

ARTICLE

ViLoc-Net: Leveraging Synthetic Defect Generation and Vision Transformers for Industrial Surface and Texture Anomaly Detection

Asim Niaz^{1,#}, Muhammad Umraiz^{2,#}, Syed Farhan Alam Zaidi³ and Kwang Nam Choi^{2,*}

¹Ludwig-Maximilians-Universität München, Munich, Germany

²Department of Computer Science and Engineering, Chung-Ang University, Seoul, Republic of Korea

³Premium Research Institute for Human Metaverse Medicine (WPI-PRIME), The University of Osaka, Suita City, Osaka, Japan

*Corresponding Author: Kwang Nam Choi. Email: knchoi@cau.ac.kr

Asim Niaz and Muhammad Umraiz contributed equally to this work

Received: 05 May 2026; Accepted: 15 June 2026; Published: 28 August 2026

ABSTRACT: Automated visual inspection in industrial settings often struggles with limited defect data and poor generalization to unseen anomalies. To overcome this challenge, we propose a hybrid anomaly detection pipeline, which integrates embedding-based, reconstruction-based, and self-supervised learning approaches. The framework also proposes a new Realistic Industrial Defect Synthesis (RIDS) module that synthesizes structured and textured synthetic anomalies based on the target masks, composite maps, and blending techniques. This helps to learn from pseudo-labeled data without the need for large annotated datasets. The pipeline further includes ViLoc-Net, a Vision Transformer-based localization network that obtains global features and then reconstructs detailed segmentation maps by a multi-scale decoder. The method is tested on various industrial datasets, demonstrating high accuracy, robustness, and generalization, making it suitable for real-world inspection tasks.

KEYWORDS: Image anomaly; industrial anomaly; vision transformers; surface anomaly detection

1 Introduction

Anomalies are unwanted behaviors or patterns that differ from the expected data distribution being observed [1]. Image anomaly detection is about finding unexpected images or areas within those images [2]. This method has many uses in industrial safety and inspection and is very important [3–6]. Detecting anomalies in these situations is especially tough because incidents that differ from established patterns are rare. This leads to a lack of anomalous samples for training computer vision models that use a supervised learning approach. This limits exploration of supervised methods further [7]. Irregularities or anomalies can vary in shape, size, and severity, and may appear anywhere on the surface of a product [8].

Current anomaly detection approaches address the limitation of data scarcity by adopting unsupervised methods [9–12], where models learn general patterns from normal images and flag anomalous images that exhibit deviations from these learned patterns. These methods often utilize techniques such as reconstruction and embedding-based representations to tackle this challenge effectively [13]. Reconstruction-based methods often involve auto-encoders operating under the assumption that a model trained on normal patterns will flag deviations when presented with anomalous patterns [14–17]. However, this assumption often breaks down as the model tends to generalize to unseen patterns, leading to compromised performance in

anomaly detection. On the other hand, discriminators can be described as mechanisms that detect deviations from the tightly packed clustering of normal textures within a deep subspace [18]. Creating such compact representations ensures that anomalous instances are not mapped near non-anomalous examples. Fig. 1 provides a visual representation of both scenarios.

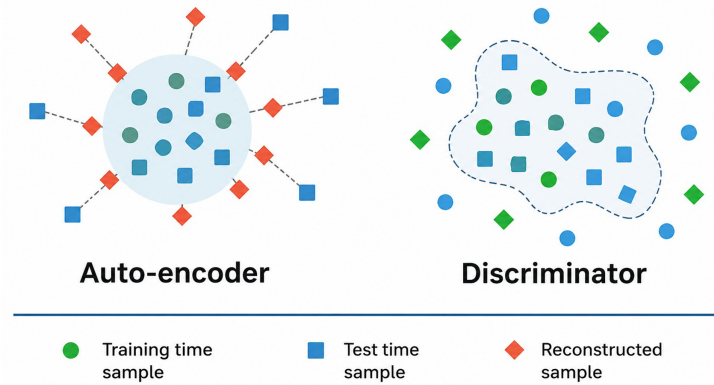


Figure 1: Autoencoders often treat anomalies like normal data, while discriminators work well on fake anomalies but fail on real ones [19].

Similarly, embedding-based methods often use convolutional neural networks (CNNs) that are pre-trained on ImageNet to achieve top performance in general anomaly detection tasks [13,20]. However, for industrial images, the distribution can be quite different from that of ImageNet, which can reduce the performance of these models on industrial datasets.

The proposed method combines elements from embedding-based, reconstruction-based, and self-supervised learning to improve anomaly detection. The Vision Transformer (ViT) encoder extracts global features to project input images into a high-dimensional embedding space, similar to embedding-based approaches. At the same time, the multi-scale decoder progressively reconstructs anomaly segmentation maps, ensuring precise localization and high-resolution outputs, which align with reconstruction-based techniques. Additionally, the RIDS module introduces synthetic anomalies that imitate real-world defect patterns. This allows the model to learn strong features through pseudo-labeled data, a key aspect of self-supervised learning. To help visualize the overall process, Fig. 2 provides a high-level overview of the proposed hybrid anomaly detection framework. This combination enables the proposed method to use the strengths of various approaches, ensuring strong performance in both anomaly segmentation and localization tasks.

- **Realistic Industrial Defect Synthesis (RIDS) Module:** Developed a data augmentation module to generate realistic industrial anomalies using target masks, anomaly masks, and blending, enabling robust feature learning with pseudo-labeled data.
- **Hybrid Anomaly Detection Framework:** Introduced a novel hybrid approach that integrates embedding-based, reconstruction-based, and self-supervised learning paradigms, enhancing the detection and segmentation of industrial anomalies.
- **Vision Transformer-Based Localization Network (ViLoc-Net):** Designed a Vision Transformer-based localization network that leverages global feature extraction and a multi-scale decoder for precise anomaly segmentation and high-resolution localization.

- **Application-Friendly Pipeline:** Designed a versatile framework for industrial anomaly detection, ensuring applicability across diverse datasets and scenarios.

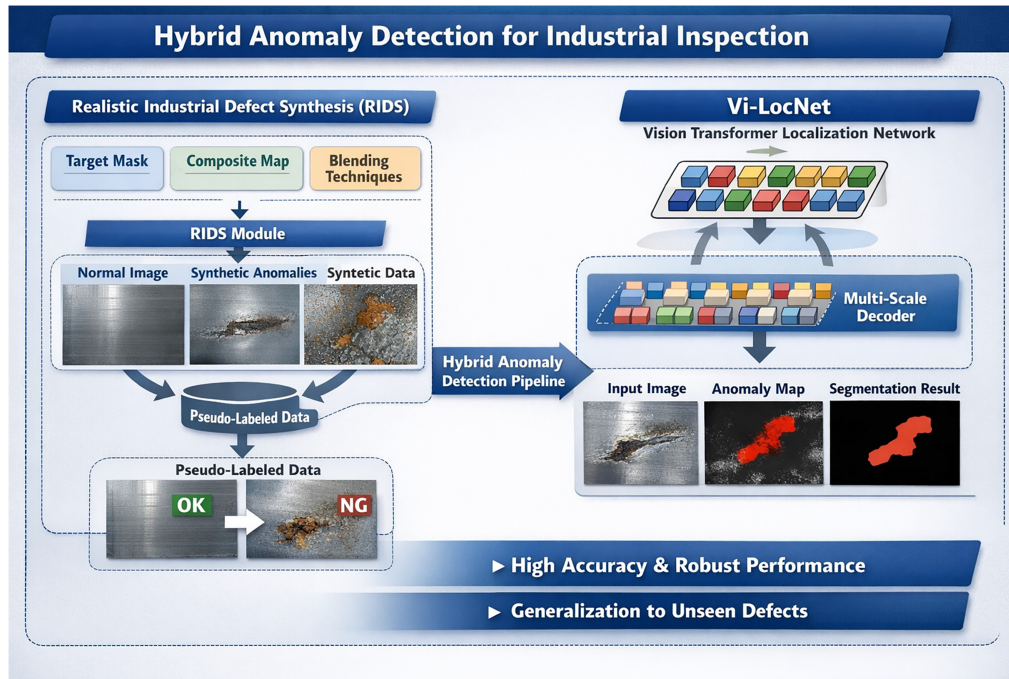


Figure 2: High-level overview of the proposed hybrid anomaly detection framework integrating RIDS-based synthetic defect generation with hybrid learning and ViLoc-Net for anomaly localization.

2 Related Work

Industrial anomaly detection and localization mainly involve two types of methods, embedding-based and reconstruction-based approaches. Self-supervised methods boost both types by improving their effectiveness. In embedding-based approaches, self-supervised tasks such as contrastive learning or clustering assist in training models to create strong feature embeddings. In reconstruction-based approaches, tasks like inpainting or denoising help models concentrate on normal patterns, which enhances their ability to spot anomalies.

Embedding-based methods detect anomalies by analyzing feature representations (embeddings) of data within a predefined feature space. Typical approaches [13,20,21] use networks that are pre-trained on ImageNet for feature extraction. For example, PaDiM [13] uses a pre-trained model to represent anomaly patch features through a multivariate Gaussian distribution. PatchCore [20] employs a memory bank of patch features during training, and assigns anomaly scores at test time based on the maximum feature distance. However, this approach may limit real-time performance because of the computational load of searching for similarities in the memory bank. CFLOW-AD [22] and DifferNet [23] proposed anomaly detection methods that use normalizing flows to change normal feature distributions into Gaussian distributions. But a limitation of these methods is that they cannot reduce feature maps. They need full-sized feature maps, which results in high memory use. JRCC-Net [24] uses an anomaly simulation strategy and joint representation learning with contrast clustering to detect and localize anomalies effectively. However, its reliance on memory samples and complex representations can cause significant memory overhead. In general, embedding-based image anomaly detection methods often struggle to distinguish subtle anomalies in complex or high-dimensional feature spaces.

Reconstruction-based methods for image anomaly detection aim at learning to reconstruct normal data patterns, but struggle with anomalous data. The techniques frequently rely on deep learning models such as autoencoders (AEs) [25,26] or generative adversarial networks (GANs) [27,28]. Some methods like [14,16,29] formulate the problem of anomaly detection as an inpainting problem. For this task, patches of an image are randomly masked, and the networks are required to fill in the missing areas. Reconstruction by Inpainting for Visual Anomaly Detection (RIAD) [14] proposed a multi-scale complementary masking method on the original image. This is done to minimize the impact of atypical regions on the reconstruction process by masking them. In the same way, the Inpainting Transformer (InTra) [30] applied a transformer-based architecture that took masked images as input and performed the task of repairing them. In general, reconstruction techniques do not work well when anomalies have similar patterns to the normal data.

To conclude, both embedding-based and reconstruction-based approaches have drawbacks in detecting small or complicated anomalies. This indicates a need for the improvement of techniques. Besides, Vision Transformers were applied to the classification of leather surface defects and to qualitatively localize anomalies using small datasets by Smith et al. [31]. Our proposed ViLoc-Net is for pixel-level anomaly detection and localization, while their work focuses primarily on image-level classification without a quantitative evaluation of localization. It adopts a multi-scale ViT-based architecture combined with synthetic defect augmentation to enhance the robustness.

3 Proposed Method

The RIDS module and the vision transformer-based localization network together constitute the proposed ViLoc-Net method.

3.1 Realistic Industrial Defect Synthesis (RIDS)

To create a strong anomaly detection framework, we propose a *RIDS* process that simulates realistic anomalies in industrial images. This method ensures that anomalies look like actual defect patterns, improving the training and evaluation of anomaly detection models. The overall process of the proposed RIDS module is shown in Fig. 3. Fig. 4 displays example anomalies and masks generated by the RIDS module. The process includes the following steps:

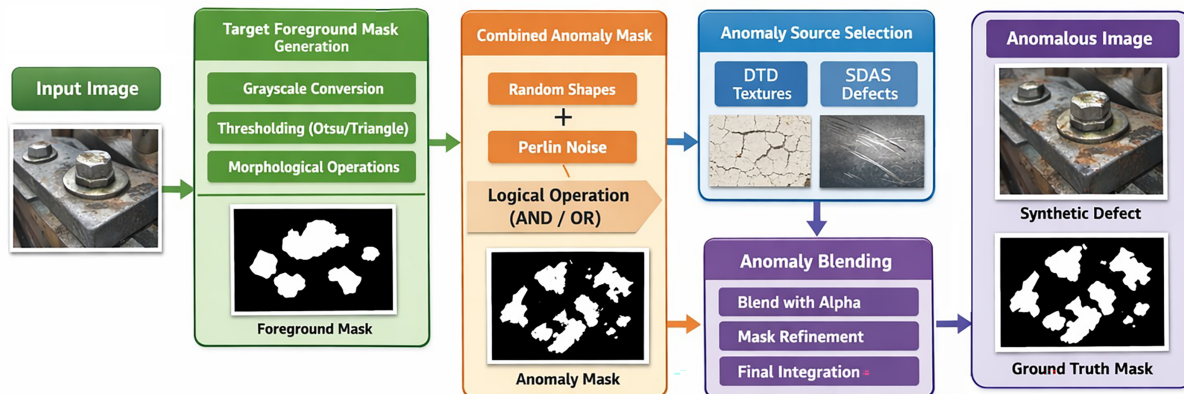


Figure 3: Flowchart of the proposed realistic industrial defect synthesis (RIDS) module. The process begins with the input image, creates relevant masks, selects anomaly sources, blends them, and finally outputs a realistic anomalous image along with its ground truth mask.

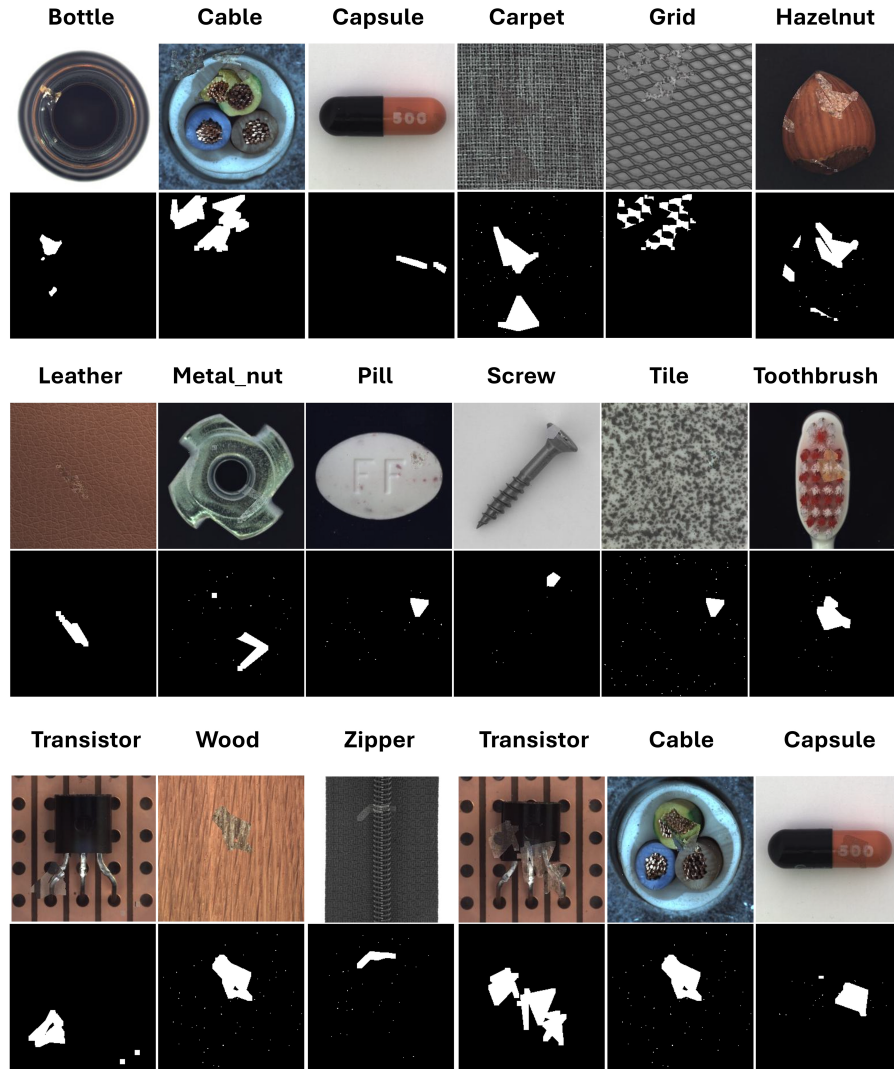


Figure 4: Illustration of anomalies generated by the RIDS module: Class names are displayed at the top, followed by anomalous images and their corresponding ground truth masks created by the RIDS module.

3.1.1 Target Foreground Mask Generation

The first step is to isolate the foreground regions of interest from the input image. A precise target foreground mask is important to ensure that anomalies are introduced only in the relevant areas of the image. Depending on the dataset and its subclass, the following techniques are used to achieve this:

- **Grayscale Conversion:** The input RGB image is converted to grayscale to reduce computational complexity and focus on intensity variations that highlight structural differences. This conversion is performed as $I_{\text{gray}} = 0.2989 \cdot I_R + 0.5870 \cdot I_G + 0.1140 \cdot I_B$, where I_R , I_G , and I_B are the red, green, and blue channels of the input RGB image, respectively.
- **Thresholding:** Depending on the nature of the data and the characteristics of the subclass, thresholding techniques are used to segment the image:
 - **Otsu's Method:** This global thresholding method automatically determines the optimal threshold by minimizing intra-class variance. It is suitable for images with bimodal histograms, where

foreground and background intensities are clearly separable. The threshold is computed as:

$$T_{\text{Otsu}} = \arg \min_T (w_0(T) \cdot \sigma_0^2(T) + w_1(T) \cdot \sigma_1^2(T)) \quad (1)$$

where $w_0(T)$ and $w_1(T)$ are the probabilities of the two classes separated by the threshold T , and $\sigma_0^2(T)$ and $\sigma_1^2(T)$ are the variances of the two classes.

- **Triangle Thresholding:** This method is ideal for images with skewed intensity distributions, where one class (foreground or background) dominates the histogram. It determines the threshold T_{Triangle} geometrically as the maximum distance from the histogram curve to a straight line connecting the histogram peak to its ends.
- **Morphological Operations:** After thresholding, morphological operations such as closing (dilation followed by erosion) and opening (erosion followed by dilation) are applied:
 - **Closing:** This operation fills small holes or gaps in the foreground, ensuring that the mask is contiguous and complete. It is defined as:

$$I_{\text{close}} = (I \oplus S) \ominus S \quad (2)$$

where $\oplus$ denotes dilation, $\ominus$ denotes erosion, and S is the structuring element.

- **Opening:** This operation removes small noise or isolated pixels from the mask, refining its boundaries and ensuring that irrelevant regions are excluded. It is defined as:

$$I_{\text{open}} = (I \ominus S) \oplus S \quad (3)$$

where $\ominus$ and $\oplus$ have the same definitions as above.

These steps produce a binary mask, termed the *target foreground mask*, which highlights the regions where anomalies will be introduced. The final mask is computed as:

$$M_{\text{foreground}} = \text{MorphologicalRefine}(I_{\text{binary}}) \quad (4)$$

where I_{binary} is the binary image obtained after thresholding, and *MorphologicalRefine* represents the combined application of closing and opening operations.

This ensures that anomalies are concentrated in areas of interest, avoiding irrelevant regions such as the background.

3.1.2 Combined Anomaly Mask Generation

To ensure diversity and complexity in the anomalies, a *combined anomaly mask* is generated by integrating multiple anomaly types:

- **Random Shapes:** Irregular polygonal shapes are generated with random sizes, positions, and vertices. These shapes simulate structured anomalies like scratches, dents, or geometric defects. The shapes are mathematically represented as:

$$M_{\text{shapes}}(x, y) = \begin{cases} 1 & \text{if } (x, y) \in \text{polygon region,} \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where (x, y) are the pixel coordinates. To enhance these shapes, morphological dilation is applied:

$$M_{\text{dilated}} = M_{\text{shapes}} \oplus S, \quad (6)$$

where $\oplus$ denotes the dilation operation, and S is the structuring element.

- **Perlin Noise:** Smooth, natural-like textures are synthesized using Perlin noise, controlled by a frequency parameter scale. The noise is thresholded to generate a binary mask:

$$M_{\text{Perlin}}(x, y) = \begin{cases} 1 & \text{if } \text{Perlin}(x, y, \text{scale}) > T_{\text{Perlin}}, \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

where T_{Perlin} is the threshold. Optional rotation θ can be applied to add spatial variability.

The combined random shape mask M_{combined} is derived based on a selected logical operation between the random shapes mask M_{shapes} and the Perlin noise mask M_{Perlin} . This can be represented as:

$$M_{\text{combined}} = \begin{cases} M_{\text{shapes}} \vee M_{\text{Perlin}}, & \text{if } \mathcal{O} = \text{OR}, \\ M_{\text{shapes}} \wedge M_{\text{Perlin}}, & \text{if } \mathcal{O} = \text{AND}. \end{cases} \quad (8)$$

The resulting mask M_{combined} introduces both structured and textured anomalies, ensuring high diversity and realism in the synthetic defects.

3.1.3 Anomaly Source Selection

To ensure the anomalies are visually realistic, external texture datasets are used as the source of anomalies. These sources add a level of realism by incorporating external patterns:

- **DTD:** Texture-based anomalies are generated by augmenting random DTD images with transformations like rotation and contrast adjustment, then resizing them to the input dimensions.
- **SDAS:** Structural anomalies (e.g., scratches) use images from a predefined defect library [32], resized and applied to the anomaly mask.

3.1.4 Blending Anomalies into the Original Image

The selected anomaly source is blended with the input image within the regions defined by the combined anomaly mask. The blending process is mathematically represented as follows:

- **Transparency Factor:** A random transparency factor $\alpha \in [0, 1]$ controls the blending intensity between the input image and the anomaly source. This ensures variability across samples and is expressed as:

$$I_{\text{blend}}(x, y) = \alpha \cdot I_{\text{anomaly}}(x, y) + (1 - \alpha) \cdot I_{\text{input}}(x, y), \quad (9)$$

where $I_{\text{blend}}(x, y)$ is the blended image at pixel (x, y) , $I_{\text{anomaly}}(x, y)$ is the anomaly source, and $I_{\text{input}}(x, y)$ is the original input image.

- **Mask Restriction:** The blending is restricted to the regions defined by the *combined anomaly mask* $M_{\text{combined}}(x, y)$ and the *target foreground mask* $M_{\text{foreground}}(x, y)$. This ensures that anomalies do not appear in irrelevant areas such as the background. The mask-constrained blending is represented as:

$$I_{\text{final}}(x, y) = \begin{cases} I_{\text{blend}}(x, y), & \text{if } M_{\text{combined}}(x, y) \wedge M_{\text{foreground}}(x, y) = 1, \\ I_{\text{input}}(x, y), & \text{otherwise.} \end{cases} \quad (10)$$

- **Mask Expansion:** For seamless integration, the combined mask may be expanded or softened using morphological operations or Gaussian blurring, ensuring smooth transitions between the original image and the anomaly regions:

$$M_{\text{smooth}} = \text{Blur}(M_{\text{combined}}, \sigma), \quad (11)$$

where σ is the standard deviation of the Gaussian kernel.

This process ensures that the anomalies are realistically blended into the image, maintaining natural variability while adhering to the spatial constraints defined by the masks.

3.1.5 Realistic Anomaly Image Creation

The final output of the proposed process consists of two essential components.

The first component is the **anomalous image** $I_{\text{anomalous}}(x, y)$, which is created by blending the original image $I_{\text{input}}(x, y)$ with the selected anomaly source $I_{\text{anomaly}}(x, y)$. This blending process is guided by the combined anomaly mask $M_{\text{combined}}(x, y)$ and the target foreground mask $M_{\text{foreground}}(x, y)$. The anomalous image is mathematically expressed as:

$$I_{\text{anomalous}}(x, y) = \begin{cases} \alpha \cdot I_{\text{anomaly}}(x, y) + (1 - \alpha) \cdot I_{\text{input}}(x, y), \\ \text{if } M_{\text{combined}}(x, y) \wedge M_{\text{foreground}}(x, y) = 1, \\ I_{\text{input}}(x, y), \quad \text{otherwise,} \end{cases} \quad (12)$$

where α is the transparency factor controlling the blending intensity. The second component is the **binary anomaly mask** $M_{\text{final}}(x, y)$, which precisely indicates the location of the introduced anomalies. It is defined as:

$$M_{\text{final}}(x, y) = M_{\text{combined}}(x, y) \wedge M_{\text{foreground}}(x, y). \quad (13)$$

The combination of the anomalous image and the binary mask offers a complete description of synthetic defects, which allows for powerful test and training scenarios. The proposed RIDS process systematically generates realistic industrial anomalies for improving the performance of anomaly detection models. The entire process is described in Algorithm 1.

Algorithm 1: Realistic industrial defect synthesis (RIDS)

Require: Input image I_{input} , anomaly sources, and α

Ensure: Anomalous image $I_{\text{anomalous}}$ and binary mask M_{final}

1: Generate $M_{\text{foreground}}$ (Eqs. (1)–(4))

2: Create M_{combined} (Eqs. (6)–(8))

3: Select I_{anomaly} from DTD or SDAS

4: Blend anomalies (Eqs. (9)–(11))

5: Generate $I_{\text{anomalous}}$ and M_{final} (Eqs. (12) and (13))

ViLoc-Net: Vision Transformer (Vi) with Localization Network.

This is because the proposed *ViLoc-Net* framework combines the power of Vision Transformers (Vi) with a powerful localization network for anomaly segmentation and localization, where the encoder is based on an ImageNet-pretrained ViT-Base/16 backbone, initialized with pretrained weights and fine-tuned during the anomaly detection training. The architecture is shown in Fig. 5.

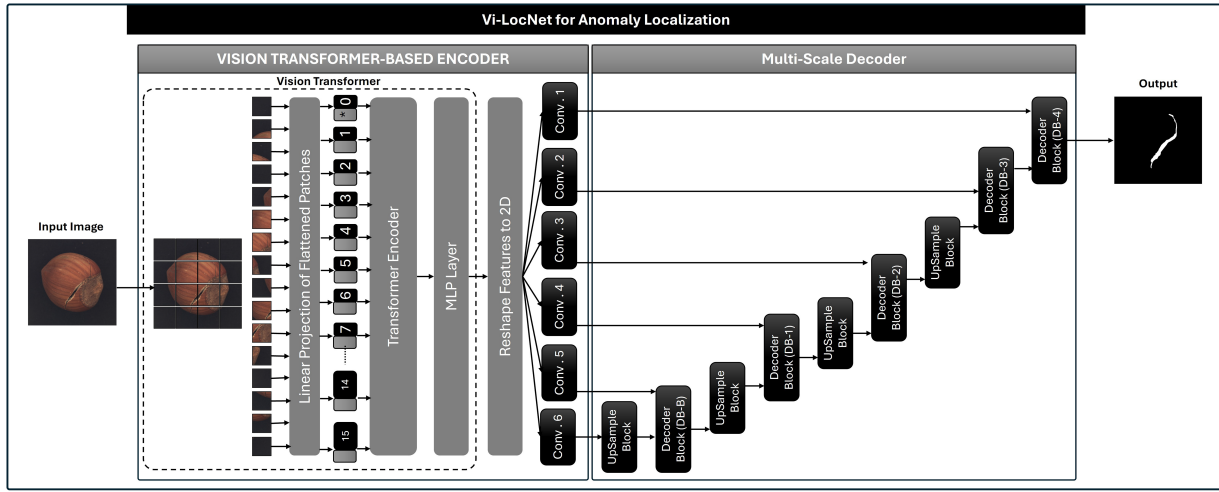


Figure 5: Block diagram illustrating the anomaly detection process in industrial images using the ViLoc-Net architecture, showcasing the encoder-decoder framework for feature extraction and anomaly localization.

The following key components are used in this architecture:

3.1.6 Vision Transformer-Based Encoder

The encoder utilizes a Vision Transformer (ViT) to extract global contextual features from the input image. The input image $I_{\text{input}} \in \mathbb{R}^{H \times W \times C}$ is divided into non-overlapping patches of size $P \times P$, where H , W , and C denote the height, width, and number of channels, respectively. Each patch $I_{\text{patch}} \in \mathbb{R}^{P \times P \times C}$ is flattened and linearly projected into an embedding space of dimension D :

$$z_0 = [\text{PatchEmbed}(I_{\text{patch}}, 1), \text{PatchEmbed}(I_{\text{patch}}, 2), \dots, \text{PatchEmbed}(I_{\text{patch}}, N)], \quad (14)$$

where $N = \frac{H \cdot W}{P^2}$ is the number of patches.

These embeddings z_0 are processed through L transformer layers, each consisting of multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks:

$$z'_\ell = \text{MSA}(\text{LN}(z_{\ell-1})) + z_{\ell-1}, \quad (15)$$

$$z_\ell = \text{MLP}(\text{LN}(z'_\ell)) + z'_\ell, \quad (16)$$

where $\ell = 1, 2, \dots, L$, and LN represents layer normalization.

The output embeddings from the Vision Transformer, $z_L \in \mathbb{R}^{N \times D}$, are reshaped into a 2D spatial structure:

$$f_{\text{ViT}} = \text{Reshape}(z_L) \in \mathbb{R}^{D \times \sqrt{N} \times \sqrt{N}}. \quad (17)$$

To enable multi-scale feature extraction, f_{ViT} is passed through a series of 1×1 convolutional layers:

$$b_i = \text{Conv}_i(f_{\text{ViT}}), \quad i = 1, 2, \dots, 6, \quad (18)$$

where $b_1, b_2, \dots, b_6$ represent hierarchical feature maps. Here, b_6 captures high-level semantic information with reduced spatial resolution, while b_1 retains finer spatial details.

3.1.7 Multi-Scale Decoder

The decoder reconstructs high-resolution segmentation maps by progressively integrating multi-scale features from the encoder. Each level of the decoder performs upsampling, concatenation of skip connections, and feature refinement through convolutional blocks.

The decoder operates in the following stages:

- **Initial Upsampling:** The most abstract feature map b_6 is upsampled and concatenated with b_5 . The resulting tensor undergoes convolutional refinement to produce a new feature map:

$$b'_6 = \text{Conv}(\text{Concat}(\text{Upsample}(b_6), b_5)). \quad (19)$$

- **Intermediate Stages:** Features b_4 , b_3 , and b_2 are sequentially integrated using bilinear upsampling and skip connections from the corresponding encoder layers:

$$b'_k = \text{Conv}(\text{Concat}(\text{Upsample}(b'_{k+1}), b_k)), \quad (20)$$

where $k = \{4, 3, 2\}$.

- **Final Stage:** The lowest-level feature b_1 is fused with the upsampled features to generate the final high-resolution anomaly segmentation map:

$$b'_1 = \text{Conv}(\text{Concat}(\text{Upsample}(b'_2), b_1)). \quad (21)$$

The decoder processes the features extracted by the encoder to enhance the anomaly segmentation and localization tasks. It operates on the anomalous image generated by the RIDS module and produces two refined outputs. The first is a processed anomalous image, which refines the representation of anomalies embedded in the input image. The second is a binary anomaly mask that highlights the precise locations of the detected anomalies. These outputs facilitate accurate evaluation and benchmarking of anomaly detection models.

3.1.8 Integration of Vision Transformer and Localization Network

ViLoc-Net integrates the global feature extraction capabilities of Vision Transformers with the precise reconstruction functionality of a localization network. The hierarchical decoder complements the Vision Transformer by refining spatial details through skip connections and multi-scale processing, ensuring that both high-level semantic information and fine-grained spatial details are preserved. This combination enables accurate anomaly segmentation and localization. The overall functionality of ViLoc-Net can be mathematically expressed as:

$$M_{\text{final}}, I_{\text{refined}} = \text{ViLoc-Net}(I_{\text{input}}, A_{\text{RIDS}}), \quad (22)$$

where $I_{\text{input}} \in \mathbb{R}^{H \times W \times C}$ represents the input image, $A_{\text{RIDS}} \in \mathbb{R}^{H \times W \times C}$ denotes the anomalous image generated by the RIDS module, $M_{\text{final}} \in \mathbb{R}^{H \times W}$ is the binary anomaly mask, and $I_{\text{refined}} \in \mathbb{R}^{H \times W \times C}$ corresponds to the refined anomalous image.

The process involves two main components: an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The encoder extracts hierarchical multi-scale features:

$$f_{\text{multi}} = \mathcal{E}(I_{\text{input}}), \quad (23)$$

where $f_{\text{multi}} = \{b_1, b_2, \dots, b_6\}$ are the multi-scale feature maps. These features are processed by the decoder to produce the outputs:

$$M_{\text{final}}, I_{\text{refined}} = \mathcal{D}(f_{\text{multi}}, A_{\text{RIDS}}). \quad (24)$$

The modular design of ViLoc-Net allows it to work with different input resolutions and types of anomalies, making it suitable for many industrial applications. By using the strengths of Vision Transformers and localization networks, ViLoc-Net shows excellent performance in anomaly detection tasks. The detailed architecture of ViLoc-Net, which includes layer operations and input-output dimensions, is provided in Table 1. The table illustrates how data moves through the network. It points out the operations done at each stage, the input and output shapes, and the key parts of the encoder-decoder framework.

Table 1: Architecture of ViLoc-Net for anomaly localization.

Layer/Block Name	Input Shape	Operation	Output Shape
Encoder (EncoderDiscriminative)			
Vision Transformer (ViT)	(B, 3, 256, 256)	Patch Embedding, Multi-Head Attention	(B, 768, 16, 16)
Conv1	(B, 768, 16, 16)	Conv2d (kernel = 1)	(B, 128, 16, 16)
Conv2	(B, 768, 16, 16)	Conv2d (kernel = 1)	(B, 256, 16, 16)
Conv3	(B, 768, 16, 16)	Conv2d (kernel = 1)	(B, 512, 16, 16)
Conv4	(B, 768, 16, 16)	Conv2d (kernel = 1)	(B, 1024, 16, 16)
Conv5	(B, 768, 16, 16)	Conv2d (kernel = 1)	(B, 1024, 16, 16)
Conv6	(B, 768, 16, 16)	Conv2d (kernel = 1)	(B, 1024, 16, 16)
Decoder (DecoderDiscriminative)			
UpB	(B, 1024, 16, 16)	Upsample (scale = 2) + Conv2d (kernel = 3)	(B, 1024, 32, 32)
DBB	(B, 2048, 32, 32)	Conv2d (kernel = 3) $\times$ 2 + BatchNorm2d	(B, 1024, 32, 32)
Up1	(B, 1024, 32, 32)	Upsample (scale = 2) + Conv2d (kernel = 3)	(B, 512, 64, 64)
DB1	(B, 1024, 64, 64)	Conv2d (kernel = 3) $\times$ 2 + BatchNorm2d	(B, 512, 64, 64)
Up2	(B, 512, 64, 64)	Upsample (scale = 2) + Conv2d (kernel = 3)	(B, 256, 128, 128)
DB2	(B, 512, 128, 128)	Conv2d (kernel = 3) $\times$ 2 + BatchNorm2d	(B, 256, 128, 128)
Up3	(B, 256, 128, 128)	Upsample (scale = 2) + Conv2d (kernel = 3)	(B, 128, 256, 256)
DB3	(B, 256, 256, 256)	Conv2d (kernel = 3) $\times$ 2 + BatchNorm2d	(B, 128, 256, 256)
DB4	(B, 256, 256, 256)	Conv2d (kernel = 3) $\times$ 2 + BatchNorm2d	(B, 128, 256, 256)
Final Output	(B, 128, 256, 256)	Conv2d (kernel = 3)	(B, 1, 256, 256)
Full Network (DiscriminativeSubNetwork)			
Encoder Output	(B, 3, 256, 256)	EncoderDiscriminative	Multi-Scale Outputs
Decoder Output	Multi-Scale Inputs	DecoderDiscriminative	(B, 1, 256, 256)

The encoder, referred to as EncoderDiscriminative, extracts multi-scale feature representations from the input image using a Vision Transformer (ViT) followed by a series of convolutional layers. The Vision Transformer processes the input image of shape (B, 3, 256, 256), applies patch embedding and multi-head attention, and outputs a feature map of shape (B, 768, 16, 16). Subsequently, six convolutional layers (Conv1 to Conv6) with a kernel size of 1×1 progressively extract features with increasing channel dimensions. The output shapes range from (B, 128, 16, 16) for Conv1 to (B, 1024, 16, 16) for Conv6.

The decoder, referred to as DecoderDiscriminative, reconstructs the feature map into a localized anomaly map. The upsampling blocks (UpB, Up1, Up2, Up3) double the spatial resolution of the input feature maps while refining the features using 3×3 convolutions. For instance, UpB increases the resolution from

$(B, 1024, 16, 16)$ to $(B, 1024, 32, 32)$. These upsampling blocks are complemented by dual-block layers (DBB, DB1, DB2, DB3, DB4) that combine 3×3 convolutions and BatchNorm layers to refine features at different resolutions. For example, DBB processes combined features from UpB and Conv5 to produce refined outputs at $(B, 1024, 32, 32)$.

The final output layer is a 3×3 convolution that reduces the number of channels to 1, producing an anomaly map with a shape of $(B, 1, 256, 256)$, matching the input image resolution. This anomaly map shows possible defects in the input image.

The complete network is a combination of the encoder and decoder. The encoder generates multi-scale features, which are used by the decoder to generate the final anomaly map. This table summarizes the entire architecture of the network and shows how features are extracted, processed, and reconstructed at each stage.

Algorithm 2 describes the step-by-step process of ViLoc-Net. It combines Vision Transformers with a multi-scale decoder to effectively segment and locate anomalies.

Algorithm 2: ViLoc-Net: Vision transformer with localization network for anomaly segmentation and localization

Require: Input image I_{input} , anomalous image A_{RIDS} from RIDS, patch size P , embedding dimension D , and number of transformer layers L

Ensure: Refined anomalous image I_{refined} , binary anomaly mask M_{final}

1: **Vision Transformer-Based Encoder:**

2: Divide $I_{\text{input}} \in \mathbb{R}^{H \times W \times C}$ into $N = \frac{H \cdot W}{P^2}$ patches of size $P \times P$.

3: Embed patches into z_0 using Eq. (14).

4: Process z_0 through L transformer layers to extract global features as described in Eqs. (15) and (16).

5: Reshape z_L to $f_{\text{ViT}} \in \mathbb{R}^{D \times \sqrt{N} \times \sqrt{N}}$ (Eq. (17)).

6: Generate multi-scale features:

$$f_{\text{multi}} = \{b_1, b_2, \dots, b_6\}, \quad b_i = \text{Conv}_i(f_{\text{ViT}}),$$

as per Eq. (18).

7: **Multi-Scale Decoder:**

8: **Initial Upsampling:** Process the most abstract feature map b_6 using Eq. (19).

9: **Intermediate Stages:** Sequentially integrate b_4, b_3 , and b_2 as described in Eq. (20).

10: **Final Stage:** Fuse b_1 with b'_2 as described in Eq. (21).

11: Generate outputs:

$$M_{\text{final}}, I_{\text{refined}} = \mathcal{D}(f_{\text{multi}}, A_{\text{RIDS}})$$

as described in Eq. (24).

12: **Output:** Return I_{refined} (refined anomalous image) and M_{final} (binary anomaly mask).

4 Experiments and Results

The model was trained using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 8. Binary cross-entropy loss was used for anomaly segmentation. All experiments were conducted using an NVIDIA RTX GPU. We evaluated the performance of ViLoc-Net on the MVTec [3] and Metal Parts Defect Detection (MPDD) [4] datasets against several state-of-the-art anomaly detection methods. The MPDD dataset serves as a benchmark for defect detection in metal parts with over 1000 annotated images, while the MVTec dataset focuses on industrial anomaly detection with more than 5000 images across 15 categories, including both defective and defect-free samples. The proposed model was trained for 1000 epochs on each dataset independently using the official training and testing splits. Specifically, MVTec AD and MPDD experiments were conducted separately without cross-dataset training. The RIDS module was applied only

during training to generate synthetic anomalous samples from normal images and was not used during inference or evaluation. Image-level anomaly detection performance was measured using the Area Under the Receiver Operating Characteristic (AUROC) metric, while Pixel AUROC was used for evaluating pixel-level anomaly localization.

The proposed ViLoc-Net is compared with several state-of-the-art methods, including SPADE [33], P-SVDD [34], CutPaste [9], PaDiM [13], DRAEM [19], CDO [35], and SLGS [36]. Table 2 highlights the superior performance of ViLoc-Net in both image-level and pixel-level anomaly detection on the MVTec dataset, achieving the highest mean AUROC scores of 99.67 and 97.86, respectively. Some baseline methods employ different backbones, image resolutions, or evaluation protocols as reported in their original publications. Therefore, the reported results are reproduced directly from the corresponding papers for reference-based comparison. Missing entries indicate results that were not reported by the original authors.

Table 2: Comparison of AUROC results for Image-level and pixel-level anomaly detection on the MVTec dataset, including ViLoc-Net (Proposed) and other state-of-the-art methods.

Class	SPADE	P-SVDD	CutPaste	PaDiM	DRAEM	CDO	SLGS	ViLoc-Net
Carpet	(-, 97.5)	(93.1, 98.3)	(92.9, 92.6)	(-, 98.9)	(97.0, 95.5)	(99.1, -)	(99.0, 96.0)	(99.97, 99.82)
Grid	(-, 93.7)	(99.9, 97.5)	(94.6, 96.2)	(-, 94.9)	(99.9, 99.7)	(98.4, -)	(100, 98.5)	(100, 99.99)
Leather	(-, 97.6)	(100, 99.5)	(90.9, 97.4)	(-, 99.1)	(100, 98.6)	(99.2, -)	(100, 99.5)	(100, 99.73)
Tile	(-, 87.4)	(93.4, 90.5)	(97.8, 91.4)	(-, 91.2)	(99.6, 99.2)	(97.2, -)	(100, 98.6)	(100, 98.91)
Wood	(-, 85.5)	(98.6, 95.5)	(96.5, 90.8)	(-, 93.6)	(99.1, 96.4)	(95.9, -)	(99.6, 96.8)	(99.99, 96.34)
Bottle	(-, 98.4)	(98.3, 97.6)	(98.6, 98.1)	(-, 98.1)	(99.2, 99.1)	(99.3, -)	(99.4, 99.1)	(100, 99.95)
Cable	(-, 97.2)	(80.6, 90.0)	(90.3, 96.8)	(-, 95.8)	(91.8, 94.7)	(97.6, -)	(98.3, 97.4)	(99.37, 96.72)
Capsule	(-, 99.0)	(96.2, 97.4)	(76.7, 95.8)	(-, 98.3)	(98.5, 94.3)	(98.6, -)	(95.5, 95.9)	(99.43, 98.72)
Hazelnut	(-, 99.1)	(97.3, 97.3)	(92.0, 97.5)	(-, 97.7)	(100, 99.7)	(99.2, -)	(99.5, 97.8)	(100, 99.83)
Metal Nut	(-, 98.1)	(99.3, 93.1)	(94.0, 98.0)	(-, 96.7)	(98.7, 99.5)	(98.5, -)	(100, 98.9)	(100, 99.61)
Pill	(-, 96.5)	(92.4, 95.7)	(86.1, 95.1)	(-, 94.7)	(98.9, 97.6)	(98.9, -)	(99.2, 98.0)	(99.96, 96.48)
Screw	(-, 98.9)	(86.3, 96.7)	(81.3, 95.7)	(-, 97.4)	(93.9, 97.6)	(99.0, -)	(89.1, 97.3)	(97.08, 91.21)
Toothbrush	(-, 97.9)	(98.3, 98.1)	(100, 98.1)	(-, 98.7)	(100, 98.1)	(98.9, -)	(100, 99.4)	(99.79, 97.68)
Transistor	(-, 94.1)	(95.5, 93.0)	(91.5, 97.0)	(-, 97.2)	(93.1, 90.9)	(95.3, -)	(97.3, 92.5)	(99.61, 97.23)
Zipper	(-, 96.5)	(99.4, 99.3)	(97.9, 95.1)	(-, 98.2)	(100, 98.8)	(98.2, -)	(100, 97.1)	(99.82, 95.69)
Mean	(85.5, 96.0)	(92.1, 95.7)	(95.2, 96.0)	(95.3, 96.7)	(98.0, 97.3)	(98.2, -)	(98.5, 97.5)	(99.67, 97.86)

The improvement achieved by the proposed method is evident across various defect classes, with ViLoc-Net consistently outperforming or matching the best results. The qualitative results on the MVTec dataset are presented in Fig. 6.

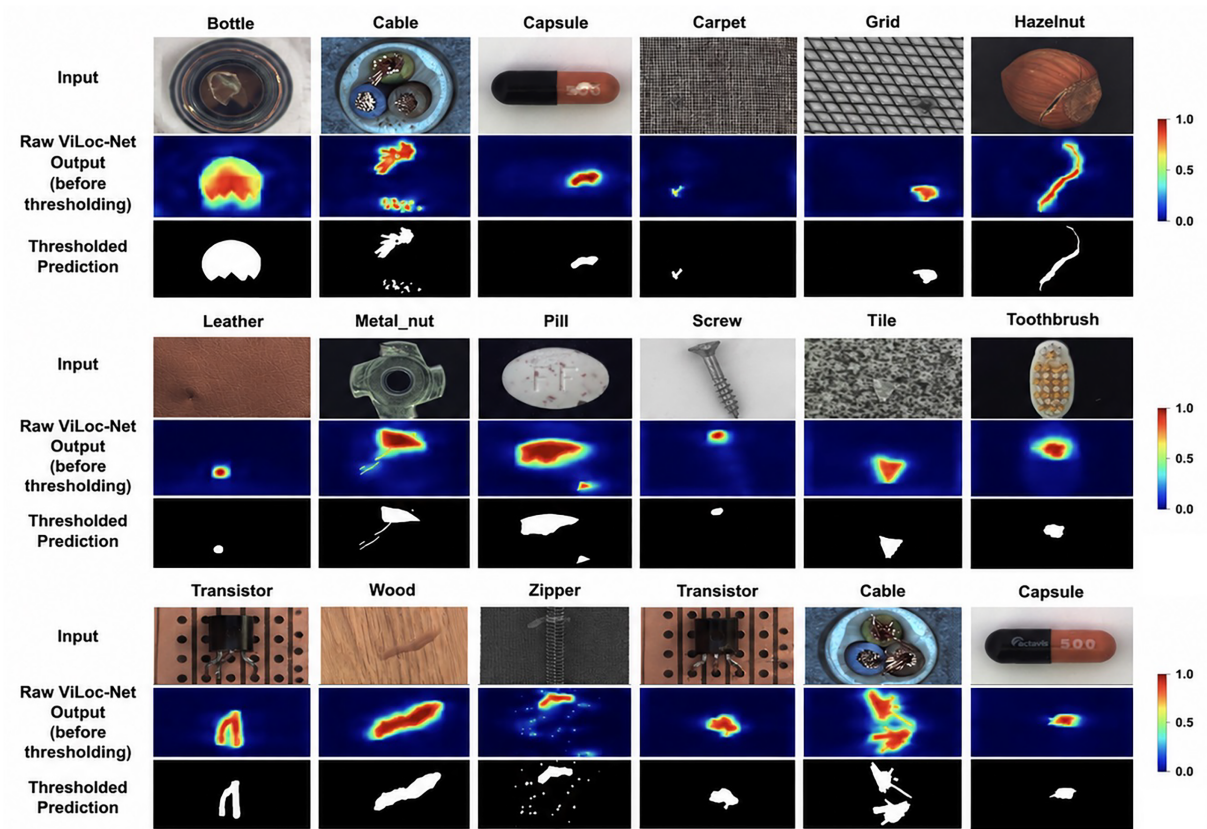


Figure 6: Results of anomaly detection using the proposed method, showcasing accurate identification of anomalous regions.

Table 3 presents a comprehensive comparison of AUROC results for Image-level and Pixel-level anomaly detection on the MPDD dataset. The results demonstrate that ViLoc-Net (Proposed) achieves superior performance in most defect categories compared to state-of-the-art methods, including RealNet [32], DTD [37], and CutPaste [9]. This highlights the robustness and effectiveness of ViLoc-Net in accurately detecting and localizing anomalies across various defect types, further emphasizing its potential for practical deployment in industrial applications.

Table 3: AUROC results on the MPDD dataset for Image-level and pixel-level anomaly detection, comparing ViLoc-Net (Proposed) with RealNet [32], DTD [37], and CutPaste [9].

Class	RealNet	DTD	CutPaste	ViLoc-Net
Bracket Black	(94.95, 99.27)	(89.49, 98.90)	(66.42, 96.67)	(92.74, 94.64)
Bracket Brown	(96.83, 97.81)	(92.99, 97.35)	(95.48, 97.54)	(97.86, 99.39)
Bracket White	(88.78, 97.44)	(86.67, 98.59)	(88.44, 96.51)	(93.21, 97.38)
Connector	(100.0, 97.46)	(99.05, 97.76)	(99.05, 98.47)	(100.00, 99.03)
Metal Plate	(100.0, 99.28)	(100.00, 99.35)	(99.95, 98.83)	(99.01, 99.25)
Tubes	(97.51, 97.94)	(92.62, 99.01)	(91.49, 98.09)	(97.89, 98.91)
Mean	(96.35, 98.20)	(93.47, 98.49)	(90.14, 97.69)	(96.79, 98.10)

Fig. 7 shows a comparison of image-level AUROC scores with radar charts for different methods, including ViLoc-Net, on two datasets: (a) MVTec and (b) MPDD. This highlights ViLoc-Net's steady performance across various industrial scenarios.

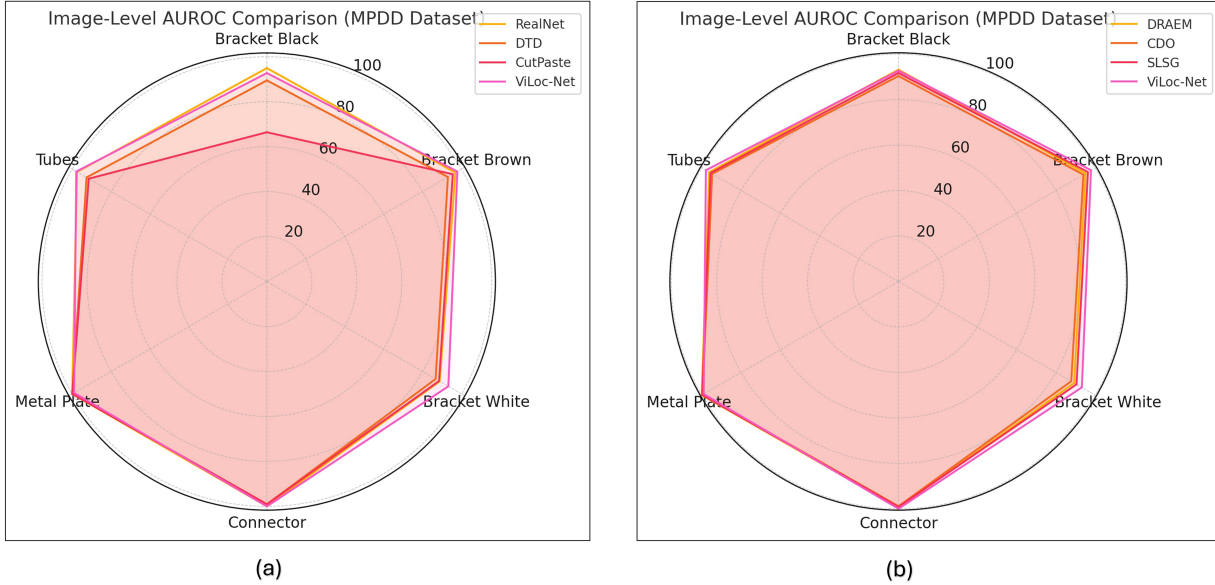


Figure 7: Image-level AUROC comparison of DRAEM, CDO, SLSG, and ViLoc-Net on (a) MVTec and (b) MPDD datasets.

4.1 Ablation Study

To better understand the contribution of each component of the proposed framework, we conducted a series of ablation experiments on the MVTec AD dataset. Since ViLoc-Net integrates multiple design choices, including synthetic anomaly generation, transformer-based encoding, and multi-scale feature fusion, the ablation analysis aims to isolate the impact of each component under a consistent experimental setting. Unless otherwise stated, all ablation variants were trained using the same optimizer, image resolution, batch size, and training schedule to ensure fair comparison.

4.1.1 Effect of the RIDS Module

The proposed RIDS module is designed as a training-time synthetic anomaly generation mechanism and is not used during inference. To evaluate its contribution, we compared the full ViLoc-Net framework with two alternative training configurations:

- **Without RIDS:** The model was trained without RIDS-generated anomalous samples. Only normal images and standard image augmentations were used during training.
- **Simple Synthetic Anomalies:** The model was trained using basic synthetic anomalies generated through simple geometric masking and Gaussian noise injection instead of the proposed RIDS pipeline.
- **Full RIDS (Proposed):** The complete RIDS module using foreground masks, random shape masks, Perlin noise, DTD textures, SDAS structural defects, and blending operations.

As illustrated in Fig. 8, the simple synthetic anomaly baseline generates artificial defects using basic geometric masks filled with Gaussian noise and blended into normal images. Unlike the proposed RIDS framework, these anomalies lack realistic structural and texture characteristics commonly observed in industrial defect patterns.

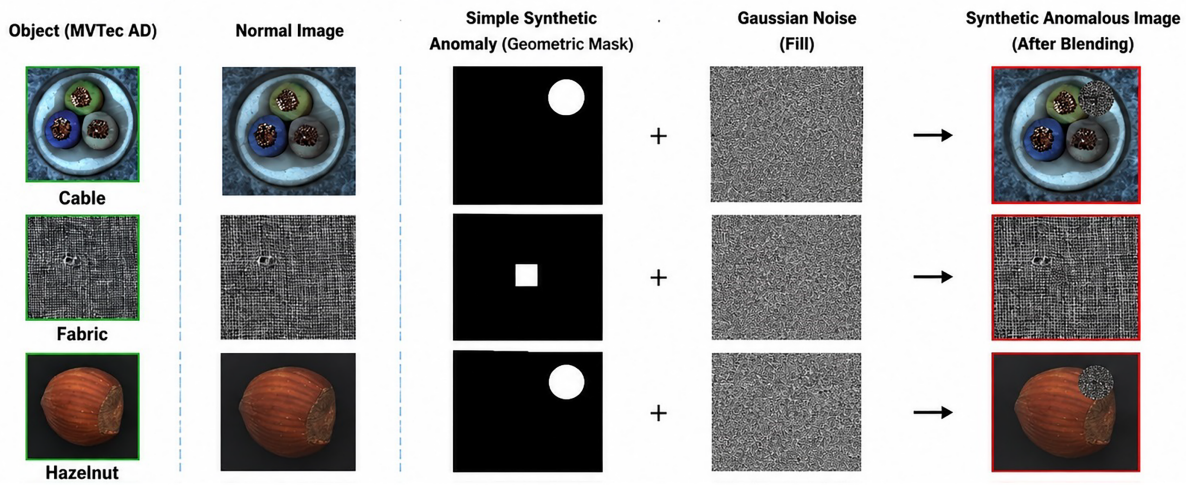


Figure 8: Examples of simple synthetic anomalies generated using geometric masks and Gaussian noise for the ablation study.

Table 4 presents the comparison results.

Table 4: Effect of the RIDS module on anomaly detection performance.

Training Configuration	Image AUROC	Pixel AUROC
Without RIDS	88.73	81.46
Simple Synthetic Anomalies	94.28	89.91
Full RIDS (Proposed)	99.67	97.86

The results indicate that the proposed RIDS module significantly improves both image-level and pixel-level anomaly detection performance. While simple synthetic anomalies provide moderate improvement compared to training without anomaly synthesis, the proposed RIDS framework achieves superior localization accuracy due to its ability to generate structurally and texturally realistic defect patterns.

4.1.2 CNN Encoder vs. Vision Transformer Encoder

To evaluate how well the Vision Transformer (ViT) encoder works, we replaced the transformer backbone with different encoder types, both CNN and transformer-based, while keeping the decoder the same. We used a ResNet18 encoder as the CNN baseline and tested various Vision Transformer configurations to see how the transformer architecture and patch size affected the results. The comparison results are shown in Table 5.

The Vision Transformer-based encoders showcased better performance than the CNN-based encoder, particularly in the localization of anomalies at the pixel level. ViT-Base/16 was the top-performing variant among those evaluated. This suggests that higher resolution of the patch representations and more robust global context modeling result in higher accuracy in anomaly segmentation. The results indicate that the feature extraction using transformers is more effective in capturing irregular and spatially-distributed industrial defect patterns than the traditional CNN representations.

Table 5: Comparison between CNN and vision transformer encoder variants.

Encoder Type	Image AUROC	Pixel AUROC
ResNet18 Encoder	98.14	94.92
ViT-Small/16	98.63	95.71
ViT-Base/32	98.91	96.18
ViT-Base/16 (Proposed)	99.67	97.86

4.1.3 Effect of Multi-Scale Decoder Fusion

The proposed decoder uses skip connections and combines features at different scales to keep spatial details when locating anomalies. To assess how this design helps, we tested a simpler version of the decoder that lacks skip connections and hierarchical multi-scale fusion. The comparison results are in [Table 6](#).

Table 6: Effect of multi-scale feature fusion in the decoder.

Decoder Configuration	Image AUROC	Pixel AUROC
Without Multi-Scale Fusion	97.83	92.64
With Multi-Scale Fusion (Proposed)	99.67	97.86

Removing multi-scale fusion noticeably reduced anomaly localization performance, particularly at the pixel level. The decrease in Pixel AUROC indicates that skip connections and hierarchical feature integration are important for preserving fine spatial details and accurately segmenting small or irregular defects. In contrast, the proposed multi-scale decoder effectively combines high-level semantic information with low-level spatial features, resulting in improved localization accuracy.

4.1.4 Effect of Anomaly Source Selection

The proposed RIDS framework combines texture-based anomalies from the Describable Textures Dataset (DTD) and structural anomalies from the SDAS defect library. To assess the contribution of each anomaly source, we did experiments with DTD-only anomalies, SDAS-only anomalies, and their combined configuration. The comparison results are shown in [Table 7](#).

Table 7: Effect of different anomaly source configurations.

Anomaly Source	Image AUROC	Pixel AUROC
DTD Only	98.42	95.11
SDAS Only	98.76	95.94
DTD + SDAS (Proposed)	99.67	97.86

The combined anomaly source configuration achieved the best overall performance. DTD-based anomalies improved robustness against texture irregularities, while SDAS-based anomalies enhanced the detection of structured defects such as scratches and surface damage. Their combination enabled better generalization across diverse anomaly categories and improved both image-level detection and pixel-level localization performance.

Overall, the ablation experiments demonstrate that each component of the proposed framework contributes positively to the final performance. The combination of realistic synthetic anomaly generation, transformer-based global feature extraction, and multi-scale localization collectively enables robust anomaly detection and accurate defect localization.

5 Discussion

The proposed RIDS pipeline design is based on well-known image processing methods and industrial anomaly detection tasks assessment. Hyperparameters such as threshold values, blending coefficients, and strategies for placing anomalies were chosen. This not only aids in generating realistic fake defects but also ensures stable optimization during the training process.

Although a thorough robustness analysis was not performed, the stochastic design of the RIDS framework creates different synthetic anomalies with variations in shape, size, texture, location, and noise characteristics. Foreground masks are used to restrict anomaly synthesis to only valid object regions to improve semantic consistency. The proposed method does not explicitly model defect meaning for objects, but the experiment and ablation results indicate that the generated anomalies are realistic enough for improving generalization and localization performance. Future research will focus on a more systematic robustness assessment while systematically varying and changing the context.

To assess how well the proposed framework works, we also compare ViLoc-Net with recent transformer-based anomaly detection methods. The results of the comparisons are presented in [Table 8](#).

Table 8: AUROC results on the MVTec AD dataset for image-level and pixel-level anomaly detection, comparing ViLoc-Net with transformer-based methods including UVAD [38], VT-ADL, and NN2ViT [39].

Class	Bib38	VT-ADL	NN2ViT	ViLoc-Net
Carpet	(99.9, 99.0)	(-, 77.3)	(-, -)	(99.97, 99.82)
Grid	(100, 99.0)	(-, 87.1)	(-, -)	(100, 99.99)
Leather	(100, 99.5)	(-, 72.8)	(-, -)	(100, 99.73)
Tile	(99.9, 98.6)	(-, 79.6)	(-, -)	(100, 98.91)
Wood	(99.0, 95.5)	(-, 78.1)	(-, -)	(99.99, 96.34)
Bottle	(100, 98.3)	(-, 94.9)	(98.3, 95.1)	(100, 99.95)
Cable	(96.0, 88.3)	(-, 77.6)	(92.2, 97.3)	(99.37, 96.72)
Capsule	(95.9, 95.2)	(-, 67.2)	(91.2, 92.4)	(99.43, 98.72)
Hazelnut	(99.9, 98.7)	(-, 89.7)	(97.9, 93.2)	(100, 99.83)
Metal Nut	(100, 96.4)	(-, 72.6)	(94.3, 94.2)	(100, 99.61)
Pill	(98.9, 85.9)	(-, 70.5)	(89.4, 97.3)	(99.96, 96.48)
Screw	(98.1, 98.9)	(-, 92.8)	(94.3, 94.2)	(97.08, 91.21)
Toothbrush	(100, 98.3)	(-, 90.1)	(97.4, 98.9)	(99.79, 97.68)
Transistor	(98.1, 83.9)	(-, 79.6)	(93.1, 95.3)	(99.61, 97.23)
Zipper	(100, 99.2)	(-, 80.8)	(95.4, 96.7)	(99.82, 95.69)
Mean	(99.05, 95.65)	(-, 80.71)	(94.3, 95.46)	(99.67, 97.86)

The proposed ViLoc-Net method is capable of achieving excellent performance in anomaly detection at both the image level and pixel level. In particular, compared to state-of-the-art transformer-based models such as UVAD and NN2ViT, ViLoc-Net performs better in terms of mean AUROC, especially for pixel-level localization. This is attributed to the usage of a multi-scale decoder, which preserves fine-grained detail

information in the reconstruction phase, as well as the design of the RIDS block, which enhances data augmentation quality.

On the other hand, approaches based on shallow transformer representation or feature fusion exhibit relatively poor performance in terms of localization in hard defect classes. Moreover, the experiments conducted confirm that incorporating realistic synthetic anomaly data generation, transformer modeling of global context, and hierarchical multi-scale decoding can benefit anomaly segmentation and localization.

6 Conclusion

In this work, we presented a practical approach to the problem of detecting and locating industrial abnormalities, namely ViLoc-Net. It is the solution to the critical issues of rare and diverse anomaly identification. The approach is a hybrid of embedding-based, reconstruction-based, and self-supervised learning methods, which integrates the Realistic Industrial Defect Synthesis (RIDS) module with a Vision Transformer-based localization network. We have tested ViLoc-Net on multiple industrial datasets, such as MVTec and MPDD, and it has proven to be superior to other approaches for image- and pixel-level anomaly detection. ViLoc-Net is a flexible solution for automated quality inspection in real industrial applications, thanks to its capability of generating realistic anomalies and its capability of locating them. This work lays the groundwork for future developments of hybrid approaches combining synthetic data and better localization systems for anomaly detection.

Acknowledgement: None.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm their contributions to the paper as follows: study conception and design: Asim Niaz and Muhammad Umraiz; data collection: Asim Niaz and Muhammad Umraiz; analysis and interpretation of results: Asim Niaz, Muhammad Umraiz, and Syed Farhan Alam Zaidi; draft manuscript preparation: Asim Niaz, Muhammad Umraiz, and Kwang Nam Choi. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data available on request from the authors. The data that support the findings of this study are available from the corresponding author, Kwang Nam Choi, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ma W, Hou Y, Jin M, Jian P. Anomaly based multi-stage attack detection method. PLoS One. 2024;19(3):e0300821. doi:10.1371/journal.pone.0300821.
2. Zhou S, Lin C, Chen Z, Guo B, Zhao Y. AdapSyn: anomaly detection based on triplet training with adaptive anomaly synthesis. Displays. 2025;86:102885.
3. Bergmann P, Fauser M, Sattlegger D, Steger C. MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 9592–600.
4. Jezek S, Jonak M, Burget R, Dvorak P, Skotak M. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In: Proceedings of the 2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT); 2021 Oct 25–27; Virtual. p. 66–71.
5. Sheng S, Jing J, Wang Z, Zhang H. Cosine similarity knowledge distillation for surface anomaly detection. Sci Rep. 2024;14(1):8150. doi:10.1038/s41598-024-58409-9.

6. Park Y, Kang S, Kim MJ, Lee Y, Kim HS, Yi J. Visual defect obfuscation based self-supervised anomaly detection. *Sci Rep.* 2024;14(1):18872. doi:10.1038/s41598-024-69698-5.
7. Liu Z, Huo Z, Li C, Dong Y, Li B. DLSE-Net: a robust weakly supervised network for fabric defect detection. *Displays.* 2021;68:102008.
8. Gao Y, Lin Q, Ye S, Cheng Y, Zhang T, Liang B, et al. Outlier detection in temporal and spatial sequences via correlation analysis based on graph neural networks. *Displays.* 2024;84:102775. doi:10.2139/ssrn.4812287.
9. Li CL, Sohn K, Yoon J, Pfister T. Cutpaste: self-supervised learning for anomaly detection and localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2021 Jun 20–25; Nashville, TN, USA. p. 9664–74.
10. Du N, Huo Y, Wang D. A video anomaly detection method based on percentile loss training and attention mechanism. *Displays.* 2022;75:102327. doi:10.1016/j.displa.2022.102327.
11. Gu H, Li G, Liu Z. Masked feature regeneration based asymmetric student-teacher network for anomaly detection. *Multimed Tools Appl.* 2024;83(42):90573–94. doi:10.1007/s11042-024-18512-5.
12. Wan D, Gao C, Zhou J, Shen X, Shen L. Unsupervised fabric defect detection with high-frequency feature mapping. *Multimed Tools Appl.* 2024;83(7):21615–32. doi:10.1007/s11042-023-16340-7.
13. Defard T, Setkov A, Loesch A, Audigier R. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In: *International Conference on Pattern Recognition*. Berlin/Heidelberg, Germany: Springer; 2021. p. 475–89.
14. Zavrtanik V, Kristan M, Skočaj D. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognit.* 2021;112:107706. doi:10.1016/j.patcog.2020.107706.
15. Ristea NC, Madan N, Ionescu RT, Nasrollahi K, Khan FS, Moeslund TB, et al. Self-supervised predictive convolutional attentive block for anomaly detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2022 Jun 18–24; New Orleans, LA, USA. p. 13576–86.
16. Haselmann M, Gruber DP, Tabatabai P. Anomaly detection using deep learning based image completion. In: *Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*; 2018 Dec 17–20; Orlando, FL, USA. p. 1237–42.
17. Gong D, Liu L, Le V, Saha B, Mansour MR, Venkatesh S, et al. Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2019 Oct 27–30; Seoul, Republic of Korea. p. 1705–14.
18. Schlegl T, Seebock P, Waldstein SM, Langs G, Schmidt-Erfurth U. f-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks. *Med Image Anal.* 2019;54:30–44. doi:10.1016/j.media.2019.01.010.
19. Zavrtanik V, Kristan M, Skočaj D. DRAEM—a discriminatively trained reconstruction embedding for surface anomaly detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2021 Oct 11–17; Montreal, BC, Canada. p. 8330–9.
20. Roth K, Pemula L, Zepeda J, Schölkopf B, Brox T, Gehler P. Towards total recall in industrial anomaly detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2022 Jun 18–24; New Orleans, LA, USA. p. 14318–28.
21. Deng H, Li X. Anomaly detection via reverse distillation from one-class embedding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2022 Jun 18–24; New Orleans, LA, USA. p. 9737–46.
22. Gudovskiy D, Ishizaka S, Kozuka K. CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*; 2022 Jan 3–8; Waikoloa, HI, USA. p. 98–107.
23. Rudolph M, Wandt B, Rosenhahn B. Same same but different: semi-supervised defect detection with normalizing flows. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*; 2021 Jan 5–9; Virtual. p. 1907–16.
24. Zhang R, Wang H, Feng M, Liu Y, Yang G. JRCC-Net: a segmentation network with joint representation and contrast clustering for surface anomaly detection. *IEEE Trans Instrum Meas.* 2023;72:5020914.

25. Mei S, Yang H, Yin Z. An unsupervised-learning-based approach for automated defect inspection on textured surfaces. *IEEE Trans Instrum Meas.* 2018;67(6):1266–77. doi:10.1109/tim.2018.2795178.
26. Niaz A, Amin SU, Soomro S, Zia H, Choi KN. Spatially aware fusion in 3D convolutional autoencoders for video anomaly detection. *IEEE Access.* 2024;12:104770–84. doi:10.1109/access.2024.3487363.
27. Akçay S, Atapour-Abarghouei A, Breckon TP. Ganomaly: semi-supervised anomaly detection via adversarial training. In: *Proceedings of the Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision*; 2018 Dec 2–6; Perth, Australia. p. 622–37.
28. Akçay S, Atapour-Abarghouei A, Breckon TP. Skip-ganomaly: skip connected and adversarially trained encoder-decoder anomaly detection. In: *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*; 2019 Jul 14–19; Budapest, Hungary. p. 1–8.
29. Ristea NC, Croitoru FA, Ionescu RT, Popescu M, Khan FS, Shah M, et al. Self-distilled masked auto-encoders are efficient video anomaly detectors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2024 Jun 17–21; Seattle, WA, USA. p. 15984–95.
30. Pirnay J, Chai K. Inpainting transformer for anomaly detection. In: *International Conference on Image Analysis and Processing*. Berlin/Heidelberg, Germany: Springer; 2022. p. 394–406.
31. Smith AD, Du S, Kurien A. Vision transformers for anomaly detection and localisation in leather surface defect classification based on low-resolution images and a small dataset. *Appl Sci.* 2023;13(15):8716. doi:10.3390/app13158716.
32. Zhang X, Xu M, Zhou X. RealNet: a feature selection network with realistic synthetic anomaly for anomaly detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2024 Jun 17–21; Seattle, WA, USA. p. 16699–708.
33. Cohen N, Hoshen Y. Sub-image anomaly detection with deep pyramid correspondences. *arXiv:2005.02357.* 2020.
34. Yi J, Yoon S. Patch SVDD: patch-level SVDD for anomaly detection and segmentation. In: *Proceedings of the Asian Conference on Computer Vision*. Berlin/Heidelberg, Germany: Springer; 2020. p. 830–46.
35. Cao Y, Xu X, Liu Z, Shen W. Collaborative discrepancy optimization for reliable image anomaly localization. *IEEE Trans Ind Inform.* 2023;19(11):10674–83. doi:10.1109/tii.2023.3241579.
36. Yang M, Liu J, Yang Z, Wu Z. SLSG: industrial image anomaly detection with improved feature embeddings and one-class classification. *Pattern Recognit.* 2024;156:110862.
37. Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A. Describing textures in the wild. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2014 Jun 23–28; Columbus, OH, USA. p. 3606–13.
38. Kim JH, Kwon GR. Unsupervised visual anomaly detection using self-supervised pre-trained transformer. *IEEE Access.* 2024;12:127604–13. doi:10.1109/access.2024.3454753.
39. Wahid JA, Ayoub M, Xu M, Jiang X, Shi L, Hussain S. NN2ViT: neural networks and vision transformers based approach for visual anomaly detection in industrial images. *Neuro Comput.* 2025;615:128845. doi:10.1016/j.neucom.2024.128845.