Expert Systems With Applications 308 (2026) 131146

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Beyond feature mapping: Dual-heterogeneous knowledge distillation with mamba for industrial anomaly detection

Muhao Xu [a], Zihan Nie [a], Baochen Fu [a,b], Zhuangzhuang Chen [c], Zijian Li [a], Hua Wei [a], Yi Wan [a], Weiye Song [a,*]

[a] Department of Mechanical Engineering, Key Laboratory of High Efficiency and Clean Mechanical Manufacture of Ministry of Education, Shandong University, Jinan, 250061, China
[b] School of Software, Shandong University, Jinan, 250061, China
[c] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

ARTICLE INFO

Keywords:
Unsupervised anomaly detection
Surface defect detection
Knowledge distillation
Dual-heterogeneous

ABSTRACT

Unsupervised anomaly detection is essential for guaranteeing product quality and reliability in industrial manufacturing, yet existing knowledge distillation frameworks remain constrained by the feature mapping problem and by limited capacity in modeling both fine-grained local details. These limitations hinder robust detection of structural anomalies, as well as logical anomalies. To overcome these challenges, we present Beyond Feature Mapping, a Dual-Heterogeneous Knowledge Distillation network enhanced with Mamba based sequence modeling. The framework employs a teacher encoder and a student network consisting of an encoder with two heterogeneous decoders: a convolutional branch with Spatial Channel Mixer Modulation to enhance local detail interaction, and a Mamba based branch with a Hybrid Self-Task Mamba Module to model long range dependencies. This dual-decoder design alleviates feature mapping problems while enabling robust detection of both structural and logical anomalies. Extensive experiments on three public benchmarks, MVTec LOCO, VisA, and BTAD, demonstrate that DHKD consistently outperforms recent state-of-the-art methods in both detection accuracy and localization robustness. Our code is available at: https://github.com/Gilbert-Et/DHKD

1. Introduction

Unsupervised anomaly detection (UAD) plays a critical role in industrial manufacturing, where ensuring product quality and production efficiency is essential (Hilal et al., 2022; Liu et al., 2023a; Ma et al., 2025; Zhou et al., 2022). Unlike supervised approaches, UAD does not rely on defect annotations but instead leverages normal data to model the distribution of standard patterns. During inference, deviations from this learned distribution are regarded as anomalies (Salehi et al., 2021). This property makes UAD particularly suitable for industrial settings, where collecting large-scale defect labels is costly, time-consuming, and often infeasible for rare or novel defects.

Unsupervised anomaly detection methods are primarily divided into three categories: reconstruction-based methods (Schlegl et al., 2019; Yan et al., 2021), representation-based methods (Guo et al., 2019; Reiss et al., 2021; Wu et al., 2023), and knowledge distillation-based approaches (Deng & Li, 2022; Guo et al., 2023). Reconstruction-based unsupervised anomaly detection methods train a generative model, such as an autoencoder (AE)(An & Cho, 2015) or a variational autoencoder (VAE)(Bergmann et al., 2018), to learn the feature distribution of normal samples. During the training phase, the model learns to encode normal samples into low-dimensional representations and reconstruct normal samples from these representations. In the testing phase, when an anomalous sample is input into the model, significant differences in its features compared to normal samples prevent the model from effectively reconstructing it, resulting in a high reconstruction error. The model identifies anomalous regions by assessing the error between the input and reconstructed images. While these methods are effective at detecting structural anomalies to a certain degree, their performance often declines when facing logical anomalies or complex backgrounds. The different kinds of anomalies are shown in the Fig. 1. Representation-based methods (Wu et al., 2021) typically use pre-trained networks to extract high-dimensional feature representations of images and train a representation model on normal samples to learn their feature distribution. In the testing phase, the model identifies anomalous samples by comparing the input's feature representation to those of normal samples

* Corresponding author.
E-mail addresses: ujnmhxu@hotmail.com (M. Xu), 202414371@mail.sdu.edu.cn (Z. Nie), fbc@mail.sdu.edu.cn (B. Fu), eezzchen@ust.hk (Z. Chen), 202320663@mail.sdu.edu.cn (Z. Li), weihua@sdu.edu.cn (H. Wei), wanyi@sdu.edu.cn (Y. Wan), songweiye@sdu.edu.cn (W. Song).

https://doi.org/10.1016/j.eswa.2026.131146
Received 2 September 2025; Received in revised form 30 December 2025; Accepted 5 January 2026
Available online 8 January 2026

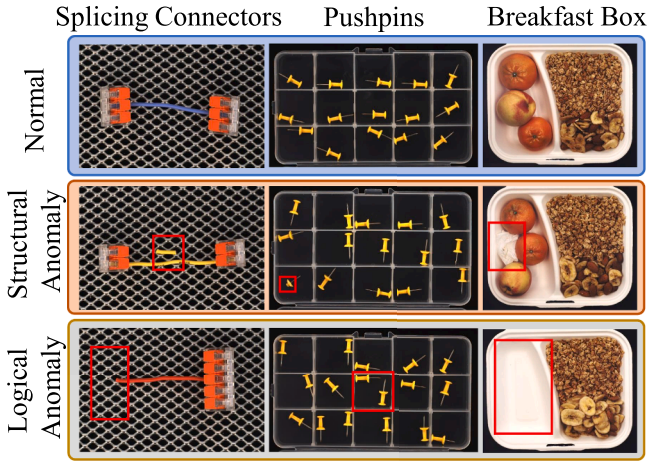

Fig. 1. Examples of samples of different types of anomalies, including normal images (top), structural anomalies (middle), and logical anomalies (bottom). Structural anomalies are those that introduce new local structures (e.g., new connectors, bad pushpins, and toilet paper), while logical anomalies violate the logical constraints of the normal image (e.g., missing connectors, extra pushpins, and missing oranges and nectarines).

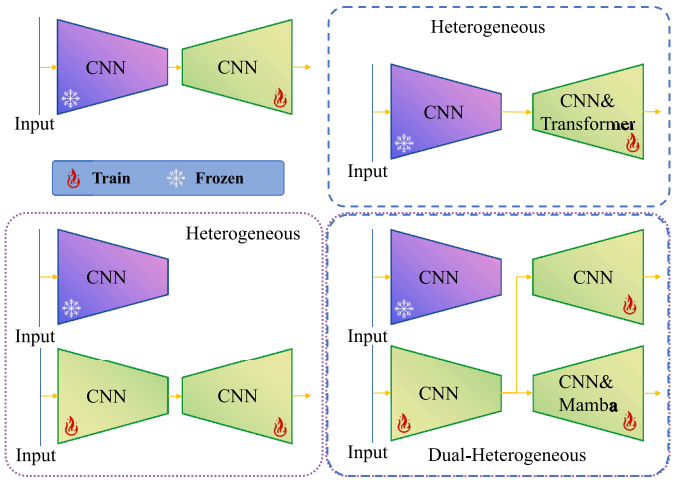

Fig. 2. Comparison of strategies: (a) Homogeneous–frozen the teacher encoder (blue-purple) CNN and train the student decoder (green) CNN; (b) Heterogeneous, top-right–frozen the teacher encoder CNN, then train the student decoder CNN & Transformer. bottom-left–frozen the teacher encoder CNN, then train the student encoder-decoder CNN; (c) Dual-heterogeneous–frozen the teacher encoder CNN, then train the student shared encoder and two different decoders CNN + CNN & Mamba.

from training, detecting deviations in the feature space. These methods are advantageous for handling various types of anomalies, especially when anomalies involve complex semantics or structures, facilitating more accurate anomaly region detection. However, in complex environments where both logical and structural anomalies are present, the feature differences between anomalous and normal samples may be negligible, leading to a decrease in detection accuracy.

In recent years, knowledge distillation (Wu et al., 2024) has emerged as one of the most effective paradigms for unsupervised anomaly detection (UAD). In comparison with reconstruction- and representation-based methods, knowledge distillation (Deng & Li, 2022) employs a teacher-student framework in which the teacher network, pre-trained on normal data, provides rich and reliable feature representations, while the student learns to approximate these representations. During the process of inference, discrepancies between teacher and student features naturally highlight anomalous regions. Recent studies have further explored reverse knowledge distillation (RKD) frameworks, where a compact student model guides the formation of a teacher through an encoder–decoder residual architecture (Lei et al., 2024).This design eliminates the need for explicit defect annotations, ensures stable training, and has led to state-of-the-art results across several industrial benchmarks, making knowledge distillation the mainstream solution for UAD. Notwithstanding the success of existing knowledge distillation approaches, two critical challenges remain to be addressed. Firstly, the eminent feature mapping problem manifests due to students frequently learning to replicate not only conventional patterns but also anomalous ones. This results in a diminution of the discriminability between normal and abnormal regions. Secondly, although introducing architectural heterogeneity between teacher and student partially mitigates this limitation, most existing student designs rely on a single heterogeneous decoder, which is inherently unable to capture fine-grained local details and long-range semantic dependencies simultaneously. As a consequence, the student tends to prioritise the dominant representational mode of the training data and neglect the weaker one, leading to the suppression of anomaly-relevant residuals either in the local space or in the global space. This limitation is detrimental to the robust detection of both structural anomalies and logical anomalies, which require distinct and complementary representational strengths.

To overcome these challenges, we propose a Dual-Heterogeneous Knowledge Distillation Network (DHKD) with Mamba for industrial anomaly detection, as illustrated in Fig. 2. The framework extends the classical teacher-student paradigm by introducing heterogeneity not only between the teacher and student but also within the student architecture itself. The teacher is a frozen encoder trained exclusively on normal samples, providing a stable reference of defect-free representations. In contrast, the student comprises a shared encoder followed by two heterogeneous decoders with fundamentally different inductive biases and representational capacities. This structural heterogeneity prevents the student from perfectly aligning with the teacher in all representational subspaces, thus preserving subtle residual discrepancies between normal and abnormal patterns and improving anomaly separability. The first decoder is a convolutional branch enhanced with the proposed Spatial Channel Mixer Modulation (SCMM), which performs non-local spatial-channel mixing to strengthen the reconstruction of fine-grained details and sharpen sensitivity to structural anomalies. The second decoder adopts a sequence-modelling perspective via the Hybrid Self-Task Mamba Module (HSTM). Built upon the recently introduced Mamba architecture, HSTM efficiently captures long-range dependencies with linear time complexity and, through a self-task mechanism, enforces semantic coherence across distant regions in the feature map. This enables the model to detect logical anomalies such as missing or swapped components that cannot be inferred solely from local appearance cues. By jointly exploiting SCMM for high-fidelity local reconstruction and HSTM for global semantic consistency, the dual-heterogeneous student architecture preserves complementary residuals across multiple representational levels rather than collapsing onto a single dominant feature mode. Consequently, DHKD can robustly differentiate normal and abnormal regions across diverse anomaly types. Extensive experiments on three widely used industrial benchmarks–MVTec LOCO, VisA, and BTAD–demonstrate that DHKD consistently outperforms recent state-of-the-art UAD approaches in both anomaly detection and localisation.

Our contributions can be summarized as follows:

- We propose an image anomaly detection model based on a dual-heterogeneous knowledge distillation network, which effectively prevents the feature mapping problem and enhances the model's anomaly detection performance.
- The Hybrid Self-Task Mamba Module (HSTM) is introduced to capture remote spatial dependencies effectively, boosting the model's accuracy and robustness in global anomaly detection.

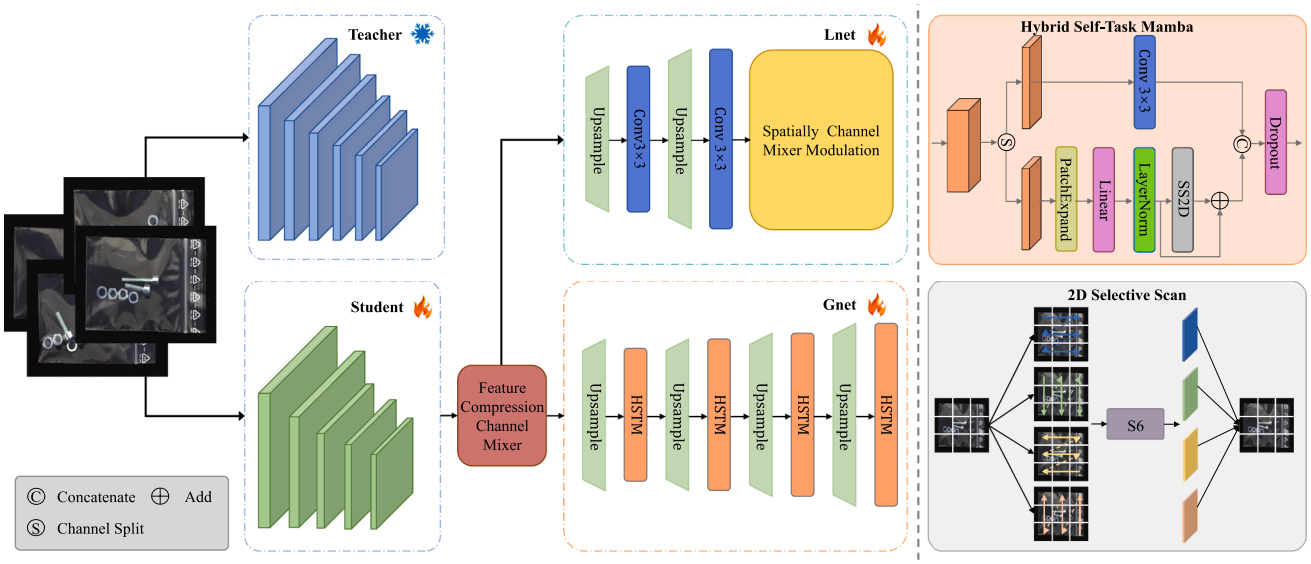

Fig. 3. Overview architecture of the proposed framework.

- The Spatial Channel Mixer Modulation (SCMM) was developed to enable non-local feature interaction, improving the model's sensitivity and precision in local anomaly detection.

2. Related works

2.1. Feature representation-based methods

Feature representation-based methods have shown significant effectiveness in both anomaly detection and localization by extracting representations (i.e., feature vectors or patches) of images from pre-trained networks and detecting anomalies based on the distance between the test image representation and those of normal cases. Reiss et al. (2021) proposed a method that leverages a pre-trained network to extract feature representations and model their distribution. During inference, anomalies are detected by measuring the deviation between the sample's feature representation and the learned distribution of normal samples. However, this approach is highly dependent on the quality of the pre-trained network and may struggle with subtle anomalies that show minimal deviation from the normal distribution. Additionally, it does not explicitly address the challenge of feature compression, which may affect the detection accuracy when dealing with complex or diverse anomaly types.

PaDiM (Defard et al., 2021) enhances this approach by storing multivariate Gaussian distributions of feature representations and calculating the Mahalanobis distance between test samples and the stored normal distributions to detect anomalies. While the use of a statistical distance metric improves performance, its reliance on the Gaussian assumption limits its effectiveness for non-Gaussian data distributions. Moreover, storing parameters for each feature map can increase memory consumption, making the method less efficient for large-scale applications. The Metaformer method presented by Wu et al. (2021) addresses the issues of model adaptation and reconstruction gap by leveraging an unsupervised universal model that is meta-learned (Finn et al., 2017)to achieve high model adaptation capability and instance-aware attention to emphasize focal regions for localizing abnormal regions. Patch Core (Roth et al., 2022) introduces an optimized patch description method using a greedy algorithm to streamline the normal feature library, selectively retaining the most representative features to balance detection accuracy and computational efficiency. However, Patch Core may encounter challenges in detecting anomalies appearing in sparsely sampled or low-frequency regions of the feature space due to the potential exclusion of subtle but critical features.

2.2. The reconstruction-based methods

Reconstruction-based methods are extensively explored in anomaly detection due to their assumption that models trained exclusively with normal images cannot effectively reconstruct anomalous ones. An and Cho (2015), Guo et al. (2023) demonstrated such models' ability to minimize reconstruction error for normal images while struggling with abnormal data during testing. Commonly, Generative Adversarial Networks (GANs) (Schlegl et al., 2019) and Variational Auto-Encoders (VAEs) (Bergmann et al., 2018) are utilized to learn and reproduce normal patterns, with deviations indicating anomalies. Challenges arise when anomalies share characteristics with normal data, leading models to inadvertently reconstruct them and fail to detect abnormalities effectively. SCADN (Yan et al., 2021) incorporates inpainting frameworks to utilize contextual information for reconstructing masked normal regions, aiding in anomaly detection by evaluating reconstructed quality. However, the generalization power of deep learning models can cause them to reproduce subtle anomalies, leading to diminished anomaly visibility and reduced detection accuracy.

2.3. Distillation-based methods

Distillation-based methods leverage a teacher-student framework where a pre-trained teacher network transfers knowledge to a student network, which then mimics the teacher's behavior. Anomalies are detected based on discrepancies between the outputs of the two networks. Bergmann et al. (2020) used this framework to detect anomalies by leveraging the intrinsic uncertainty of the student networks as an anomaly scoring function, but this approach may struggle with subtle or complex anomalies where the output discrepancies are insufficient for detection. In an effort to further refine anomaly detection, Salehi et al. (2021) introduced a multiresolution knowledge distillation paradigm that capitalizes on the disparities in intermediate activation values between a pre-trained expert network and a less complex cloner network, consequently augmenting the discriminative capacity between normal and anomalous data, which ameliorates the efficacy of anomaly detection. Nonetheless, the methodology may encounter challenges in the transference of knowledge to the cloner network in scenarios where the test data exhibits substantial divergence from the training data

distribution, potentially compromising its generalizability. Reverse distillation, as demonstrated in RD (Deng & Li, 2022), introduces an encoder-decoder structure to improve performance by combining the distillation framework with reconstruction principles. While effective, encoder-decoder designs still face challenges related to feature compression and loss of detailed information, potentially impacting the detection of finely detailed anomalies. MHKD (Xu et al., 2025b) proposes a multitask learning strategy combined with hybrid distillation loss to enhance the student's representation ability for anomaly detection. However, it employs a homogeneous student architecture, without explicitly modeling the complementary characteristics of different anomaly types. MFKD (Lu et al., 2025) for Improved Anomaly Detection incorporates auxiliary textual information during training and distills multimodal knowledge into a single image-based model for inference. Although effective, its framework focuses primarily on multimodal feature complementarity rather than heterogeneous decoder design within a unimodal setting. KD-LightMAD (Yousefimehr et al., 2025) introduces a lightweight multimodal distillation framework designed for practical industrial deployment and resource efficiency. Lei et al. (Lei et al., 2024) propose a reverse knowledge-distillation framework in which the student first extracts features that are subsequently used to construct the teacher. This inversion prevents direct teacher-to-student feature regression but still employs a single architectural pathway for reconstruction.

Several recent detection works have advanced contextual modeling and anomaly localization, such as the confluent triple-flow network for detection (Tang et al., 2024) and attention-guided pyramidal representations for few-shot detection (Tang et al., 2022). Additionally, asymmetric distillation strategies have been explored for improving detection performance in anomaly and object localization tasks (Xing et al., 2024). To mitigate these issues, THFR (Guo et al., 2023) proposed a template-based reconstruction method that references normal images to guide the recovery process and improve reconstruction detail. While this approach enhances anomaly detection by aiding the recovery of abnormal images, it also poses a drawback: differences between the normal template and input image can degrade the reconstruction quality of normal images, reducing detection accuracy.

3. Proposed method

To address the challenges of anomaly detection in complex scenarios, we propose a Dual-Heterogeneous Knowledge Distillation Network, as depicted in Fig. 3.

During the training phase, the teacher network $F_T(\cdot)$ supervises the learning process of the student network components, including the encoder $F_S(\cdot)$, the Feature Compression Channel Mixer $M(\cdot)$, the Logical Decoder Branch $Gnet(\cdot)$, and the Structural Decoder Branch $Lnet(\cdot)$, by extracting features from normal images x. In the testing phase, structural anomaly score maps S_l are calculated by comparing the feature representations f_T produced by the teacher network with those generated by the Structural Decoder Branch f_L. Concurrently, logical anomaly score maps S_g are obtained by evaluating the consistency between the outputs f_G of the Logical Decoder Branch $G(\cdot)$ and the outputs f_S of the student network's encoder $F_S(\cdot)$. Finally, the structural and logical anomaly scores are combined to produce the final anomaly score S, providing a robust and comprehensive evaluation of anomalies.

3.1. Teacher network

The teacher network $F_T(\cdot)$ is composed of six convolutional layers and two average pooling layers. This architecture is inspired by the Efficient Anomaly Detection (AD) framework, where each neuron in the output layer of $F_T(\cdot)$ corresponds to a receptive field of 33×33 pixels. The network is obtained via the distillation of a pre-trained WideResNet-101 model (He et al., 2016) using the ImageNet dataset. During this process, the mean square error (MSE) loss function is utilized to refine the weights in $F_T(\cdot)$. Moreover, the feature post-processing strategy from PatchCore (Roth et al., 2022) is incorporated. The method combines features from two different network layers within the teacher's network, transforming them into a 384-dimensional compressed representation to improve feature expressiveness. The patch description network (PDN) offers superior computational efficiency compared to the direct utilization of WideResNet-101. Its architecture is specifically designed to confine anomaly detections to their respective localized regions, preventing unintended interactions with unrelated areas of the image.

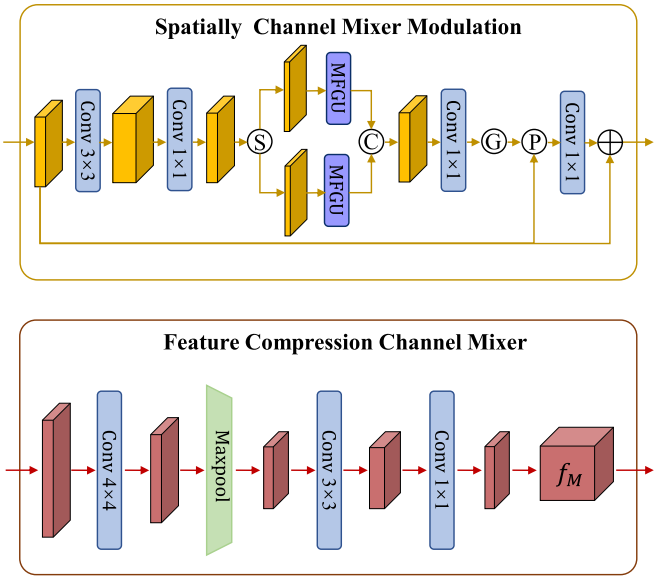

Fig. 4. Overview architecture of spatial channel mixer modulation and feature compression channel mixer.

The features $f_T \in \mathbb{R}^{h \times w \times c}$ are obtained by feeding the training images x into the $F_T(\cdot)$, as defined by:

$$f_T = F_T(x). \tag{1}$$

3.2. Student network

The proposed student network consists of four primary components: the student encoder network $F_S(\cdot)$, the Feature Compression Channel Mixer module(FCCM), the structural anomaly detection branch $Lnet(\cdot)$, and the logical anomaly detection branch $Gnet(\cdot)$. The architecture of the student encoder network includes six convolutional layers and two max-pooling layers. In comparison to the teacher encoder network, the student encoder adopts max-pooling instead of average pooling and features different convolutional parameters. This asymmetric design is intended to enhance the network's sensitivity to anomalous images.

During training, normal sample images x were processed by the student encoder $F_S(\cdot)$ to produce shallow features f_S. These features were then passed to the Feature Compression Channel Mixer (FCCM) to obtain the compressed feature f_M, as illustrated in Fig. 4. FCCM is intentionally organised as a four-stage pipeline designed to (i) remove spatial redundancy, (ii) aggregate robust structural cues, and (iii) produce a compact channel-wise representation suitable for subsequent decoding. Concretely, 4×4 convolution is applied first to achieve early spatial compression and to enlarge the effective receptive field with modest depth; this choice yields coarse structural descriptors while keeping computation efficient compared with stacking many small kernels. The MaxPool layer follows to aggregate information over a wider neighbourhood, which suppresses isolated noise and stabilises dominant activations, thereby improving robustness to local perturbations. The 3×3 convolution then refines the pooled features, restoring spatial specificity and enabling non-linear combination of neighbouring contexts; this stage may also expand channel dimensionality to increase

Table 1
Quantitative results on MVTec LOCO AD dataset for anomaly localization, as measured on pixel-sPRO [%]. The best results are marked in bold.

Method	Breakfast Box	Screw Bag	Pushpins	Splicing Connectors	Juice Bottle	Mean
VM (Steger et al., 2018)	16.8	25.3	25.4	12.5	32.5	22.5
f-AnoGAN (Schlegl et al., 2019)	22.3	34.8	33.6	19.5	56.9	33.4
MNAD (Park et al., 2020)	8.0	34.4	35.7	44.2	47.2	33.9
AE (An & Cho, 2015)	18.9	28.9	32.7	47.9	60.5	37.8
VAE (Bergmann et al., 2018)	16.5	30.2	31.1	49.6	63.6	38.2
SPADE (Cohen & Hoshen, 2020)	37.2	33.1	23.4	51.6	80.4	45.1
S-T (Bergmann et al., 2020)	49.6	60.2	52.3	69.8	81.1	62.6
RD (Deng & Li, 2022)	56.0	53.5	57.7	70.1	83.7	64.2
PaDiM (Defard et al., 2021)	50.9	46.1	29.5	46.7	77.9	50.2
Patch Core (Roth et al., 2022)	45.1	56.2	42.3	59.8	69.4	54.6
GCAD (Bergmann et al., 2022)	50.2	55.8	73.9	79.8	91.0	70.1
DRAEM (Zavrtanik et al., 2021)	49.9	49.0	49.3	67.3	80.0	59.1
DSKD (Zhang et al., 2024)	56.8	62.7	82.5	76.7	86.5	73.0
THFR(Guo et al., 2023)	58.3	61.5	76.3	84.8	89.6	74.1
NPGMF (Xu et al., 2025a)	64.9	66.6	75.5	82.2	89.6	75.8
Efficient AD (Batzner et al., 2024)	–	–	–	–	–	79.8
DHKD	**69.1**±0.1	**69.8**±0.3	**90.8**± 0.2	**90.3**±0.1	**96.5**±0.1	**83.3**±0.1

representational capacity for ensuing mixing. Finally, 1×1 convolution performs point-wise channel mixing and dimensional alignment, producing f_M at the target decoder dimensionality. The arranged sequence was selected to balance robustness, locality and computational cost, and to preserve anomaly-relevant signals while avoiding over-sensitivity to high-frequency noise.

To effectively separate the intricate tasks of structural and logical anomaly detection, a two-way decoder framework is meticulously designed. Within this framework, the Lnet branch is dedicated to detecting structural anomalies. Conversely, the Gnet branch targets logical anomaly detection by extracting and analyzing global image features to identify deviations from standard image composition and consistency.

The LDecoder consists of upsampling stages, convolutional blocks, and the Spatial Channel Mixer Modulation (SCMM) unit. Its architecture is shown in Fig. 4. The SCMM output f_M was upsampled twice and convolved to yield $\hat{f}_M$, which was then expanded by a 3×3 convolution that doubled the channel dimensionality and projected back with a 1×1 convolution; this temporary expansion facilitates richer cross-channel interactions without permanently increasing the decoder footprint. The resulting tensor was split channel-wise into two groups and routed into a Multiscale Feature Generation Unit (MFGU). The spatial branch applied an efficient depth-wise 3×3 convolution to emphasise local structure and high-frequency detail. In contrast, the context branch performed pooling-based summarisation to capture coarser, non-local context. An adaptive max-pooling operation was applied to the context branch to select discriminative activations and attenuate irrelevant responses. The two branch outputs were fused (via learnt multiplicative modulation and additive residual connection) to produce a modulation map, which reweighted $\hat{f}_M$ and injected spatially selective, channel-aware contextual priors into the decoder. In aggregate, SCMM implements disentangled spatial and channel processing followed by guided modulation: this yields non-localised contextual cues without sacrificing local reconstruction fidelity, and does so with modest additional computational cost due to the use of depth-wise operations and point-wise mixing. Given the input features $\hat{f_M}$, the process can be formulated as:

$$
\begin{aligned}
&\hat{f_{M'}} = Conv_{1\times1}(Conv_{3\times3}(\hat{f_M})),\\
&[f_{M0}, f_{M1}] = Split(\hat{f_{M'}}),\\
&\hat{f_{M0}} = DW - Conv_{3\times3}(f_{M0}),\\
&\hat{f_{M1}} = \uparrow_p (DW - Conv_{3\times3}(\downarrow_{\frac{p}{2}} (f_{M1}))),
\end{aligned} \tag{2}
$$

where $Split(\cdot)$ refers to the channel splitting operation, $DW - Conv_{3\times3}(\cdot)$ denotes a 3×3 depth-wise convolution, $\uparrow_p (\cdot)$ denotes upsampling features to the original resolution p using nearest interpolation, while $\downarrow_{\frac{p}{2}} (\cdot)$ indicates pooling the input features to a size of $\frac{p}{2}$.

After concatenating the extracted features across channels, use a 1 ± 1 convolution to fuse them. Then, perform element-wise multiplication with the estimated attention map to adaptively recalibrate the inputs. Finally, the final features are output after 1 ± 1 convolution and residual concatenation. This process can be written as:

$$
\begin{aligned}
&f_M{}' = Conv_{1\times1}(Concat[\hat{f_{M0}}, \hat{f_{M1}}]),\\
&\bar{f_M} = Conv_{1\times1}(\hat{f_M} \odot f_M{}'),\\
&f_L = \bar{f_M} + \hat{f_M},
\end{aligned} \tag{3}
$$

where $\odot$ is the element-wise product.

The Gnet architecture is designed to refine input features f_M and generate output features f_G through a progressive multi-scale feature refinement pipeline. Initially, the input feature map f_M is transformed using a 3×3 convolution to produce the initial feature $f_M{}^1$, which serves as the starting point for the refinement process. Subsequently, the feature map is iteratively enhanced over three stages, where each stage consists of the Hybrid Self-Task Mamba (HSTM) module and an upsampling operation to progressively increase spatial resolution. This iterative process allows the architecture to gradually refine the features while incorporating multi-scale information. Mathematically, the initial step and iterative refinement are described as follows:

$$
\begin{aligned}
&f_M{}^1 = Conv_{3\times3}(f_M),\\
&f_M{}^{k+1} = \uparrow p\big(HSTM(f_M{}^k)\big), \quad k \in \{1,2,3,4\}
\end{aligned} \tag{4}
$$

The Hybrid Self-Task Mamba Module (HSTM) plays a central role in this architecture, which integrates long-range dependency modelling and contextual aggregation. As a result, f_G serves as a contextually strengthened and anomaly-suppressed target that reflects global structural priors unavailable to the encoder alone. This forms an auxiliary task created by the student themselves, enabling the encoder to internalise global semantics and reducing its tendency to collapse towards purely local or dominant-mode representations. The Mamba branch captures global dependencies using Patch Expansion, Linear Transformation, LayerNorm, and the Structured 2D-Selective-Scan Module (SS2D). The convolutional branch enhances local spatial features using 3×3 convolutions. The outputs of the two branches are fused and normalized to produce refined features. The HSTM output is formulated as follows:

$$
\begin{aligned}
&[f_{M1}{}^k, f_{M2}{}^k] = Split(f_M{}^k),\\
&f_{li} = Linear(PatchExpand(f_{M1}{}^k)),\\
&f_{ss2d} = SS2D(LayerNorm(f_{li})),\\
&f_{Mamba} = f_{ss2d} + f_{li},\\
&f_C = Conv_{3\times3}(f_{M2}{}^k),
\end{aligned}
$$

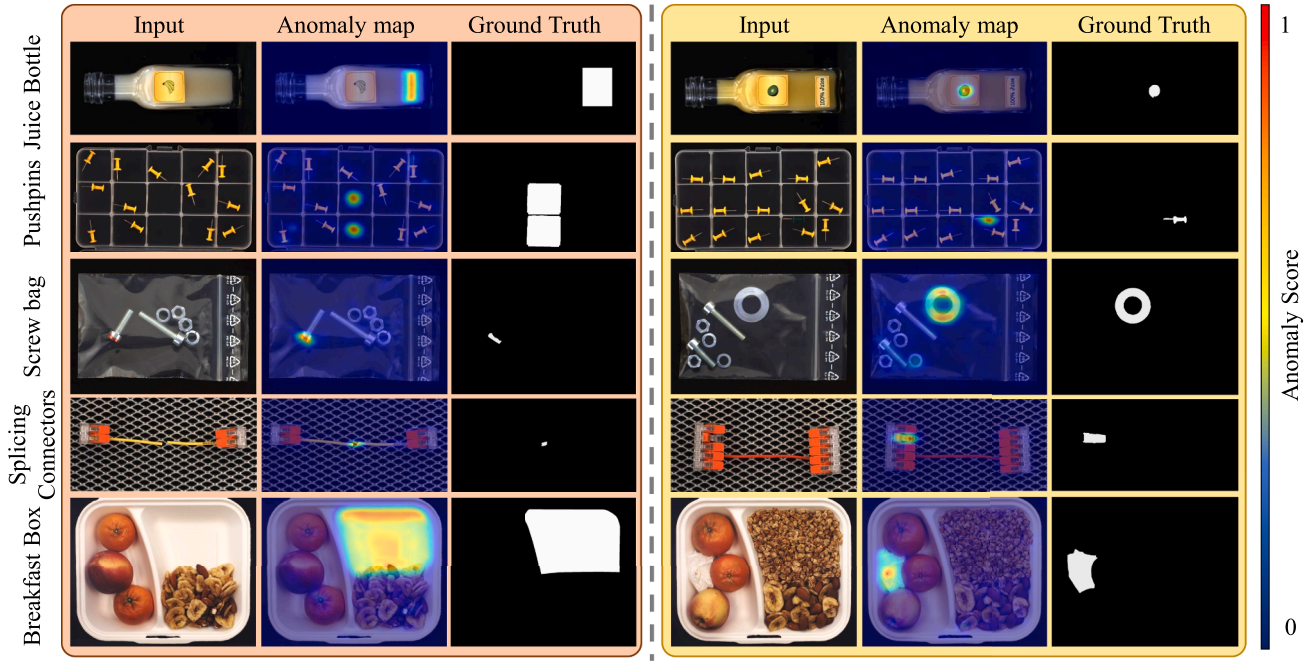

Fig. 5. Visualization of anomaly localization results on different categories of MVTec LOCO AD dataset.

$$f_M{}^{k+1} = Dropout\big(concat\big(f_{Mamba}, f_C\big)\big), \tag{5}$$

After three iterative refinement steps, the final feature $f_M{}^4$ undergoes a concluding 3×3 convolution to produce the output feature f_G. This architecture effectively combines global and local feature enhancements, progressively increasing spatial resolution and multi-scale feature representation.

The loss functions for the model components are defined as follows: The loss for $Gnet(\cdot)$ is $\mathcal{L}_G = (hwc)^{-1} \sum_c \|f_T - f_G\|_F^2$. The loss for $Lnet(\cdot)$ is $\mathcal{L}_L = (hwc)^{-1} \sum_c \|f_T - f_L\|_F^2$, where h, w,and c represent the height, width, and number of channels. Additionally, f_G is used as supervisory information to guide the output feature f_S of $F_S(\cdot)$, with the corresponding loss function $\mathcal{L}_D$ defined as $\mathcal{L}_D = (hwc)^{-1} \sum_c \|f_G - f_S\|_F^2$. The total loss $\mathcal{L}_{total}$ of the model defined as

$$\mathcal{L}_{total} = \alpha \mathcal{L}_L + \beta \mathcal{L}_G + \gamma \mathcal{L}_D. \tag{6}$$

where α, β, and γ are used to adjust the relative contributions of the three loss terms.

3.3. Anomaly map

In the testing phase, the logical anomaly score S_g is calculated as the difference between the student encoder network $F_S(\cdot)$ and $Gnet(\cdot)$. The structural anomaly score S_l is obtained by comparing the output feature f_s of $Lnet(\cdot)$ with the feature f_t from the teacher network F_T.

To align the scales of both scores, they are normalized to prevent noise interference. Following Efficient AD, pixel anomaly scores are computed for both maps, and p-quantiles are calculated for each. Linear transformations are applied, and the scores are normalized accordingly. The normalized scores are then interpolated to the original image size and summed to yield the overall anomaly score S, as shown below:

$$S = \varphi(S_l) + \varphi(S_g), \tag{7}$$

where φ represents the linear interpolation.

4. Experiments

4.1. Dataset and experimental setup

MVTec LOCO AD dataset (Bergmann et al., 2022), developed by MVTec in Germany, is purpose-built for unsupervised anomaly detection. It serves as a comprehensive benchmark for evaluating advanced anomaly detection techniques. Widely recognized as one of the most demanding datasets in this domain, it replicates practical industrial inspection scenarios using data derived directly from manufacturing processes. Structural irregularities, such as scratches, dents, and contamination, frequently occur in industrial environments. In contrast, logical anomalies refer to misplaced items or the appearance of valid components in unsuitable positions. The dataset encompasses five genuine sub-datasets, including 1772 normal images for training, 304 for validation, and 1568 for testing. The test set comprises samples illustrating normal scenarios, structural anomalies, and logical inconsistencies.

VisA dataset(Zou et al., 2022) consists of 12 subsets that represent a diverse range of objects and challenging structural features. These subsets include complex printed circuit boards (PCBs) and those with significant variability in object pose and location, such as capsules, candles and macaroni. The dataset consists of 10,821 images, including 9621 normal samples and 1200 anomalous samples. The anomalies encompass both surface defects, such as scratches, dents, stains, and cracks, as well as structural defects, including misaligned or missing components.

BeanTech AD dataset (BTAD) (Mishra et al., 2021) includes three distinct categories of industrial products, each exhibiting both body and surface defects. It offers comprehensive labeling at both the image and pixel levels, allowing for in-depth defect analysis. The training sets for each category consist of 400, 399, and 1000 normal images, while the corresponding test sets contain 21/49, 30/200, and 410/31 normal and abnormal images. The BeanTech AD dataset serves as a crucial resource for developing and testing advanced automated methods for defect detection in industrial environments.

Training details. The proposed method is implemented using the PyTorch framework and trained from scratch on a server equipped with an NVIDIA A100 GPU. Input images are uniformly resized to 256×256

Table 2
Quantitative results on VisA dataset for anomaly detection/localization, as measured on image-AUROC/pixel-AUROC[%].

Category		PatchCore (Roth et al., 2022)	RD (Deng & Li, 2022)	DMAD (Liu et al., 2023b)	SimpleNet (Zavrtanik et al., 2021)	DRAEM (Liu et al., 2023c)	FastFLow (Yu et al., 2021)	Efficient AD (Batzner et al., 2024)	DHKD
VisA	candles	98.6/99.5	92.2/97.9	92.7/98.1	98.7/97.7	94.4/97.3	92.8/94.9	97.4/98.6	98.7/98.8
	capsules	81.6/99.5	90.1/89.5	88.0/99.2	89.9/99.0	76.3/99.1	71.2/75.3	92.3/99.4	94.1/99.6
	cashew	97.3/98.9	99.6/95.8	95.0/95.3	97.5/98.8	90.7/88.2	91.0/91.4	97.8/99.1	97.2/99.3
	chewinggum	99.1/99.1	99.7/99.0	97.4/97.9	99.8/98.3	94.2/97.1	91.4/98.6	98.9/98.6	99.0/98.6
	fryum	96.2/93.8	96.6/94.3	98.0/97.0	98.1/91.1	97.4/92.7	88.6/97.3	97.1/97.3	97.2/97.4
	macaroni1	97.5/99.8	98.4/97.7	94.3/99.7	99.4/99.6	95.0/99.7	98.3/97.3	97.9/99.8	98.2/99.9
	macaroni2	78.1/99.1	97.6/87.7	90.4/99.7	82.4/98.9	96.2/99.9	86.3/89.2	96.6/99.6	96.7/99.9
	pcb1	98.5/99.9	97.6/75.0	95.8/99.8	99.0/99.6	54.8/90.5	77.4/75.2	99.3/99.9	99.7/99.9
	pcb2	97.3/99.0	91.1/64.8	96.9/99.0	99.1/98.3	77.8/90.5	61.9/67.3	99.6/99.1	99.9/99.2
	pcb3	97.9/99.2	95.5/95.5	98.3/99.3	98.5/99.2	94.5/98.6	74.3/94.8	98.8/99.0	98.7/99.2
	pcb4	99.6/98.6	96.5/92.8	99.7/98.8	99.6/93.9	93.4/88.0	80.9/89.9	99.5/98.5	99.8/98.8
	pipe_fryum	99.8/99.1	97.0/92.0	99.0/99.3	99.7/98.9	99.4/90.9	72.0/87.3	99.6/99.5	100/99.6
	Mean	95.1/98.8	96.0/90.1	95.5/98.6	96.8/97.8	88.7/94.4	82.2/88.2	97.9/98.0	98.3/99.2

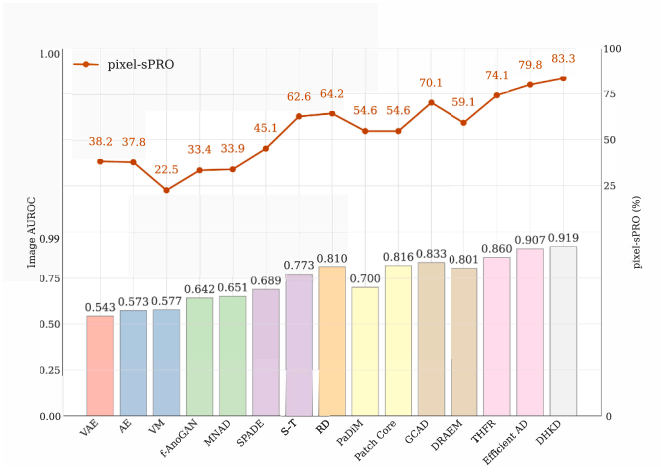

Fig. 6. Quantitative results on MVTec LOCO AD dataset for anomaly detection, as measured on the average image-AUROC and pixel-sPRO.

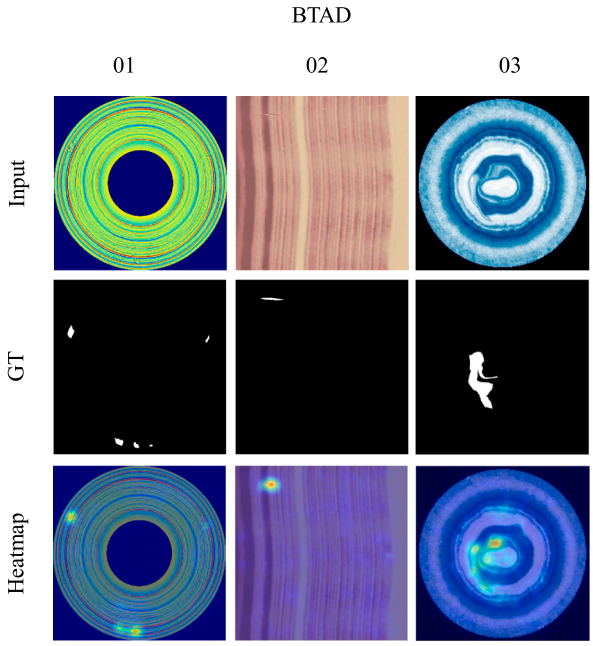

Fig. 7. Visualization of anomaly localization results on different categories of BTAD dataset.

pixels. Following prior research, one-model-per-category approach is adopted. Each student network is trained for 100,000 epochs with a batch size of 1. The Adam optimizer is used with an initial learning rate of 10^{-4} and weight decay of 10^{-5}, with the learning rate reduced to 10^{-5} after 76,000 epochs. The weights α,β, and γ are set in a ratio of 1:1:2, respectively.During training, a dropout rate of 0.2 is applied. The anomaly score map S is rescaled back to the original image resolution using bilinear interpolation. To ensure statistical reliability, all experiments were repeated three times using different random seeds, and we report the mean values.

Evaluation metrics. The AUROC, measuring an algorithm's ability to distinguish normal from anomalous samples, serves as the primary metric for threshold-free image-level anomaly detection. For anomaly localization, AUROC remains effective in assessing structural anomaly detection. However, logical anomalies, such as missing objects, are difficult to annotate and segment on a per-pixel basis precisely. To address these challenges, localization performance is evaluated using the saturated per-region overlap (sPRO) metric (Bergmann et al., 2022), an enhancement of the PRO metric.

4.2. Anomaly detection and localization

Results on MVTec LOCO AD We evaluated our proposed method against state-of-the-art techniques for anomaly detection/localization using the MVTec LOCO AD dataset. Quantitative results, summarized in Table 1 and Fig. 6 demonstrates the superior performance of our approach, with comparison data extracted from the original papers of the respective methods. As shown in Table 1, distillation-based methods, including GCAD (Bergmann et al., 2022), THFR (Guo et al., 2023), and Efficient AD (Batzner et al., 2024), consistently outperform representation- and reconstruction-based approaches in detecting structural and logical anomalies. Our method achieves the highest performance, with a mean pixel-level sPRO score of 83.3% and an image-level AUROC of 91.9%. Our method offers significant advantages in addressing both structural and logical anomalies across diverse and challenging scenarios. By leveraging a dual-heterogeneous knowledge distillation framework, the model effectively resolves the feature mapping issue, enhancing generalization and enabling precise anomaly localization. The integration of the Mamba module facilitates the detection of global dependencies, allowing the model to identify logical inconsistencies in complex scenes, such as the "breakfast box" category. Additionally, the dual decoders collaborate to capture both local and global anomaly features, substantially reducing false positives in challenging categories like "screw bag" and "pushpins," where irregular arrangements often hinder detection. The Multi-scale Feature Generation Unit further strengthens the model's ability to reconstruct fine-grained details, resulting in more accurate localization of structural anomalies, as exemplified by the "splicing connectors" category. These innovations collectively enable our method to outperform state-of-the-art approaches consistently,

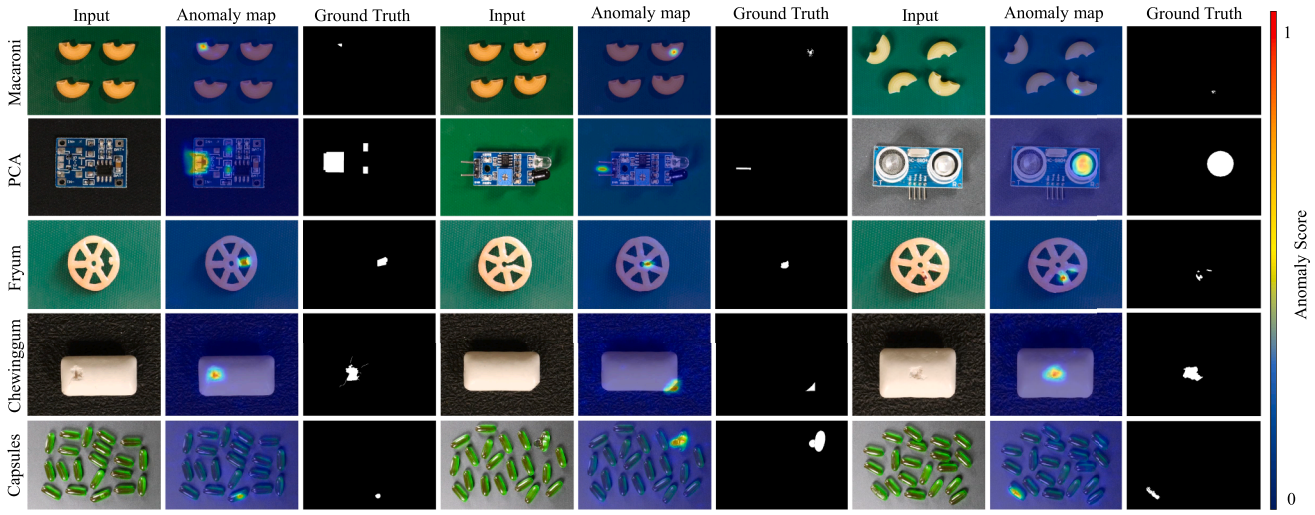

Fig. 8. Visualization of anomaly localization results on different categories of VisA dataset.

Table 3
Quantitative results on the BTAD dataset for anomaly detection/localization, as measured on AUROC [%]. The best results are marked in bold.

Dataset	BTAD			
Method	1	2	3	Avg
Patch-SVDD (Yi & Yoon, 2020)	95.7/92.3	74.6/93.7	82.8/90.8	84.4/92.2
IGD (Chen et al., 2022)	92.3/79.9	63.5/83.9	91.7/90.3	82.5/84.7
U-Student (Bergmann et al., 2020)	93.4/83.6	88.8/95.4	98.7/81.8	93.6/86.9
MKD (Salehi et al., 2021)	93.6/85.4	75.6/86.1	**100**/99.1	89.8/90.2
STFPM (Wang et al., 2021)	90.5/94.7	81.2/97.3	99.3/99.4	90.3/97.1
RD(Deng & Li, 2022)	98.1/96.4	82.1/**96.6**	**100**/**99.7**	93.4/97.5
FYD (Zheng et al., 2021)	99.4/96.2	88.1/95.4	98.7/99.4	95.4/97.0
DRAEM (Zavrtanik et al., 2021)	78.6/68.1	80.5/89.1	99.5/87.1	86.2/81.4
Panda (Reiss et al., 2021)	96.4/96.4	82.2/95.9	99.9/99.3	92.8/97.2
PaDiM (Defard et al., 2021)	99.4/97.2	82.5/95.2	99.9/99.7	93.9/97.4
Efficient AD (Batzner et al., 2024)	99.1/95.8	91.2/95.5	99.9/99.4	96.7/96.9
DHKD	**99.9**/**97.2**	**93.5**/96.3	**100**/99.5	**97.8**/**97.7**

Table 4
Computational efficiency analysis experiments on the MVTec LOCO AD dataset.

Method	Pixel-sPRO	Number of Parameters [$\times 10^6$]	FLOPs[$\times 10^9$]	Latency[ms]
GCAD (Bergmann et al., 2022)	0.701	65	416	16
PatchCore (Roth et al., 2022)	0.546	150 + 8	159 + kNN	45
S-T (Bergmann et al., 2020)	0.626	26	4468	88
FastFlow (Yu et al., 2021)	0.568	92	85	24
SimpleNet (Liu et al., 2023c)	0.363	73	38	18
EfficientAD (Batzner et al., 2024)	0.798	21	235	6.5
Ours	0.833	33	196	7.6

achieving superior accuracy, robustness, and adaptability across various industrial tasks. Qualitative results, illustrated in the Fig. 5 further validates the effectiveness of our method. These findings emphasize the generalizability and reliability of our method, underscoring its suitability for real-world industrial anomaly detection applications.

Results on VisA dataset To validate the generalisability of the proposed method, we conducted experiments on the VisA dataset, which is currently the largest industrial anomaly detection dataset with diverse object types. The results presented in Table 2 show that the proposed method outperforms state-of-the-art approaches in both image-level and pixel-level anomaly localization (AUROC). In particular, the method shows significant improvements on subsets characterized by complex structural dependencies and irregular object arrangements, such as PCBs and capsules, where long-range spatial dependencies are critical for accurate detection. For PCB subsets, which feature intricate arrangements of transistors, capacitors, and chips, the method achieves near-perfect AUROC scores of 99.7% and 99.9% for PCB1 and PCB2, respectively. Similarly, the Capsules and Macaroni2 subsets, which are notable for their variability in pose and location, show robust results with AUROC values consistently exceeding those of existing models. The qualitative results are shown in the Fig. 8. The model's ability to address both local and global anomalies contributes to its robustness, even under the challenging conditions presented by this dataset. These results highlight the adaptability of the method to diverse and complex industrial scenarios and confirm its effectiveness in real-world anomaly detection and localization tasks.

Results on BTAD dataset To evaluate the effectiveness of our proposed method, we conducted experiments on the BTAD dataset and compared it with state-of-the-art techniques for anomaly detection and localization. Fig. 7 and Table 3 summarize the qualitative and quantitative results, underscoring the consistent superiority of our approach across all dataset subsets. Our method achieves a mean AUROC of 97.8% for anomaly detection and 97.7% for anomaly localization, outperforming all competing methods. Notably, it delivers near-perfect detection results (AUROC of 99.9%) on subsets 1 and 3, while maintaining high localization accuracy across all subsets. Compared to existing approaches, our method exhibits remarkable robustness and precision. For instance, while Efficient AD achieves competitive results with an average detection AUROC of 96.7% and localization AUROC of 96.9%, our method surpasses it by 1.1% and 0.8%, respectively.

This approach exploits the differences in feature extraction between the teacher and student networks to achieve high accuracy in anomaly detection. The Mamba module enhances global anomaly detection by capturing long-range spatial dependencies, thereby improving robustness in complex scenarios. To further refine anomaly detection, the Multi-scale Feature Generation Unit (MFGU) facilitates non-local feature interaction, strengthening the model's ability to reconstruct fine structural details and increasing sensitivity to local anomalies. These results highlight the robustness and adaptability of the method, confirming its applicability to real-world anomaly detection and localization tasks.

4.3. Computational efficiency analysis

In the context of real-time performance, Table 4 summarises the key metrics of several representative unsupervised anomaly detection methods on the MVTec LOCO AD dataset, including Pixel-sPRO, parameter count, FLOPs, and latency. All baseline latency values were

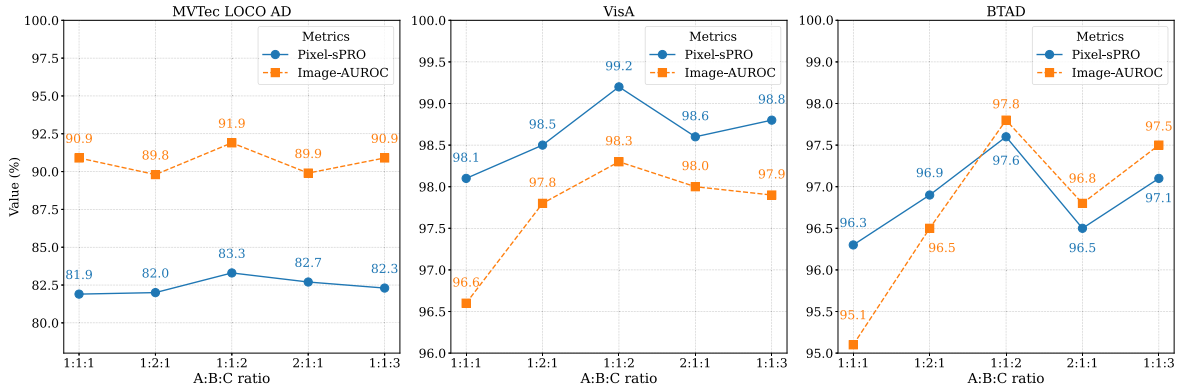

Fig. 9. Results for different proportions of the hyperparameters $\alpha : \beta : \gamma$ on the MVTec LOCO AD, VisA, and BTAD dataset.

Table 5
Ablation experiments on the MVTec LOCO AD dataset, as measured on pixel-sPRO [%].

Baseline	Pixel-sPro			Image-AUROC
	Structural	Logical	Mean	
baseline	0.821	0.725	0.773	0.854
baseline + FCCM	0.846	0.751	0.799	0.882
baseline + SCMM + FCCM	0.878	0.758	0.818	0.901
baseline + HSTM + FCCM	0.856	0.783	0.820	0.908
baseline + SCMM + HSTM	0.862	0.767	0.814	0.911
baseline + SCMM + HSTM + FCCM	0.880	0.786	0.833	0.919

measured by us using the official implementations or officially released best-performing models from the original papers, evaluated under a unified hardware and software environment. DHKD achieves a latency of 7.6 ms, which is markedly faster than heavy-weight methods such as PatchCore and S-T, demonstrating its suitability for high-speed manufacturing scenarios where rapid response is required. In terms of model size and computational cost, DHKD uses 33 million parameters and 196×10^9 FLOPs, placing it between lightweight and heavyweight competitors. Although DHKD contains more parameters and exhibits slightly higher latency than EfficientAD (7.6 ms vs. 6.5 ms), it delivers a higher Pixel-sPRO score, indicating that the performance improvement does not come at the expense of prohibitive computational overhead. Overall, rather than claiming superior efficiency, DHKD strikes a favourable trade-off between anomaly detection performance and computational efficiency. The integration of SCMM for local fidelity and the Mamba-based HSTM for global semantics enables robust feature distillation without sacrificing throughput, making DHKD well-suited for real-time industrial inspection settings in which both accuracy and responsiveness are crucial.

4.4. Ablation analysis and discussion

In this section, we present the results of ablation experiments conducted on the MVTec LOCO AD dataset to evaluate the effectiveness of various components integrated into our proposed model. The experiments were designed to systematically assess the contribution of each module, including the Hybrid Self-Task Mamba Module (HSTM), Spatial Channel Mixer Modulation (SCMM), and the Feature Compression Channel Mixer (FCCM). The results of these ablation experiments, shown in Table 5, demonstrate the significant impact of each component on the overall anomaly detection performance.

The performance of the baseline serves as a reference point for understanding the improvements introduced by each added component. This initial result highlights the potential of the model but also shows that it lacks some critical mechanisms for effectively handling complex anomaly detection tasks. When the baseline is combined with the Feature Compression Channel Mixer (FCCM), the Image-AUROC improves to 0.882 sPRO, which demonstrates the importance of feature compression for enhancing the detection ability of the model. The Feature Compression Channel Mixer leads to a consistent improvement in both anomaly types by enhancing feature compactness and stabilising teacher-student alignment. Incorporating the Spatial Channel Mixer Modulation (SCMM) alongside FCCM further enhances the model's performance, reaching an Image-AUROC of 0.901 sPRO. SCMM allows for non-local feature interactions, which are crucial for improving sensitivity and precision, particularly in local anomaly detection. Building upon this, the introduction of the Spatial Channel Mixer Modulation produces a substantial gain in structural pixel-sPRO, demonstrating that the enriched spatial-channel mixing mechanism improves the reconstruction of fine-grained textures and local defects. Adding the Hybrid Self-Task Mamba Module (HSTM) to the baseline, along with FCCM, results in an Image-AUROC of 0.908 sPRO, showcasing the importance of capturing remote spatial dependencies for global anomaly detection. The HSTM allows the model to better understand the spatial relationships across large regions of an image, improving the detection of anomalies that may span across distant areas. The baseline exhibits stronger sensitivity to structural deviations than to logical inconsistencies, indicating that the student-teacher backbone alone struggles to capture long-range semantic relations. In contrast, adding the Hybrid Self-Task Mamba Module results in the largest improvement in logical pixel-sPRO among all single-module additions. This confirms that long-range dependency modelling is critical for identifying semantic inconsistencies such as missing or misplaced components, which cannot be captured by purely local reconstruction mechanisms. As shown in Table 5, SCMM and HSTM are evaluated by incrementally extending the same baseline + FCCM model. The results suggest that the two branches introduce different inductive biases that primarily benefit their respective anomaly types, while maintaining complementary interactions, instead of enforcing a hard separation between structural and logical detection. The combination of SCMM and HSTM further boosts the performance to an Image-AUROC of 0.911 sPRO, as both local and global anomaly patterns are captured. The final model, incorporating all three modules–SCMM, HSTM, and FCCM–achieves the highest Image-AUROC of 0.919 sPRO. This configuration fully integrates the advantages of each component, making the model more robust and accurate in detecting anomalies in dynamic and complex environments. The combination of SCMM and HSTM further boosts the performance to an Image-AUROC of 0.911 sPRO, as both local and global anomaly patterns are captured. The final model, incorporating all three modules–SCMM, HSTM, and FCCM–achieves the highest Image-AUROC of 0.919 sPRO. This configuration fully integrates the advantages of each component, making the model more robust and accurate in detecting anomalies in dynamic and complex environments.

We analysed the effects of varying the loss-weight ratios $\alpha : \beta : \gamma$ on the performance of our model by conducting experiments on three

benchmark datasets: MVTec LOCO AD, VisA and BTAD, as shown in Fig. 9. Specifically, the ratio $\alpha : \beta : \gamma = 1 : 1 : 2$ consistently produced higher Image-AUROC scores across all three datasets, demonstrating that appropriately balancing the contributions of the three loss terms was crucial for strengthening global-local consistency and improving overall discrimination. The terms $\mathcal{L}_L$ and $\mathcal{L}_G$ constrained the student with respect to local and global teacher representations, respectively, whereas $\mathcal{L}_D$ directly aligned the student encoder output f_S with the globally refined representation f_G. Because f_G integrated decoder-enhanced global semantics, weighting $\mathcal{L}_D$ more heavily was effective at suppressing representational drift in the student encoder and at stabilising optimisation; this, in turn, improved robustness to subtle or spatially distributed anomalies.

To substantiate these claims, we expanded the sensitivity analysis beyond the original five configurations. Across the three datasets and multiple weight settings, the 1:1:2 configuration consistently yielded the best or near-best performance for both pixel-sPRO and Image-AUROC. While pixel-sPRO remained relatively stable across different ratios, Image-AUROC benefited noticeably from emphasising $\mathcal{L}_D$, confirming that enforcing student alignment to the globally enriched f_G was essential for maximising anomaly discrimination. These results collectively demonstrated that the $1 : 1 : 2$ weighting offered a robust and well-balanced choice across heterogeneous anomaly detection scenarios.

4.5. Limitations and failure modes

Although the proposed dual-heterogeneous distillation framework achieves strong performance in terms of both anomaly detection and localisation, it still exhibits several limitations. When structural and logical anomalies occur in the same spatial region, the two decoder branches may emphasise different aspects of the input. The SCMM branch prioritises fine-grained texture and structural cues, while the HSTM branch focuses on long-range semantic consistency. In tightly coupled anomaly cases, these differing inductive biases can lead to partially conflicting residual responses, which may reduce the reliability of the fused anomaly map. The Mamba-based decoder introduces specific failure modes. Due to its reliance on global context, it is less effective at detecting very small, texture-like anomalies that lack coherent semantic information. These limitations suggest that, although long-range reasoning improves logical anomaly detection, it cannot entirely replace local texture sensitivity in all scenarios. Overall, the dual-heterogeneous formulation brings clear benefits but also presents new challenges related to cross-branch consistency and robustness under complex anomaly distributions. Future work may explore adaptive fusion strategies and multi-scale Mamba variants to mitigate these limitations.

5. Conclusion

This study proposes a Dual-Heterogeneous Knowledge Distillation Network with Mamba for industrial anomaly detection, effectively capturing both local and global anomalies. By exploiting structural and parametric heterogeneity, the framework mitigates feature mapping conflicts across branches. The hybrid self-task Mamba module, together with a Multi-scale Feature Generation Unit, further enhances the detection of global and local anomalies. Extensive experiments on three industrial datasets demonstrate state-of-the-art performance, confirming the method's robustness and practical relevance for manufacturing applications.

CRediT authorship contribution statement

Muhao Xu: Conceptualization, Methodology, Software, Writing – original draft; **Zihan Nie:** Writing – review & editing, Data curation, Visualization; **Baochen Fu:** Writing – review & editing; **Zhuangzhuang Chen:** Writing – review & editing; **Zijian Li:** Writing – review & editing; **Hua Wei:** Writing – review & editing; **Yi Wan:** Writing – review & editing; **Weiye Song:** Writing – review & editing, Project administration, Funding acquisition.

Data availability

I shared the code link in the manuscript.

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.

Acknowledgment

This work was supported in part by the Youth Project of Natural Science Foundation of Shandong Province, China (ZR2023QC262); the National Natural Science Foundation of China (62205181); the Natural Science Foundation of Shandong Province (ZR2022QF017); and the Shandong Province Outstanding Youth Science Fund Project (Overseas) (2023HWYQ-023); the Key R&D Program of Shandong Province, China (2024CXGC010106); the Taishan Scholar Foundation of Shandong Province (tsqn202211038).

References

An, J., & Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, *2*(1), 1–18.

Batzner, K., Heckler, L., & König, R. (2024). EfficientAD: Accurate visual anomaly detection at millisecond-level latencies. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 128–138).

Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., & Steger, C. (2022). Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, *130*(4), 947–969.

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).

Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., & Steger, C. (2018). Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv preprint arXiv:1807.02011.

Chen, Y., Tian, Y., Pang, G., & Carneiro, G. (2022). Deep one-class classification via interpolated gaussian descriptor. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)* (pp. 383–392). (*vol. 36*).

Cohen, N., & Hoshen, Y. (2020). Sub-image anomaly detection with deep pyramid correspondences. arXiv preprint arXiv:2005.02357.

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)* (pp. 475–489). Springer.

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).

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th international conference on machine learning* (pp. 1126–1135). PMLR (*vol. 70*). Proceedings of Machine Learning Research.

Guo, H., Ren, L., Fu, J., Wang, Y., Zhang, Z., Lan, C., Wang, H., & Hou, X. (2023). Template-guided hierarchical feature restoration for anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 6447–6458).

Guo, J., Zheng, P., & Huang, J. (2019). Efficient privacy-preserving anomaly detection and localization in bitstream video. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(9), 3268–3281.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 770–778).

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications*, *193*, 116429.

Lei, X., Sun, M., Zhao, R., Wu, H., Zhou, Z., Dong, Y., & Sun, L. (2024). Unsupervised vision-based structural anomaly detection and localization with reverse knowledge distillation. *Structural Control and Health Monitoring*, *2024*(1), 8933148.

Liu, R., Liu, W., Zheng, Z., Wang, L., Mao, L., Qiu, Q., & Ling, G. (2023a). Anomaly-GAN: A data augmentation method for train surface anomaly detection. *Expert Systems with Applications*, *228*, 120284.

Liu, W., Chang, H., Ma, B., Shan, S., & Chen, X. (2023b). Diversity-measurable anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12147–12156).

Liu, Z., Zhou, Y., Xu, Y., & Wang, Z. (2023c). SimpleNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20402–20411).

Lu, M., Chai, Y., Xu, K., Chen, W., Ao, F., & Ji, W. (2025). Multimodal fusion and knowledge distillation for improved anomaly detection. *The Visual Computer*, *41*(8), 5311–5322.

Ma, Z., Li, J., & Wong, W. K. (2025). Patch distance based auto-encoder for industrial anomaly detection. *Expert Systems with Applications*, *270*, 126537.

Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., & Foresti, G. L. (2021). VT-ADL: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th international symposium on industrial electronics (ISIE)* (pp. 01–06). IEEE.

Park, H., Noh, J., & Ham, B. (2020). Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 14372–14381).

Reiss, T., Cohen, N., Bergman, L., & Hoshen, Y. (2021). PANDA: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2806–2814).

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).

Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M. H., & Rabiee, H. R. (2021). Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 14902–14912).

Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., & Schmidt-Erfurth, U. (2019). f-anoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, *54*, 30–44.

Steger, C., Ulrich, M., & Wiedemann, C. (2018). Machine vision algorithms and applications[M]. John Wiley & Sons.

Tang, H., Li, Z., Zhang, D., He, S., & Tang, J. (2024). Divide-and-conquer:Confluent triple-flow network for RGB-t salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*, (pp. 1958–1974).

Tang, H., Yuan, C., Li, Z., & Tang, J. (2022). Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition*, *130*, 108792.

Wang, G., Han, S., Ding, E., & Huang, D. (2021). Student-teacher feature pyramid matching for anomaly detection. arXiv preprint arXiv:2103.04257.

Wu, J.-C., Chen, D.-J., Fuh, C.-S., & Liu, T.-L. (2021). Learning unsupervised metaformer for anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 4369–4378).

Wu, K., Zhu, L., Shi, W., Wang, W., & Wu, J. (2023). Self-attention memory-augmented wavelet-CNN for anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(3), 1374–1385.

Wu, Q., Li, H., Tian, C., Wen, L., & Li, X. (2024). AEKD: Unsupervised auto-encoder knowledge distillation for industrial anomaly detection. *Journal of Manufacturing Systems*, *73*, 159–169.

Xing, P., Tang, H., Tang, J., & Li, Z. (2024). ADPS: Asymmetric distillation postsegmentation for image anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(4), 7051–7064.

Xu, M., Zhou, X., Gao, X., Song, W., Feng, G., & Niu, S. (2025a). Normality prior-guided multisemantic fusion network for unsupervised image anomaly detection. *IEEE Transactions on Instrumentation and Measurement*, *74*, 1–12.

Xu, M., Zhu, C., Feng, G., & Niu, S. (2025b). Multitask hybrid knowledge distillation for unsupervised anomaly detection. *IEEE Transactions on Industrial Informatics*, *21*, 5666–5676.

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* (pp. 3110–3118). (*vol. 35*).

Yi, J., & Yoon, S. (2020). Patch SVDD: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the asian conference on computer vision (ACCV)*.

Yousefimehr, B., Ghatee, M., & Razavi-Far, R. (2025). Multi-teacher knowledge distillation framework for lightweight anomaly detection. *Neural Networks*, *195*, 108267.

Yu, J., Zheng, Y., Wang, X., Li, W., Wu, Y., Zhao, R., & Wu, L. (2021). FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows. arXiv preprint arXiv:2111.07677.

Zavrtanik, V., Kristan, M., & Skočaj, D. (2021). 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).

Zhang, J., Suganuma, M., & Okatani, T. (2024). Contextual affinity distillation for image anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 149–158).

Zheng, Y., Wang, X., Deng, R., Bao, T., Zhao, R., & Wu, L. (2021). Focus your distribution: Coarse-to-fine non-contrastive learning for anomaly detection and localization. arXiv preprint arXiv:2110.04538.

Zhou, Q., He, S., Liu, H., Chen, T., & Chen, J. (2022). Pull & push: Leveraging differential knowledge distillation for efficient unsupervised anomaly detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(5), 2176–2189.

Zou, Y., Jeong, J., Pemula, L., Zhang, D., & Dabeer, O. (2022). Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision* (pp. 392–408). Springer.