87.4	**99.0/78.7/72.8/95.0**	99.0/48.5/59.2/91.7	99.0/49.1/60.8/92.5	99.0/48.4/60.2/92.0
Transistor	85.9/42.3/45.2/74.7	95.8/58.2/56.0/83.2	95.1/15.6/31.7/90.0	96.5/69.4/**67.1**/87.0	**96.9/69.7**/67.0/**91.6**	96.3/67.2/65.4/90.9
Zipper	98.5/53.9/60.3/94.1	97.9/53.4/54.6/90.7	96.2/60.7/60.0/91.6	98.4/**60.4/61.7**/94.3	98.3/59.1/61.2/**94.6**	98.5/61.8/62.2/95.1
Carpet	99.0/58.5/60.4/95.1	97.4/38.7/43.2/90.6	98.6/42.2/46.4/90.6	99.2/60.0/63.3/96.7	**99.3/65.1/64.5/97.2**	99.2/62.3/63.4/**97.2**
Grid	96.5/23.0/28.4/97.0	96.8/20.5/27.6/88.6	97.0/42.1/46.9/86.8	99.2/47.4/47.7/97.0	**99.2/49.1/48.5/97.5**	99.2/48.0/49.1/97.3
Leather	99.3/38.0/45.1/97.4	98.7/28.5/32.9/92.7	98.8/56.1/62.3/91.3	99.4/50.3/53.3/98.7	**99.4/50.4/53.5/98.7**	99.4/51.0/51.2/98.7
Tile	95.3/48.5/60.5/85.8	95.7/60.5/59.9/90.6	92.4/65.7/64.1/90.7	93.8/45.1/54.8/80.0	93.8/45.5/55.0/81.7	**94.9/47.8/57.8/82.7**
Wood	95.3/47.8/51.0/90.0	91.4/34.8/39.7/76.3	93.3/43.3/43.5/97.5	94.4/46.2/48.2/91.2	**94.7/48.3/49.4/99.8**	93.6/44.3/45.3/91.4
Mean	96.9/45.9/49.7/86.5	93.1/54.3/50.9/64.8	96.8/52.6/55.5/90.7	97.7/56.3/59.2/93.1	**97.8/57.3/59.7/93.5**	**97.8**/56.3/59.0/93.4

4.2. Comparison with state-of-the-art methods

Table 1 demonstrates a quantitative comparison of the proposed method against state-of-the-art methods for anomaly detection on the MVTec-AD. Our proposed method shows competitive performance in the anomaly detection setting a new state-of-the-art across all anomaly detection metrics. Compared to the previous best (Iqbal and Martinel, 2025), our model achieves the highest mean AU-ROC (+0.1), AP (+0.1), and $F1_{max}$ (+0.2).

Notably, the proposed method surpasses or matches the best results in most of the classes (11/15). It also shows perfect detection (100% across all metrics) for 7/15 classes, across multiple types of defects. The enhanced anomaly detection results can be attributed to the integration of domain transfer techniques and SE-Enhanced Convolution Block combined with Pyramid-mamba, leading to improved performance beyond what was achievable with the Pyramid-mamba architecture alone. Furthermore, in Table 1, we show that our model has an equal number of Flops compared to MambaAD (He et al., 2024b), but significantly lower than RD4AD (Deng and Li, 2022), DiAD (He et al., 2024a), and Pyramid-Mamba (Iqbal and Martinel, 2025). Despite fewer Flops, our model still achieves higher efficiency compared to baseline models.

Table 2 demonstrates a comprehensive comparison of our proposed method against SOTA models for anomaly localization on the MVTec-AD. On overall pixel-level localization, our single-level P_1 design matches Pyramid-Mamba's mean Pixel AU-ROC (97.8) while demonstrating substantial performance leads on fine-structured categories that suffer under P_2 spatial collapse. For instance, on *Cable*, our model achieves **97.9 AU-ROC/52.1 AP** (compared to Pyramid-Mamba's 96.8/42.9, representing a **+9.2 AP gain**). Similarly, on *Pill* and *Tile*, our model delivers notable localization AP improvements of **+7.3** (73.1 vs. 65.8) and **+2.3** (47.8 vs. 45.5), respectively. These gains confirm that preserving spatial fidelity at level P_1 via multi-kernel enhancement is significantly more effective for precise defect localization than deeper downsampling pyramids.

Tables 3 and 4 compare our proposed model against state-of-the-art methods for multi-class anomaly detection and localization on the VisA dataset, respectively.

The model shows particular strength in challenging categories, with significant improvements in $F1_{max}$ for pcb2 (+1.4) and pcb3 (+1.5) compared to the previous best (Iqbal and Martinel, 2025). Table 4 reveals even more pronounced advancements in anomaly localization. Our approach achieves new state-of-the-art results in mean AP (+3.2%) and $F1_{max}$ (+0.9%) compared to baseline model (He et al., 2024b). PAD-SSM exhibits exceptional performance in complex categories such as cashew($F1_{max}$: +9.2), pipe-fyrum (AP: +8.2), and macaroni1 (AU-PRO: +0.9) compared to previous best (Iqbal and Martinel, 2025). The consistently strong performance across both detection and localization tasks underscores the versatility of our novel model.

Qualitative Results.

Fig. 3 provides the visual assessment of the performance of PAD-SSM as compared to other SOTA models. From the figure, it can be observed that our model effectively detects anomalous regions in most cases. For example, in a cable heatmap image, most of the SOTA methods miss the localization of some portion of the anomaly, while

Table 3
Comparison with SoTA methods on VisA dataset for multi-class anomaly detection with AU-ROC/AP/$F1_{max}$ metrics.

Class	SimpleNet	DeSTeg	DiAD	MambaAD	Pyramid-Mamba	PAD-SSM(Ours)
pcb1	91.6/91.9/86.0	87.6/83.1/83.7	88.1/88.7/80.7	95.4/93.0/91.6	96.0/94.6/**94.1**	**96.1**/**95.0**/93.3
pcb2	92.4/93.3/84.5	86.5/85.8/82.6	91.4/91.4/84.7	94.2/93.7/89.3	95.2/95.1/90.5	**96.7**/**96.7**/**91.9**
pcb3	89.1/91.1/82.6	93.7/95.1/87.0	86.2/87.6/77.6	93.7/94.1/86.7	95.3/95.3/88.9	**95.9**/**96.1**/**90.4**
pcb4	97.0/97.0/93.5	97.8/97.8/92.7	99.6/99.5/97.0	**99.9**/**99.9**/**98.5**	99.8/99.8/**98.5**	99.7/99.7/97.1
macaroni1	85.9/82.5/73.1	76.6/69.0/71.0	85.7/85.2/78.8	91.6/89.8/81.6	**92.2**/**90.5**/**83.3**	91.5/89.5/82.7
macaroni2	68.3/54.3/59.7	68.9/62.1/67.7	62.5/57.4/69.6	81.6/78.0/73.8	**83.1**/**82.3**/**74.8**	80.0/75.8/ 73.1
capsules	74.1/82.8/74.6	87.1/93.0/84.2	58.2/69.0/78.5	91.8/**95.0**/88.8	**91.0**/94.5/89.1	90.9/ 94.2/**89.1**
candle	84.1/73.3/76.6	94.9/94.8/89.2	92.8/92.0/87.6	96.8/96.9/90.1	**97.4**/**97.6**/**92.5**	95.2/95.7/89.1
cashew	88.0/91.3/84.7	92.0/96.1/88.1	91.5/95.7/89.7	94.5/97.3/91.1	**96.4**/**97.3**/**92.0**	94.0/ 97.2/ 91.5
chewinggum	96.4/98.2/93.8	95.8/98.3/94.7	99.1/99.5/95.9	97.7/98.9/94.2	**98.1**/**99.1**/95.2	97.7/ 98.9/95.4
fryum	88.4/93.0/83.3	92.1/96.1/89.5	89.8/95.0/87.2	95.2/97.7/90.5	**96.7**/**98.4**/**94.1**	**96.7**/ **98.4**/93.4
pipe_fryum	90.8/95.5/88.6	94.1/97.1/91.9	96.2/98.1/93.7	98.7/99.3/97.0	**98.8**/**99.4**/**97.5**	98.7/99.3/97.0
Mean	87.2/87.0/81.8	88.9/89.0/85.2	86.8/88.3/85.1	**94.3**/**94.5**/89.4	94.1/94.4/**89.9**	93.9/94.2/89.5
FLOPs (G)	16.1	35.2	451.5	8.7	9.7	8.7

Table 4
Comparison with SoTA methods on VisA dataset for multi-class anomaly localization with AU-ROC/AP/$F1_{max}$/AU-PRO metrics.

Class	SimpleNet	DeSTeg	DiAD	MambaAD	Pyramid-Mamba	PAD-SSM(Ours))
pcb1	99.2/86.1/78.8/83.6	95.8/46.4/49.0/83.2	98.7/49.6/52.8/80.2	99.8/77.1/72.4/92.8	**99.8**/80.2/73.5/94.6	**99.8**/**81.6**/**74.1**/**94.9**
pcb2	96.6/8.9/18.6/85.7	97.3/14.6/28.2/79.9	95.2/7.5/16.7/67.0	98.9/13.3/23.4/89.6	**99.0**/15.7/24.3/91.3	**99.0**/**15.8**/23.8/**91.8**
pcb3	97.2/31.0/36.1/85.1	97.7/28.1/33.4/62.4	96.7/8.0/18.8/68.9	99.1/18.3/27.4/89.1	**99.2**/**19.4**/**28.1**/**92.4**	99.2/18.4/**28.1**/91.8
pcb4	93.9/23.9/32.9/61.1	95.8/53.0/53.2/76.9	97.0/17.6/27.2/85.0	98.6/47.0/46.9/87.6	**98.8**/47.1/47.7/90.3	98.7/48.2/48.1/91.0
macaroni1	98.9/3.5/8.4/92.0	99.1/5.8/13.4/62.4	94.1/10.2/16.7/68.5	99.5/17.5/27.6/95.2	**99.6**/**19.7**/**29.7**/95.4	**99.6**/17.6/26.7/**96.3**
macaroni2	93.2/0.6/3.9/77.8	98.5/6.3/14.4/70.0	93.6/0.9/2.8/73.1	99.5/9.2/16.1/96.2	**99.4**/**10.7**/**17.5**/**96.3**	99.3/7.7/14.9/96.1
capsules	97.1/52.9/53.3/73.7	96.9/33.2/9.1/76.7	97.3/10.0/21.0/77.9	**99.1**/**61.3**/**59.8**/91.8	**99.1**/60.6/59.8/**93.4**	**99.1**/56.6/57.4/91.7
candle	97.6/8.4/16.5/87.6	98.7/39.9/45.8/69.0	97.3/12.8/22.8/89.4	99.0/23.2/32.4/95.5	**99.1**/**23.7**/**32.7**/**96.1**	**99.1**/23.4/32.4/95.8
cashew	98.9/68.9/66.0/84.1	87.9/47.6/52.1/66.3	90.9/53.1/60.9/61.8	94.3/46.8/51.4/87.8	95.8/52.6/55.8/**89.1**	**99.0**/**66.8**/**65.0**/**89.1**
chewinggum	97.9/26.8/29.8/78.3	98.8/86.9/81.0/68.3	94.7/11.9/25.8/59.5	98.1/57.5/59.9/79.7	98.2/62.9/62.2/**80.1**	**98.6**/58.4/60.7/**80.1**
fryum	93.0/39.1/45.4/85.1	88.1/35.2/38.5/47.7	97.6/**58.6**/**60.1**/81.3	96.9/47.8/51.9/91.6	**97.3**/50.1/53.0/92.2	97.4/51.6/55.2/91.8
pipe_fryum	98.5/65.6/63.4/83.0	98.9/78.8/72.7/45.9	**99.4**/**72.7**/**69.9**/89.9	99.1/53.5/58.5/95.1	99.2/56.0/58.3/**95.3**	99.2/64.2/62.4/93.2
Mean	96.8/34.7/37.8/81.4	96.1/**39.6**/43.4/67.4	96.0/26.1/33.0/75.2	98.5/39.4/44.0/91.0	98.6/40.1/44.3/**91.7**	**98.9**/**40.7**/**44.4**/91.2

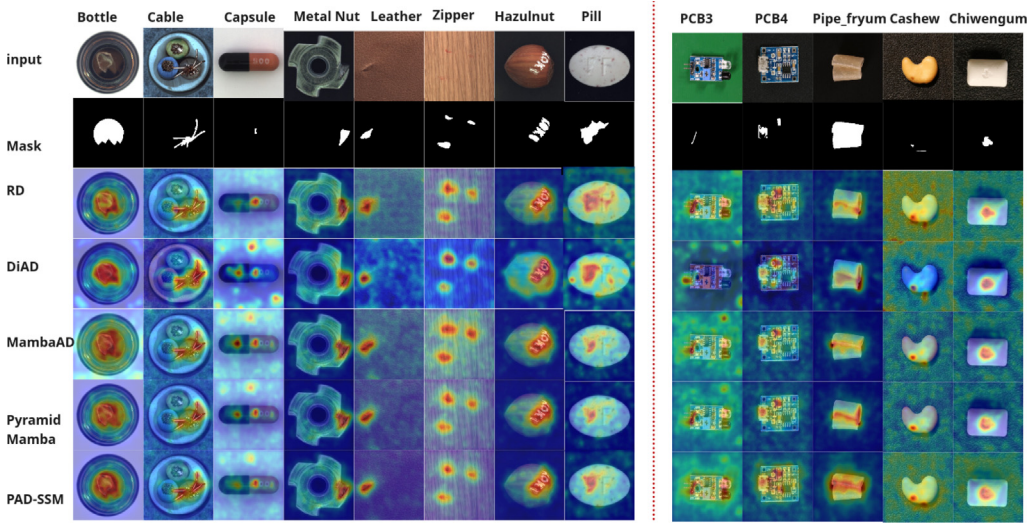


Fig. 3. Qualitative visualization for pixel-level anomaly segmentation on MVTec and VisA datasets.

PAD-SSM accurately localizes that anomaly. Compared to other SOTA models, PAD-SSM shows fewer false positives in small and challenging anomalies. For instance, in a capsule image heatmap, most of the SOTA methods falsely localize the normal region as an anomaly; on the other hand, the PAD-SSM localizes this anomaly with reasonable accuracy.

4.3. Ablation study

Through the ablation study, we want to answer different questions that would help us understand the importance of each proposed component of our architecture.

Table 5
Comparison of mean performance values for anomaly detection and localization tasks across different ablations.

Category	MVTec Anomaly detection (AU-ROC/AP/$F1_{max}$)
w/o DSEB	98.5/99.4/97.7
w/o PSA	96.5/98.7/96.7
PAD-SSM	**98.7**/**99.6**/**98.1**

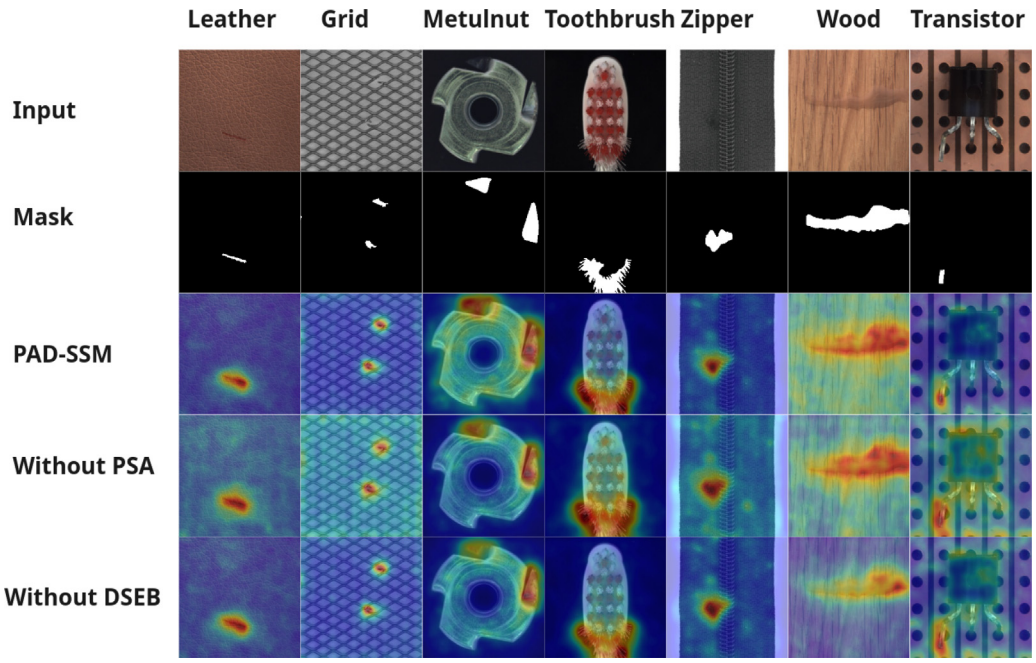


Fig. 4. Qualitative comparison of pixel-level anomaly segmentation across different component configurations.

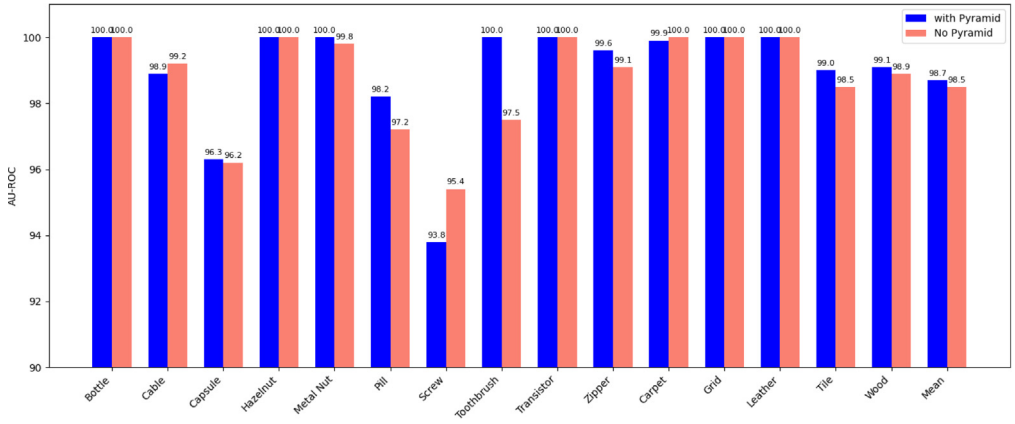


Fig. 5. Ablation studies on multiclass anomaly localization for different pyramid levels. Results are on the MVTec dataset.

What is the impact of local (DSEB) and global (PSA) features? Table 5 and Fig. 4 together demonstrate the importance of local and global features in the anomaly detection task. Removing the DSEB block leads DSEB leads to a 0.4 $F1_{\max}$ drop in detection, demonstrating that DSEB is critical for capturing local information and enhancing anomaly detection. On the other hand, removing the PSA block significantly drops the detection accuracy. Specifically, the AU-ROC is dropped by 2.2 while $F1_{\max}$ decreases 1.4. These performance gaps signify that for accurate anomaly detection in multiclass settings, both local and global features are vital.

Fig. 4 shows a qualitative comparisons to illustrate the contribution of each proposed module under different component/ablation settings. Following the quantitative performance shown in Table 5, the visual results further demonstrate that removing the PSA module leads to a noticeable increase in false positives across multiple samples (e.g., the wood and transistor examples). Operating without the DSEB component results in unstable predictions, causing both over-detection and completely missed anomalies depending on the sample type (e.g., the wood image example shows only a partial detection of the defective region). Our full method shows precise anomaly localization producing final segmentation maps that tightly match the annotated ground truth.

Ablation Study on DSEB Architecture. To evaluate the effectiveness of the proposed Dual-branch Spatial Enhancement Block (DSEB), we conduct a comparative ablation study within the complete PAD-SSM framework. Specifically, we compare DSEB against three alternative module configurations: (1) the **Original Baseline** (He et al., 2024b) (2) the **Prior LEC Module** from existing literature (Iqbal and Martinel, 2025), and (3) a **Standard Depthwise Bottleneck (DW 3 × 3)** consisting of a lightweight depthwise separable block ($1 \times 1 \rightarrow 3 \times 3$ DW Conv $\rightarrow 1 \times 1$).

As presented in Table 6, integrating DSEB yields clear empirical accuracy gains over baseline variants. Replacing DSEB with the uncalibrated LEC module results in a performance drop, reducing image-level Detection AU-ROC from 98.7 to 98.3, and $F1_{\max}$ 98.1 → 97.7, while pixel-level Localization AP drops by 1.0 (56.3 → 55.3) and $F1_{\max}$ drops by 0.5 (59.0 → 58.5). Furthermore, replacing the multi-branch structure with a simple Depthwise 3×3 bottleneck leads to additional degradation in pixel-level boundary localization ($F1_{\max}$ drops to 58.1, AU-PRO drops to 93.0). These results confirm that DSEB's multi-scale receptive fields and dynamic channel recalibration are essential for distinguishing subtle defect textures from complex backgrounds.

Importance of the pyramid levels. Fig. 5 depicts the performance of the proposed method across multiple object categories for

Table 6
Ablation study comparing the proposed DSEB against the prior LEC module and standard lightweight Depthwise (3 × 3) Bottleneck baselines within PAD-SSM. Bold indicates the best performance.

Module Variant	Detection (AUROC/ AP/$F1_{max}$)	Localization (AUROC/ AP/$F1_{max}$/AUPRO)
Proposed DSEB	**98.7/99.6/98.1**	**97.8/56.3**/59.0/93.4
Original Baseline (He et al., 2024b)	98.6/99.6/97.8	97.7/**56.3/59.2**/93.1
PAD-SSM w/ LEC	98.3/99.5/97.7	97.7/55.3/58.5/93.2
DW 3 × 3 Bottleneck	98.2/99.4/97.6	97.6/55.1/58.1/93.0

Table 7
Mean performance values for anomaly detection and localization tasks on the MVTec-AD dataset with various types of artificial noise.

Setting	Detection (AU-ROC/AP/$F1_{max}$)	Localization (AU-ROC/AP/$F1_{max}$/AU-PRO)
No Noise, No Pyramid	98.5/99.5/97.8	97.7/55.9/58.5/93.1
No Noise, Pyramid	98.3/99.4/97.6	97.5/55.3/58.1/92.8
Feature Masking, No Pyramid	98.3/99.4/97.9	97.7/56.2/58.6/93.2
Feature Masking, Pyramid	98.4/99.6/97.9	97.7/56.0/58.4/92.2
Feature Noise, No Pyramid	98.3/99.4/97.9	97.7/56.2/58.5/93.2
Feature Noise, Pyramid	**98.7/99.6/98.1**	**97.8/56.3/59.0/93.4**

Table 8
Comparison of mean performance values for anomaly detection and localization tasks across different scanning methods.

Category	Det. (AU-ROC/AP/$F1_{max}$)	Loc. (AU-ROC/AP/$F1_{max}$/AU-PRO)
Zigzag	98.7/99.5/97.8	97.8/56.5/58.9/93.1
Scan	98.5/99.4/97.8	97.8/56.5/59.0/93.0
Hilbert	98.6/99.5/97.9	97.3/53.5/56.2/92.0
Sweep	**98.7/99.6/98.1**	**97.8/56.3/59.0/93.4**

multiclass anomaly localization, using the AU-ROC metric with and without pyramid level. Utilizing the pyramid level (blue bars) consistently improves performance, yielding the highest scores across most (7 out of 15) categories compared to the no-pyramid (orange bars) configurations. Categories like *Zipper*, *Pill*, and *Tile* demonstrate a significant improvement in AU-ROC when pyramid-based operation is implemented.

This clear performance margin validates the importance of pyramid-based processing, as it detects both small and large anomalies at multiple scales.

Effect of different types of synthetic anomalies. Table 7 demonstrates the average performance metrics for the multiclass anomaly detection and localization tasks, examined across different noise conditions and methods such as feature-level noise and feature-level masking conducted with and without pyramid pre-processing. For anomaly detection, the highest mean AU-ROC/AP/$F1_{max}$ values are observed when the pyramid-based approach is employed along with the feature-level noise, achieving scores of 98.7/99.6/98.1. Table 7 also shows that this approach achieves the highest scores for the anomaly localization task. especially, the AP performance drops by about 0.9% when the experiments were conducted with no-pyramid level and the no-noise, which indicates that both feature-level noise and pyramid-based processing are essential to capture fine-grained details.

Effectiveness comparison of different scanning methods.

We compare four scanning strategies: Zigzag, Scan, Hilbert scanning, and Sweep scanning. The quantitative results are reported in Table 8, and the corresponding traversal patterns are illustrated in Fig. 6.

As shown in Table 8, Sweep scanning achieves the best overall performance across all anomaly detection and most of the localization metrics, indicating its effectiveness even compared to existing advanced scanning schemes.

Feature Adapter Analysis.

In Table 9, we analyze the performance of different adapters aligning the decoder representations with the encoder feature space in the target domain. Results show that removing the feature adapter (no feature adapter) yields to a noticeable detection and localization performance degradation. This indicates that feature adaptation is important to mitigate the ImageNet-industrial imaging domain shift. Replacing the proposed $\texttt{Conv}_{1\times1}$, with an MLP or a more complex CBAM-based

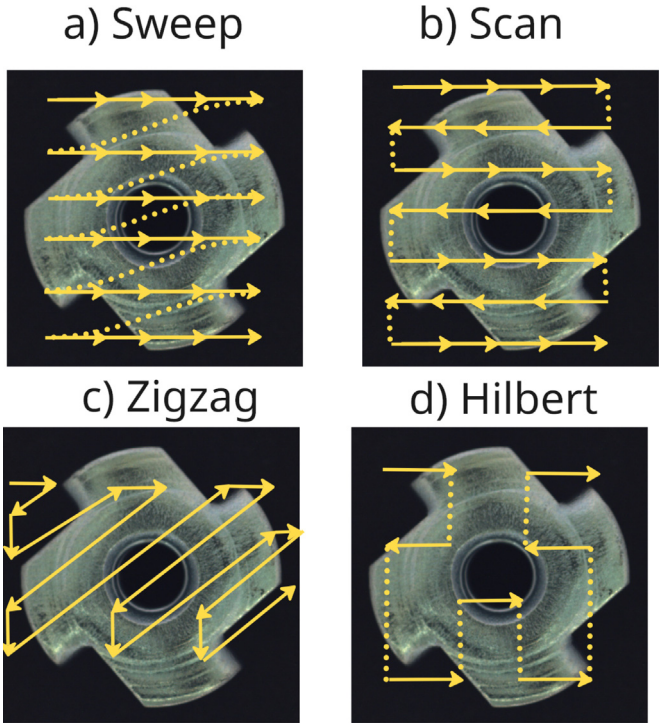


Fig. 6. Ablation on Different scanning methods.

Table 9
Ablation study on different types of feature adapters. Bold indicates the best-performing result.

Setting	Det.(AUROC /AP/$F1_{max}$)	Loc. (AUROC /AP/$F1_{max}$ /AUPRO)	GFLOPs	Params	Speed (samples/s)
No feature Adapter	98.4/99.5/98.0	97.7/56.5/58.7/93.1	8.7 G	**16.38 M**	44.54
MLP	98.5/99.5/97.9	**97.8**/56.5/58.9/**93.5**	8.7 G	16.52 M	43.46
CBAM	98.6/99.5/97.9	**97.8**/**57.0**/**59.2**/93.4	8.7 G	16.65 M	43.74
$\texttt{Conv}_{1\times1}$	**98.7**/**99.6**/**98.1**	**97.8**/56.3/59.0/93.4	8.7 G	16.65 M	**45.27**

Table 10
Computational complexity comparison of existing methods on MVTec Anomaly Detection (AU-ROC/AP/$F1_{max}$), and MVTec Anomaly Localization (AU-ROC/AP/$F1_{max}$/AU-PRO).

Mthods	DeSTeg	Pyramid mamba	PAD-SSM -level2	PAD-SSM (Ours)
Detection	89.2/95.5/91.6	98.6/99.5/97.9	**98.7**/**99.6**/**98.1**	**98.7**/**99.6**/98.1
Localization	96.9/45.9/49.7/86.5	**97.8**/**57.3**/**59.7**/**93.5**	97.7/56.1/58.6/93.3	**97.8**/56.3/59.0/93.4
Flops	35.2 G	9.7 G	9.8 G	8.7G
Params	35.154 M	14.93 M	16.65 M	16.65M
Latency	9.23 ms	52.86 ms	54.57 ms	22.21 ms
Throughput (samples/s)	108.31	18.60	18.87	45.27

adapter, yields to similar performance but with lower throughput. This substantiates the fact that proposed adapter is an effective yet efficient approach.

4.4. Complexity analysis.

To evaluate the computational efficiency and practical deployment feasibility of the proposed framework, we conduct a comprehensive complexity analysis comparing `PAD-SSM` against state-of-the-art anomaly detection architectures.

Table 10 summarizes the comparative computational footprint, inference latency, and throughput across models. To ensure complete technical rigor and fair evaluation, all computational complexity metrics (FLOPs) are profiled using an end-to-end evaluation harness built with DeepSpeed's `FlopsProfiler` (Rasley et al., 2020).

As reported in Table 10, under this comprehensive profiling setup, `PAD-SSM` requires only **8.7 G FLOPs**, compared to **9.7 G FLOPs** for Pyramid-Mamba (Iqbal and Martinel, 2025) and **9.8 G FLOPs** for the multi-level baseline variant (`PAD-SSM-level2`). Eliminating the computationally redundant second-level pyramid branch (P_2) yields a **10.3% net reduction in FLOPs** while preserving multi-scale feature representations.

As detailed in Table 10:

- **Inference Latency:** `PAD-SSM` drastically reduces average per-sample latency from **52.86 ms** (Pyramid-Mamba) down to **22.21 ms** (a **58.0% latency reduction**).
- **Processing Throughput:** Our single-level pyramid architecture achieves a throughput of **45.27 samples/s**, delivering a 2.43× **speedup** over Pyramid-Mamba (**18.60 samples/s**).

These empirical results demonstrate that the proposed architecture provides an optimal, highly scalable balance between pixel-level detection accuracy and real-time operational efficiency.

5. Conclusion

We introduce a novel approach, PAD-SSM, for multi-class anomaly detection and localization in industrial images. The proposed method consists of a pre-trained encoder, a synthetic anomaly generator, which perturbs the image representation in feature space, a feature adapter to transform the encoded features to a target domain, and a novel pyramid-based Mamba decoder. Within each Mamba decoder, the DSSC module includes a PSA block that performs a pyramidal-based scanning, which integrates the long-range modeling capabilities of Mamba at different levels. The proposed PSA block divides feature maps into multiple patches, and scans the entire image and each patch independently in four different directions. This approach ensures that both global and local dependencies are captured, which is vital for accurate localization of small anomalies in high-resolution industrial images. In addition, we introduced the lightweight DSEB block inside the DSSC module. The combination of the PSA block and the DSEB captures both local and global features, enabling the detection of fine-grained anomalies. Extensive experiments conducted on two datasets and evaluated with seven different metrics demonstrate that with minimal FLOPs, PAD-SSM achieves superior performance on multiclass anomaly detection while achieving competitive performance on multi-class anomaly localization.

CRediT authorship contribution statement

Nasar Iqbal: Writing – original draft. **Niki Martinel:** Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author(s) used Gemini (Google) in order to refine the text formatting, shorten the highlights, and check journal submission compliance. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Università degli Studi di Udine through a PhD grant awarded to N.I. under the National Recovery and Resilience Plan (PNRR), Mission 4, Component 2, Investment 3.3 (D.M. 117/2023), funded by the European Union – NextGenerationEU (CUP G23C23001180005).

The authors acknowledge EYE-Tech S.r.l., Pordenone, Italy, for its support and involvement in N.I.'s doctoral research on the study and design of artificial intelligence algorithms for quality control of production processes.

References

Bergmann, P., Fauser, M., Sattlegger, D., Steger, C., 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9592–9600. http://dx.doi.org/10.1109/CVPR.2019.00982.

Bergmann, P., Fauser, M., Sattlegger, D., Steger, C., 2020. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 4183–4192. http://dx.doi.org/10.1109/CVPR42600.2020.00424.

Defard, T., Setkov, A., Loesch, A., Audigier, R., 2021. Padim: A patch distribution modeling framework for anomaly detection and localization. In: International Conference on Pattern Recognition. ICPR, Springer, pp. 475–489.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, http://dx.doi.org/10.1109/CVPR.2009.5206848.

Deng, H., Li, X., 2022. Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9737–9746. http://dx.doi.org/10.1109/CVPR52688.2022.00951.

Fu, D.Y., Dao, T., Saab, K.K., Thomas, A.W., Rudra, A., Ré, C., 2022. Hungry hungry hippos: Towards language modeling with state space models. In: Proceedings of the International Conference on Learning Representations. ICLR, http://dx.doi.org/10.48550/arXiv.2212.14052.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., van den Hengel, A., 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 1705–1714. http://dx.doi.org/10.1109/ICCV.2019.00179.

Gu, A., Dao, T., 2023. Mamba: Linear-time sequence modeling with selective state spaces. ArXiv Preprint. arXiv:2312.00752URL https://arxiv.org/abs/2312.00752.

Gu, A., Goel, K., Ré, C., 2021. Efficiently modeling long sequences with structured state spaces. In: Proceedings of the International Conference on Learning Representations. ICLR, http://dx.doi.org/10.48550/arXiv.2111.00396.

Haselmann, M., Gruber, D.P., Tabatabai, A.u., 2018. Anomaly detection using deep learning based image completion. In: 2018 17th IEEE International Conference on Machine Learning and Applications. ICMLA, IEEE, pp. 1237–1242. http://dx.doi.org/10.1109/ICMLA.2018.00201.

He, H., Zhang, J., Chen, H., Chen, X., Li, Z., Chen, X., Wang, Y., Wang, C., Xie, L., 2024a. A diffusion-based framework for multi-class anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, http://dx.doi.org/10.1609/aaai.v38i8.286.

He, H., et al., 2024b. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. ArXiv Preprint. arXiv:2404.06564. URL https://arxiv.org/abs/2404.06564.

Huang, C., Guan, H., Jiang, A., Zhang, Y., Spratling, M., Wang, Y.F., 2022. Registration based few-shot anomaly detection. In: Proceedings of the European Conference on Computer Vision. ECCV, Springer, pp. 303–319.

Huang, T., Pei, X., You, S., Wang, F., Qian, C., Xu, C., 2024. Localmamba: Visual state space model with windowed selective scan. ArXiv Preprint. arXiv:2403.09338. URL https://arxiv.org/abs/2403.09338.

Iqbal, N., Martinel, N., 2025. Pyramid-based mamba multi-class unsupervised anomaly detection. arXiv:2504.03442.

Li, C.-L., Sohn, K., Yoon, J., Pfister, T., 2021. Cutpaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9664–9674. http://dx.doi.org/10.1109/CVPR46437.2021.00954.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y., 2024. Vmamba: Visual state space model. ArXiv Preprint. arXiv:2401.10166 . URL https://arxiv.org/abs/2401.10166.

Liu, Z., Zhou, Y., Xu, Y., Wang, Z., 2023. Simplenet: A simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 11090–11099. http://dx.doi.org/10.1109/CVPR52729.2023.01954.

Lu, R., Wu, Y., Tian, L., Wang, D., Chen, B., Liu, X., Hu, R., 2023. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS).

Mehta, H., Gupta, A., Cutkosky, A., Neyshabur, B., 2022. Long range language modeling via gated state spaces. In: Proceedings of the International Conference on Learning Representations. ICLR, http://dx.doi.org/10.48550/arXiv.2206.13947.

Rasley, J., Rajbhandari, S., Ruwase, O., He, Y., 2020. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3505–3506.

Ristea, N.-C., Madan, N., Ionescu, R.T., Nasrollahi, K., Khan, F.S., Moeslund, T.B., Shah, M., 2022. Self-supervised predictive convolutional attentive block for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 13576–13586. http://dx.doi.org/10.1109/CVPR52688.2022.01321.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P., 2022. Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 14318–14328. http://dx.doi.org/10.1109/CVPR52688.2022.01392.

Ruan, J., Xiang, S., 2024. Vm-unet: Vision mamba unet for medical image segmentation. ArXiv Preprint. arXiv:2402.02491. URL https://arxiv.org/abs/2402.02491.

Rudolph, M., Wehrbein, T., Rosenhahn, B., Wandt, B., 2022. Fully convolutional cross-scale-flows for image-based defect detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, pp. 1088–1097. http://dx.doi.org/10.1109/WACV51458.2022.00189.

Shi, Y., Xia, B., Jin, X., Wang, X., Zhao, T., Xia, X., Xiao, X., Yang, W., 2024. Vmambair: Visual state space model for image restoration. ArXiv Preprint. arXiv:2403.11423. URL https://arxiv.org/abs/2403.11423.

Smith, J.T., Warrington, A., Linderman, S.W., 2022. Simplified state space layers for sequence modeling. In: Proceedings of the International Conference on Learning Representations. ICLR.

Wang, Z., Zheng, J.-Q., Zhang, Y., Cui, G., Li, L., 2024. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. ArXiv Preprint. arXiv:2402.05079. URL https://arxiv.org/abs/2402.05079.

Yan, X., Zhang, H., Xu, X., Hu, X., Heng, P.-A., 2021. Learning semantic context from normal samples for unsupervised anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, URL https://ojs.aaai.org/index.php/AAAI/article/view/16420.

You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., Le, X., 2022. A unified model for multi-class anomaly detection. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 29561–29574.

Zavrtanik, V., Kristan, M., Skočaj, D., 2021. Reconstruction by inpainting for visual anomaly detection. Pattern Recognit. 112, 107706. http://dx.doi.org/10.1016/j.patcog.2020.107706.

Zhang, X., Li, S., Li, X., Huang, P., Shan, J., Chen, T., 2023. DestSeg: Segmentation guided denoising student-teacher for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9761–9770. http://dx.doi.org/10.1109/CVPR52729.2023.00381.

Zhang, X., Xu, M., Zhou, X., 2024. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. arXiv preprint arXiv:2403.05897. URL https://arxiv.org/abs/2403.05897.

Zhao, Y., 2023. Omnial: A unified CNN framework for unsupervised anomaly localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, http://dx.doi.org/10.1109/CVPR52729.2023.00382.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X., 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. ArXiv Preprint. arXiv:2401.09417. URL https://arxiv.org/abs/2401.09417.

Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O., 2022. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 6–7. http://dx.doi.org/10.48550/arXiv.2207.14315.