Revisiting Deep Feature Reconstruction for Logical and Structural Industrial Anomaly Detection

Anonymous authors
Paper under double-blind review

Abstract

Industrial anomaly detection is crucial for quality control and predictive maintenance but is challenging due to limited training data, varied anomaly types, and changing external factors affecting object appearances. Existing methods detect structural anomalies, such as dents and scratches, by relying on multi-scale features of image patches extracted from a deep pre-trained network. Nonetheless, extensive memory or computing requirement hinders their adoption in practice. Furthermore, detecting logical anomalies, such as images with missing or surplus elements, necessitates understanding spatial relationships beyond traditional patch-based methods. Our work focuses on Deep Feature Reconstruction (DFR), which offers a memory- and compute-efficient way of detecting structural anomalies. Moreover, we extend DFR to develop a unified framework for detecting structural and logical anomalies, called `ULSAD`. Specifically, we improve the training objective of DFR to enhance the capability to detect structural anomalies and introduce an attention-based loss using a global autoencoder-like network for detecting logical anomalies. Empirical results on five benchmark datasets demonstrate the effectiveness of `ULSAD` in the detection and localization of both structural and logical anomalies compared to eight state-of-the-art approaches. Moreover, an in-depth ablation study showcases the importance of each component in enhancing overall performance. Our code can be accessed here: https://anonymous.4open.science/r/ULSAD-2024.

1 Introduction

Anomaly detection (AD) is a widely studied problem in machine learning that is used to identify rare events or unusual patterns (Salehi et al., 2022). It enables the detection of abnormalities, potential threats, or critical system failures across diverse applications such as predictive maintenance (PdM) (Tang et al., 2020; Choi et al., 2022), fraud detection (Ahmed et al., 2016; Hilal et al., 2022), and medicine (Tibshirani & Hastie, 2007; Fernando et al., 2021). Despite its importance and widespread applicability, it remains a challenging task as the anomalous samples are not known a priori (Ruff et al., 2021). Typically, AD is therefore addressed as an unsupervised representation learning problem (Pang et al., 2020) where the training data contains predominantly normal samples. Therefore, the aim is to learn the normal behaviour using the samples in the training set and identify anomalies as deviations from this normal behaviour. This setting is also known as one-class classification (Ruff et al., 2018).

Our study concentrates on Industrial Anomaly Detection (IAD) (Bergmann et al., 2019), specifically targeting the detection and localization of anomalies in images from industrial manufacturing processes. Over the years, it has garnered attention in both industry and academia as AD can be used for various tasks like quality control or predictive maintenance, which are of primal interest to industries. Despite the necessity, addressing IAD is difficult given the following challenges: (i) evolving processes resulting in different manifestations of anomaly (Gao et al., 2023), and (ii) varying object appearances due to external factors such as background, lighting conditions and orientation (Jezek et al., 2021). Furthermore, the anomalies in IAD can be broadly categorized into: (i) **structural anomalies** where subtle localized structural defects can be observed (Bergmann et al., 2019), and (ii) **logical anomalies** where violations of logical constraints result in anomalies (Bergmann et al., 2022). Examples of both structural and logical anomalies can be seen in Figure 1.

Methods proposed for the detection of **structural anomalies** leverage multi-scale features of image patches obtained using deep convolutional neural networks (Salehi et al., 2021). PatchCore (Roth et al., 2022) and

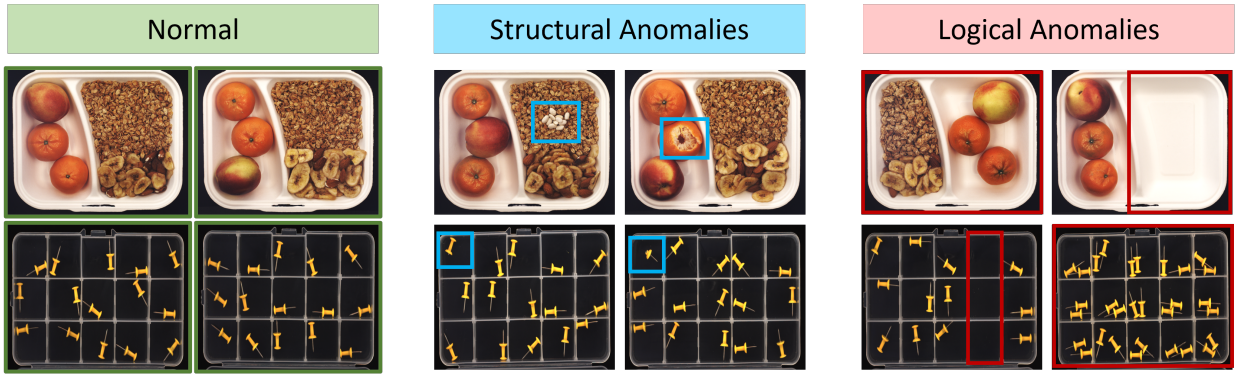


Figure 1: Types of anomalies. (Left) First, two normal samples are shown from the categories "breakfast box" and "pushpin" of the MVTecLOCO dataset. Then, we show examples of structural (Middle) and logical anomalies (Right) with the anomalies highlighted in blue and red, respectively.

CFA(Lee et al., 2022) achieved state-of-the-art (SOTA) performance by storing the extracted features in a memory bank and comparing the features of the test image with their closest neighbour from the memory bank. However, such approaches require considerable storage to accumulate the extracted features, which can be challenging for large-scale datasets. The first alternative is knowledge distillation-based approaches (Bergmann et al., 2020; 2022), where a student network is trained to mimic the teacher for normal samples. During inference, anomalies are identified based on the discrepancy between the student and teacher output. A key requirement of these distillation-based approaches is that the student network must be less expressive than the teacher to prevent it from mimicking the teacher on anomalous samples. Thus, regularization methods, such as penalty based on an external dataset or hard-mining loss (Batzner et al., 2024) are applied, which slows down the training and increases the requirement of computing resources. Moreover, excessive regularization also prevents learning representations for normal images. The second alternative is to model the features of normal images using a multivariate Gaussian distribution (Defard et al., 2021) or learn to reconstruct the features using a deep feature reconstruction (DFR) network (Yang et al., 2020).

Besides structural anomalies, **logical anomalies** occur when elements in the images are missing, misplaced, in surplus or violate geometrical constraints (Bergmann et al., 2022). Methods relying on multi-scale features of image patches would fail as they would still be considered normal. It is the combination of objects in the image that makes the image anomalous. Thus, to detect such logical anomalies, it is necessary to look beyond image patches and develop a global understanding of the spatial relationships within normal images. Distillation-based methods, which are predominantly used for the detection of logical anomalies, rely on an additional network to learn the spatial relationships between items in the normal image (Batzner et al., 2024).

In this paper, we focus on DFR, the benefits of which are four-fold. First, it does not need large memory for storing the features, unlike PatchCore (Roth et al., 2022). Second, unlike PaDiM (Defard et al., 2021), it does not make any assumption about the distribution of features. Third, learning to reconstruct features in the latent space of a pre-trained network is less impacted by the curse of dimensionality than learning to reconstruct high-dimensional images. Fourth, deep networks trained to reconstruct normal images using the per-pixel distance suffer from the loss of sharp edges of the objects or textures in the background. As a consequence, AD performance deteriorates due to an increase in false positives, i.e., the number of normal samples falsely labelled anomalies. On the contrary, using the distance between the multi-scale features and their corresponding reconstructions during training is less likely to result in such errors (Assran et al., 2023).

We revisit DFR to develop a unified framework for the detection of structural and logical anomalies. First, to improve the performance of DFR for the detection of structural anomalies, we modify the training objective by considering a combination of ℓ_2 and cosine distances between each feature and the corresponding reconstruction. The incorporation of the cosine distance addresses the curse of dimensionality, where high-dimensional features become uniformly distant from each other in Euclidean space (Aggarwal et al., 2001). Second, to allow for the detection of logical anomalies, we introduce an attention-based loss using a global autoencoder-like network. We empirically demonstrate that with our proposed improvements, not only do

the detection and localization capabilities of DFR improve for structural anomalies, but also it delivers competitive results on the detection of logical anomalies. Our contributions can be summarized as:

- We propose **U**nified framework for **L**ogical and **S**tructural **A**nomaly **D**etection referred as `ULSAD`, a framework for identifying and localizing both structural and logical anomalies building on DFR.
- We improve DFR for structural AD by considering the angular difference between the extracted and reconstructed feature vectors in addition to the difference in values.
- For learning the logical constraints, we propose a novel attention-based loss for learning the features of an autoencoder which improves the detection and localization of logical anomalies during inference.
- We demonstrate the effectiveness of `ULSAD` by comparing it with 8 SOTA methods across 5 widely adopted IAD benchmark datasets.
- Through extensive ablation study, we show the effect of each component of `ULSAD` on the overall performance of the end-to-end architecture.

2 Related Work

Several methods have been proposed over the years for addressing Industrial AD (Bergmann et al., 2019; 2022; Jezek et al., 2021). They can be broadly categorized into feature-embedding based and reconstruction-based methods. We briefly highlight some relevant works in each of the category. For an extended discussion on the prior works we refer the readers to the survey by Liu et al. (2024).

Feature Embedding-based There are mainly three different types of IAD methods which utilize feature embeddings from a pre-trained deep neural network: memory bank (Defard et al., 2021; Roth et al., 2022; Lee et al., 2022), student-teacher (Zhang et al., 2023; Batzner et al., 2024), and density-based (Gudovskiy et al., 2021; Yu et al., 2021). The main idea of memory bank methods is to extract features of nominal images and store them in a memory bank during the training phase. During the testing phase, the feature of a test image is used as a query to match the stored nominal features. There are two main constraints in these methods: *how to learn useful features* and *how to reduce the size of the memory bank*. PatchCore (Roth et al., 2022) introduces a coreset selection algorithm to reduce the memory bank size. CFA (Lee et al., 2022) proposes contrastive supervision based on a coupled hypersphere to learn target-oriented features and a compression scheme to reduce memory size. The performance of the memory bank methods heavily depends on the completeness of the memory bank, which requires a large number of nominal images. Moreover, the memory size is often related to the number of training images, which makes these methods not preferable for large datasets or very high-dimensional images. In the student-teacher approach, the student network learns to extract features of the nominal samples, similar to the teacher model. For anomalous images, the features extracted by the student network should be different from the teacher network. Batzner et al. (2024) propose to use an autoencoder model in addition to the student network to identify logical anomalies. For leveraging the multiscale feature from the teacher network to detect anomalies at various scales, Deng & Li (2022) propose a reverse-distillation approach. Zhang et al. (2023) extended it by proposing to utilize two student networks to deal with structural and logical anomalies. Yang et al. (2020) propose to learn a deep neural network for learning to reconstruct the features of the normal images extracted using the pre-trained backbone. On the other hand, density-based methods detect anomalies based on the likelihood of an extracted feature of the test sample, given the estimated distribution of the features of the normal samples. PaDiM (Defard et al., 2021) uses a multivariate Gaussian to learn the probabilistic representation of the nominal class while FastFlow (Yu et al., 2021) and CFLOW (Gudovskiy et al., 2021) utilize normalizing flows.

Reconstruction-based methods. Reconstruction-based methods assume that encoder-decoder models trained on normal samples will exhibit poor performance for anomalous samples. However, relying solely on the reconstruction objective can result in the model collapsing to an identity mapping. To address this, structural assumptions are made regarding the data generation process. One such assumption is the Manifold Assumption, which posits that the observed data resides in a lower-dimensional manifold within the data space. Methods leveraging this assumption impose a bottleneck by restricting the encoded space to a lower dimensionality than the actual data space. Common deep reconstruction models used include

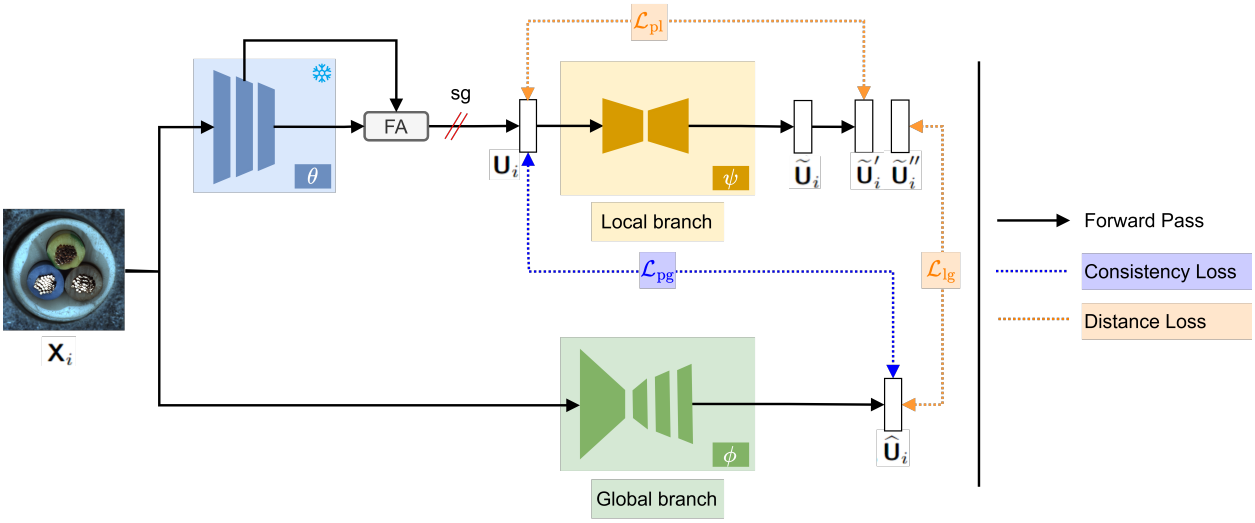


Figure 2: Overview of the end-to-end architecture of ULSAD

AE or VAE-based approaches. Advanced strategies encompass techniques like reconstruction by memorised normality (Gong et al., 2019), model architecture adaptation (Lai et al., 2019) and partial/conditional reconstruction (Yan et al., 2021; Nguyen et al., 2019). Generative models like GANs are also widely employed for anomaly detection, as the discriminator inherently calculates reconstruction loss for samples (Zenati et al., 2018). Variants of GANs, such as denoising GANs (Sabokrou et al., 2018) and class-conditional GANs (Perera et al., 2019), improve anomaly detection performance by increasing the challenge of reconstruction. Some methods utilize the reconstructed data from GANs in downstream tasks to enhance the amplification of reconstruction errors for anomaly detection (Zhou et al., 2020). Lastly, DRÆM (Zavrtanik et al., 2021) trains an additional discriminative network alongside a reconstruction network to improve the AD performance.

In this paper, we focus on feature embedding-based methods motivated by their effectiveness in the current SOTA methods. Specifically, we build on DFR (Yang et al., 2020), which has several benefits. First, it is memory-efficient as it does not rely on a memory bank of extracted features, unlike (Roth et al., 2022). Second, unlike (Defard et al., 2021), it does not make any assumption about the distribution of the extracted features. Third, it is computationally efficient and less impacted by the curse of dimensionality as it operates in the lower-dimensional latent space of a deep neural network. Last, by avoiding the use of per-pixel distance in its reconstruction objective, it is less prone to false positives (Assran et al., 2023).

3 The ULSAD Framework for Anomaly Detection

We propose ULSAD, a framework for simultaneously detecting and localising anomalies in images as shown in Figure 2. Firstly, we utilize a feature extractor network for extracting low-dimensional features from high-dimensional images, which we discuss in Section 3.1. Then, for the detection of structural and logical anomalies, we rely on a dual-branch architecture. The local branch detects structural anomalies with the help of a feature reconstruction network applied to the features corresponding to patches in the input image. We elaborate on this in Section 3.2. Conversely, the global branch, as discussed in Section 3.3, detects logical anomalies using an autoencoder-like network, which takes as input the entire image. Lastly, we provide an overview of the training algorithm of ULSAD in Section 3.4 followed by a discussion on the inference process in Section 3.5.

We consider a dataset $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{n}$ with n samples where $\mathbf{X}_i \in \mathcal{X}$ is an input and $y_i \in \mathcal{Y} := \{0, 1\}$ is the corresponding label. It is important to note that for this work, we specifically focus on image-based AD where each input $\mathbf{X}_i$ is an image. We refer to the normal class with the label 0 and the anomalous class with the label 1. The samples belonging to the anomalous class can contain either logical or structural anomalies or a combination of both. We denote the disjoint train and test partitions of $\mathcal{D}$ as $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$. The training set contains only normal samples, i.e., $y_i = 0$. Therefore, for the sake of simplicity, we refer to it as $\mathcal{D}_N = \{\mathbf{X}_i \mid (\mathbf{X}_i, y_i) \in \mathcal{D}_{\text{train}}\}$. The test set $\mathcal{D}_{\text{test}}$ includes both normal and anomalous samples.

3.1 Feature Extractor

High-dimensional images pose a significant challenge for AD (Reiss et al., 2022). Recent studies have shown that deep convolutional neural networks (CNNs) trained on ImageNet (Russakovsky et al., 2015) capture discriminative features for several downstream tasks. Typically, AD methods (Salehi et al., 2021; Defard et al., 2021; Yoon et al., 2023) leverage such pre-trained networks to extract features corresponding to partially overlapping regions or patches in the images. A key factor determining the efficiency of such networks is the size of the image patches being used, as anomalies can occur at any scale. To overcome this challenge, features are extracted from multiple layers of the CNNs and fused together (Salehi et al., 2021; Roth et al., 2022; Yang et al., 2020). As features from each layer have a different receptive field, each element of the extracted feature would correspond to a patch of a different size in the image. Thus, combining features from multiple layers results in multi-scale features of image patches, referred to as **patch feature**.

Similar to DFR, we extract low-dimensional feature maps by combining features from multiple layers of a pre-trained network θ. However, with the increasing number of layers, the computation becomes increasingly expensive as the resulting tensor becomes high-dimensional. In order to overcome this, we consider two intermediate or mid-level features. Our choice is guided by the understanding that the initial layers of such deep networks capture generic image features, while the latter layers are often biased towards the pre-training classification task (Roth et al., 2022). We denote the features extracted at a layer j for an image $\mathbf{X}_i$ as $\theta_j(\mathbf{X}_i)$. Following this convention, we express the feature map $\mathbf{U}_i \in \mathcal{U} = \mathbb{R}^{c^* \times h^* \times w^*}$ produced by the *Feature Aggregator* (FA) as a concatenation of $\theta_j(\mathbf{X}_i)$ and $\theta_{j+1}(\mathbf{X}_i)$ obtained from layers j and $j+1$ of θ. Furthermore, to facilitate the concatenation of features extracted from multiple layers of the extractor θ, the features at the lower resolution layer $j+1$ are linearly rescaled by FA to match the dimension of the features at layer j. Furthermore, we define an invertible transformation $f : \mathbb{R}^{c^* \times h^* \times w^*} \rightarrow \mathbb{R}^{c^* \times (h^* \times w^*)}$. Now, using f, we express $\boldsymbol{Z}_i = f(\mathbf{U}_i)$. Moreover, $\boldsymbol{Z}_i[:, k] = \mathbf{U}_i[:, h, w]$ with $k = (h-1) \times w^* + w$ where $h \in \{1, 2, \ldots, h^*\}$, $w \in \{1, 2, \ldots, w^*\}$, and $k \in \{1, 2, \ldots, h^* \times w^*\}$. The function f can be computed in practice by reshaping the tensor to obtain a 2D matrix.

3.2 Detecting Structural Anomalies

Given a transformed feature map $\boldsymbol{Z}_i$ as described in the previous section, we elaborate on the local branch of `ULSAD` for the detection of structural anomalies. Specifically, our goal is to train a neural network to reconstruct the patch features given the dataset $\mathcal{D}_N$ comprising solely of normal images. Therefore, we can identify the structural anomalies when the network fails to reconstruct a patch feature during inference.

We refer to each vector in the feature map $\boldsymbol{Z}_i$ as $\boldsymbol{z}_{ik} = \boldsymbol{Z}_i[:, k] \in \mathbb{R}^{c^* \times 1}$. Now, recall that θ is a CNN with multiple convolutional and pooling layers. Therefore, each c^*-dimensional feature vector $\boldsymbol{z}_{ik}$ of the output feature map $\boldsymbol{Z}_i$ has a receptive field greater than one. In other words, each feature vector $\boldsymbol{z}_{ik}$ corresponds to a patch on the original image $\mathbf{X}_i$ and hence can be considered as a patch feature. The size of the patch in the original image is determined by the receptive field of θ.

Feature Reconstruction Network (FRN). `ULSAD` utilizes a convolutional encoder-decoder architecture with a lower-dimensional bottleneck for learning to reconstruct the feature map $\mathbf{U}_i$ given the normal images in training dataset $\mathcal{D}_N$, as illustrated in Figure 3. First, the encoder network ψ_e compresses the feature $\mathbf{U}_i$ to a lower dimensional space, which induces the information bottleneck. It acts as an implicit regularizer, preventing generalization to features corresponding to anomalous images. The encoded representation is then mapped back to the latent space using a decoder network ψ_d. Therefore, we can express the FRN as $\psi = \psi_e \circ \psi_d$. The output of FRN is $\widetilde{\mathbf{U}}_i = \psi(\mathbf{U}_i) \in \mathbb{R}^{2c^* \times h^* \times w^*}$. Note that the number of channels is doubled to simultaneously generate two feature maps $\widetilde{\mathbf{U}}_i^{'}$ and $\widetilde{\mathbf{U}}_i^{''}$, both having dimension $c^* \times h^* \times w^*$. For the local objective, we utilize $\widetilde{\mathbf{U}}_i^{'}$, while $\widetilde{\mathbf{U}}_i^{''}$ is used in the global objective. Inspired by Salehi et al. (2021), we consider the difference between patch feature and their corresponding reconstruction considering both the

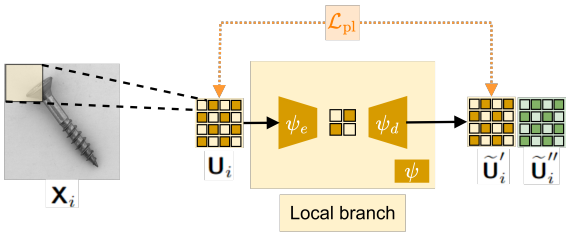


Figure 3: Feature Reconstruction Network

value and direction in the local loss, which is defined as

$$\mathcal{L}_{\text{pl}}(\widetilde{\boldsymbol{Z}}_i', \boldsymbol{Z}_i) = \frac{1}{h^* \times w^*} \sum_{k=1}^{h^* w^*} l_v(\widetilde{\boldsymbol{z}}_{ik}', \boldsymbol{z}_{ik}) + \lambda_l \, l_d(\widetilde{\boldsymbol{z}}_{ik}', \boldsymbol{z}_{ik}), \tag{1}$$

where $\widetilde{\boldsymbol{Z}}_i' = f(\widetilde{\mathbf{U}}_i')$, $\boldsymbol{Z}_i = f(\boldsymbol{Z}_i)$ and $\lambda_l \geq 0$ is a hyperparameter. Furthermore, $l_v(\widetilde{\boldsymbol{z}}_{ik}', \boldsymbol{z}_{ik})$ measures the difference in values between the vectors $\boldsymbol{z}_{ik}$ and $\widetilde{\boldsymbol{z}}_{ik}'$, while $l_d(\widetilde{\boldsymbol{z}}_{ik}', \boldsymbol{z}_{ik})$ measures the angular distance between them. We define l_v and l_d as

$$l_v(\widetilde{\boldsymbol{z}}_{ik}', \boldsymbol{z}_{ik}) = \frac{1}{c^*} \|\widetilde{\boldsymbol{z}}_{ik}' - \boldsymbol{z}_{ik}\|_2^2, \quad \text{and} \quad l_d(\widetilde{\boldsymbol{z}}_{ik}', \boldsymbol{z}_{ik}) = 1 - \frac{(\widetilde{\boldsymbol{z}}_{ik}')^T \boldsymbol{z}_{ik}}{\|\widetilde{\boldsymbol{z}}_{ik}'\|_2 \|\boldsymbol{z}_{ik}\|_2}. \tag{2}$$

3.3 Detecting Logical Anomalies

Although the feature reconstruction task discussed in Section 3.2 allows us to detect structural anomalies, it is not suited for identifying logical anomalies that violate the logical constraints of normal images. Recall that such violations appear in the form of misplaced, misaligned, or surplus objects found in normal images. If we consider the example of misaligned objects, the previously discussed approach will fail as it focuses on the individual image patches, which would be normal. It is the overall spatial arrangement of objects in the image which is anomalous. Thus, to identify such anomalies, our goal is to learn the spatial relationships among the objects present in the normal images of the training dataset $\mathcal{D}_N$. We achieve this with the global branch of `ULSAD`, shown in Figure 4, which leverages the entire image and not just its individual patches.

In order to achieve our goal, we start with an observation of the features extracted using the pre-trained network θ. Pre-trained CNNs tend to have similar activation patterns for semantically similar objects (Tung & Mori, 2019; Zagoruyko & Komodakis, 2017). In Figure 5, we visualize four attention maps computed from the features of a pre-trained Wide-Resnet50-2 network. It can be seen that in the first map, all the items for the semantic class "fruits" receive a high attention score. The remaining attention maps focus on individual semantic concepts like "oranges", "cereal" and "plate", respectively.

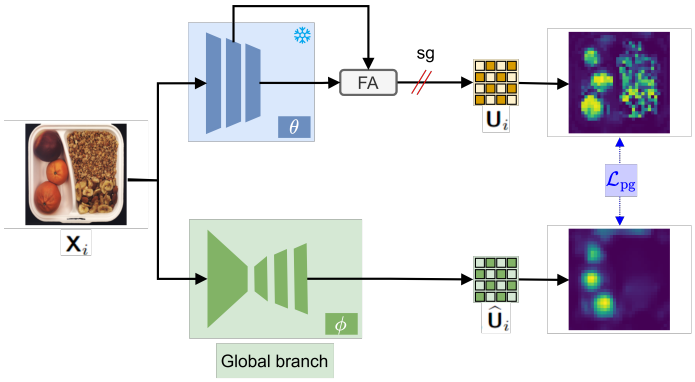


Figure 4: Global Branch of `ULSAD`

Based on this observation and inspired by the attention-transfer concept for knowledge distillation (Zagoruyko & Komodakis, 2017; Tung & Mori, 2019), we propose to learn the spatial relationships (Dosovitskiy et al., 2021) among the feature vectors in $\mathbf{U}_i$ obtained from normal images. Recall that each feature vectors correspond to a patch in the original image. Therefore, learning the spatial relationships among the feature vectors would allow us to learn the same among the patches in the original image. Starting from $\boldsymbol{Z}_i = f(\mathbf{U}_i)$, we first compute the self-attention weight matrix $\boldsymbol{W}_i \in \mathbb{R}^{(h^* \times w^*) \times (h^* \times w^*)}$, such that

$$\boldsymbol{W}_i[p, q] = \frac{\exp(\boldsymbol{z}_{ip}^T \, \boldsymbol{z}_{iq} / \sqrt{c^*})}{\sum_{r=1}^{h^* \times w^*} \exp(\boldsymbol{z}_{ir}^T \, \boldsymbol{z}_{iq} / \sqrt{c^*})}. \tag{3}$$

Then, the attention map $\boldsymbol{A}_i \in \mathbb{R}^{c^* \times (h^* \times w^*)}$ is computed as $\boldsymbol{A}_i = \boldsymbol{Z}_i \boldsymbol{W}_i$. For learning the spatial relations using $\boldsymbol{A}_i$ as our target, we use a convolutional autoencoder-like network $\phi = \phi_e \circ \phi_d$ where ϕ_e is the encoder and ϕ_d is the decoder. Similar to a standard autoencoder, ϕ_e compresses the input image $\mathbf{X}_i$ to a lower dimensional space. However, ϕ_d maps the encoded representation to the feature space $\mathcal{U}$, which has a lower dimension than the input space $\mathcal{X}$. We denote the output of ϕ as $\widehat{\mathbf{U}}_i = \phi(\mathbf{X}_i)$.

A direct approach would be to compute the self-attention map for $\widehat{\mathbf{U}}_i$ and minimize its distance from $\boldsymbol{A}_i$. However, it makes the optimization problem computationally challenging as each vector in $\widehat{\mathbf{U}}_i$ is coupled

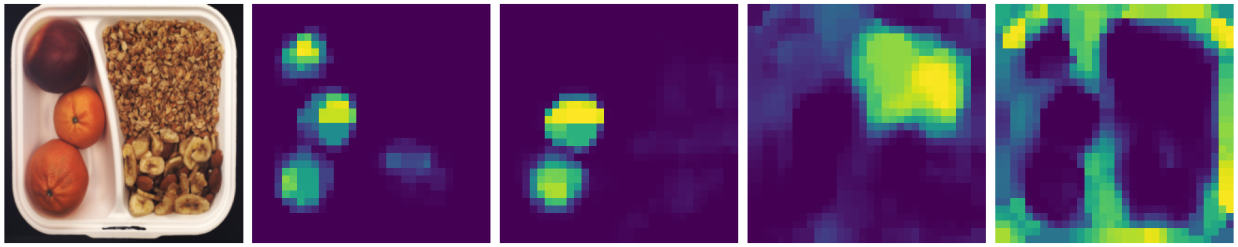


Figure 5: (First) Example image belonging to the category "breakfast box" in the MVTecLOCO dataset. (Rest) Visualization of attention maps computed using the intermediate features from a pre-trained model.

with every other vector by the network weights ϕ (Zhang et al., 2023). To overcome this, we compute the cross-attention map $\widehat{\boldsymbol{A}}_i \in \mathbb{R}^{c^* \times (h^* \times w^*)}$ between $\mathbf{U}_i$ and $\widehat{\mathbf{U}}_i$, by first computing $\widehat{\boldsymbol{W}}_i$ using $\widehat{\mathbf{Z}}_i = f(\widehat{\mathbf{U}}_i)$ as

$$\widehat{\boldsymbol{W}}_i[p,q] = \frac{\exp(\boldsymbol{z}_{ip}^T \, \widehat{\boldsymbol{z}}_{iq}/\sqrt{c^*})}{\sum_{r=1}^{h^* \times w^*} \exp(\boldsymbol{z}_{ir}^T \, \widehat{\boldsymbol{z}}_{iq}/\sqrt{c^*})}. \tag{4}$$

Then, the attention map $\widehat{\boldsymbol{A}}_i$ can be computed as $\widehat{\boldsymbol{A}}_i = \boldsymbol{Z}_i \widehat{\boldsymbol{W}}_i$. Given, the self-attention map $\boldsymbol{A}_i$ and the cross-attention map $\widehat{\boldsymbol{A}}_i$, the global loss is defined as

$$\mathcal{L}_{\text{pg}}(\widehat{\boldsymbol{A}}_i, \boldsymbol{A}_i) = \frac{1}{h^* \times w^*} \sum_{k=1}^{h^* w^*} l_v(\widehat{\boldsymbol{a}}_{ik}, \boldsymbol{a}_{ik}) + \lambda_g \, l_d(\widehat{\boldsymbol{a}}_{ik}, \boldsymbol{a}_{ik}), \tag{5}$$

where $\boldsymbol{a}_{ik} = \boldsymbol{A}_i[:,k]$, $\widehat{\boldsymbol{a}}_{ik} = \widehat{\boldsymbol{A}}_i[:,k]$ and $\lambda_{\text{g}} \geq 0$ is a hyperparameter. It is important to note that the loss in Equation 5 is minimized when the attention map $\widehat{\boldsymbol{A}}_i$ is the mean over all the training samples. As a result, the global branch is prone to false positives in the presence of sharp edges or heavily textured surfaces due to the loss of high-frequency details (Dosovitskiy & Brox, 2016; Assran et al., 2023). To address this, we utilize the FRN ψ in the local branch to learn the output $\widehat{\mathbf{U}}_i$. Recall that the output of FRN $\widetilde{\mathbf{U}} \in \mathbb{R}^{2c^* \times h^* \times w^*}$ has $2c^*$ number of channels to simultaneously generate two feature maps $\widetilde{\mathbf{U}}_i^{'}$ and $\widetilde{\mathbf{U}}_i^{''}$, both having dimension $c^* \times h^* \times w^*$. Out of which, $\widetilde{\mathbf{U}}_i^{'}$ is used for learning the patch features. Here, we define the loss $\mathcal{L}_{lg}$ to relate the local feature map $\widetilde{\mathbf{U}}_i^{''}$ with the global feature map $\widehat{\mathbf{U}}_i$ as

$$\mathcal{L}_{\text{lg}}(\widetilde{\boldsymbol{Z}}_i^{''}, \widehat{\boldsymbol{Z}}_i) = \frac{1}{h^* \times w^*} \sum_{k=1}^{h^* w^*} l_v(\widetilde{\boldsymbol{z}}_{ik}^{''}, \widehat{\boldsymbol{z}}_{ik}) + \lambda_g \, l_d(\widetilde{\boldsymbol{z}}_{ik}^{''}, \widehat{\boldsymbol{z}}_{ik}), \tag{6}$$

where $\widetilde{\boldsymbol{Z}}_i^{''} = f(\widetilde{\mathbf{U}}_i^{''})$. Therefore, during inference, a difference between the $\widetilde{\mathbf{U}}_i^{''}$ and $\widehat{\mathbf{U}}_i$ indicates the presence of logical anomalies with fewer false positives. The benefits of such a framework are two-fold: (1) it allows for learning the spatial relationships in the normal images while minimizing false positives, and (2) doubling the channels in the decoder allows sharing the encoder architecture, reducing the computational costs.

3.4 Algorithmic Overview

An overview of `ULSAD` is outlined in Algorithm 1, which can simultaneous detect structural and logical anomalies. Firstly, we pass a normal image $\mathbf{X}_i$ from the training dataset $\mathcal{D}_N$ through the feature extractor θ to obtain multi-scale feature $\mathbf{U}_i$. We normalize the features (line 4, Algorithm 1) with the channel-wise mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ computed over all the features. We do not include this step in Algorithm 1 as the calculation is trivial. Instead, we consider the values $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ to be given as input parameters for the sake of simplicity. Secondly, we obtain $\widetilde{\mathbf{U}}_i$ by passing $\mathbf{U}_i$ through the feature reconstruction network ψ (line 7, Algorithm 1). Recall that, $\widetilde{\mathbf{U}}_i$ has a dimension $2c^* \times h^* \times w^*$ which can be decomposed into two feature maps $\widetilde{\mathbf{U}}_i^{'}$ and $\widetilde{\mathbf{U}}_i^{''}$ each with a dimension $c^* \times h^* \times w^*$. The feature reconstruction loss $\mathcal{L}_{\text{pl}}$ is then computed

Algorithm 1: Unified Logical and Structural AD (`ULSAD`) // Local branch Global branch

Require: Training dataset $\mathcal{D}_N$, Feature extractor θ, Feature reconstruction network ϕ
Global autoencoder ψ, Number of epochs e, Learning rate η, Pretrained feature statictics $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$

1 **for** (epoch $\in 1, \cdots, e$) and ($\mathbf{X}_i \in \mathcal{D}_N$) **do**
2 Extract normalized features maps using the pre-trained network:
3 $\mathbf{U}_i \leftarrow \theta(\mathbf{X}_i)$
4 $\mathbf{U}_i \leftarrow (\mathbf{U}_i - \boldsymbol{\mu})/\boldsymbol{\sigma}$
5 $\boldsymbol{Z}_i \leftarrow f(\mathbf{U}_i)$
6 Reconstruct the features maps using the local branch:
7 $\widetilde{\mathbf{U}}_i \leftarrow \psi(\mathbf{U}_i)$
8 $\widetilde{\boldsymbol{Z}}_i \leftarrow f(\mathbf{U}_i)$
9 Compute local loss (Eq. 1):
10 $l_l \leftarrow \mathcal{L}_{\text{pl}}(\widetilde{\boldsymbol{Z}}_i', \boldsymbol{Z}_i)$
11 Obtain the output of the global autoencoder:
12 $\widehat{\mathbf{U}}_i \leftarrow \phi(\mathbf{X}_i)$
13 $\widehat{\boldsymbol{Z}}_i \leftarrow f(\widehat{\mathbf{U}}_i)$
14 Compute global loss (Eq. 5):
15 $l_g \leftarrow \mathcal{L}_{pg}(\widehat{\boldsymbol{Z}}_i, \boldsymbol{Z}_i)$
16 Compute local-global loss (Eq. 6):
17 $l_{\text{lg}} \leftarrow \mathcal{L}_{lg}(\widehat{\boldsymbol{Z}}_i, \widetilde{\boldsymbol{Z}}_i'')$
18 Compute overall loss:
19 $l \leftarrow l_l + l_g + l_{\text{lg}}$
20 Update model parameters:
21 $\phi \leftarrow \phi - \eta \nabla_\phi l$
22 $\psi \leftarrow \psi - \eta \nabla_\psi l$
23 **end**

Return: ϕ, ψ

between $\boldsymbol{Z}_i$ and $\widetilde{\boldsymbol{Z}}_i'$, where $\boldsymbol{Z}_i = f(\mathbf{U}_i)$ and $\widetilde{\boldsymbol{Z}}_i' = f(\widetilde{\mathbf{U}}_i^{'})$. Thirdly, we obtain the features $\widehat{\mathbf{U}}_i$ by passing the input sample $\mathbf{X}_i$ through the autoencoder ϕ. Then for learning the spatial relationships from the normal images, we compute $\mathcal{L}_{\text{pg}}$ between the self-attention map of $\boldsymbol{Z}_i$ and the cross-attention map between $\boldsymbol{Z}_i$ and $\widehat{\mathbf{Z}}_i = f(\widehat{\mathbf{U}}_i)$ (line 15, Algorithm 1). In the fourth step, we compute the loss $\mathcal{L}_{\text{lg}}$ between $\widehat{Z}_i$ and $\widetilde{\boldsymbol{Z}}_i'' = f(\widetilde{\mathbf{U}}_i^{''})$. Finally, the model parameters ψ and ϕ are updated based on the gradient of the total loss (line $21 - 22$, Algorithm 1). The end-to-end pipeline is illustrated in Figure 2.

3.5 Anomaly Detection and Segmentation

First, the local anomaly map $\boldsymbol{M}_i^l \in \mathbb{R}^{h^* \times w^*}$ is computed between the output of the local branch $\widetilde{\mathbf{U}}_i^{'}$ and the feature $\mathbf{U}_i$, such that

$$\boldsymbol{M}_i^l[h, w] = l_v(\widetilde{\mathbf{U}}_i^{'}[:, h, w],\ \mathbf{U}_i[:, h, w]) + \lambda_l\ l_d(\widetilde{\mathbf{U}}_i^{'}[:, h, w],\ \mathbf{U}_i[:, h, w]), \tag{7}$$

where $\widetilde{\mathbf{U}}_i^{'} = f^{-1}(\widetilde{\boldsymbol{Z}}_i')$ and $\mathbf{U}_i = f^{-1}(\boldsymbol{Z}_i)$. Similarly, the global anomaly map $\boldsymbol{M}_i^g$ is calculated using the output from the global autoencoder $\widehat{\boldsymbol{Z}}_i$ and the local reconstruction branch $\widetilde{\boldsymbol{Z}}_i''$:

$$\boldsymbol{M}_i^g[h, w] = l_v(\widetilde{\mathbf{U}}_i^{''}[:, h, w], \widehat{\mathbf{U}}_i[:, h, w]) + \lambda_g\ l_d(\widetilde{\mathbf{U}}_i^{''}[:, h, w], \widehat{\mathbf{U}}_i[:, h, w]), \tag{8}$$

Recall that λ_l and λ_g are hyperparameters. To obtain the final anomaly map m_i for the image x_i, the local and global anomaly maps are averaged. As local and global anomaly maps can have different ranges of anomaly scores, prior to averaging, each map is normalized individually. Besides ensuring a similar range of

scores for both maps, it prevents noise from one of the maps, making the anomaly detected from the other map indistinguishable. In this work, we follow a quantile-based normalization. The benefit is that it does not make any assumptions about the distribution of anomaly scores.

First, we generate two sets of anomaly maps: $\mathbb{M}^l = \{\boldsymbol{M}_i^l \mid \mathbf{X}_i \in \mathcal{D}_{\text{valid}}\}$ and $\mathbb{M}^g = \{\boldsymbol{M}_i^g \mid \mathbf{X}_i \in \mathcal{D}_{\text{valid}}\}$ utilizing images from the validation dataset $\mathcal{D}_{\text{valid}}$. Subsequently, we calculate the quantiles at significance levels α and β for each of these sets. The quantiles corresponding to the scores in $\mathbb{M}^g$ are denoted as q_α^g and q_β^g, whereas those for the scores in $\mathbb{M}^l$ are represented as q_α^l and q_β^l. Ultimately, employing these computed quantiles, we establish two linear transformations, $t^g(\cdot)$ and $t^l(\cdot)$, for the global and local anomaly maps, respectively. These transformations map the quantile at level α to an anomaly score of 0 and at β to a score of 0.1, as

$$t^l(\boldsymbol{M}_i^l) = 0.1\left(\boldsymbol{M}_i^l - \left(\frac{q_\alpha^l}{q_\beta^l - q_\alpha^l}\right)\mathbf{1}_{h^*\times w^*}\right), \quad \text{and} \quad t^g(\boldsymbol{M}_i^g) = 0.1\left(\boldsymbol{M}_i^g - \left(\frac{q_\alpha^g}{q_\beta^g - q_\alpha^g}\right)\mathbf{1}_{h^*\times w^*}\right), \tag{9}$$

where $\mathbf{1}_{h^*\times w^*}$ is a matrix of all ones. During inference, the global and local anomaly maps for each sample in $\mathcal{D}_{\text{test}}$ are averaged after normalizing them with their respective transformations, which can be expressed as

$$\boldsymbol{M}_i = \frac{t^l(\boldsymbol{M}_i^l) + t^g(\boldsymbol{M}_i^g)}{2}.$$

For AD, the image anomaly score

$$s_i = \max_{h,\,w} \; \boldsymbol{M}_i(h, w) \tag{10}$$

is computed as the maximum value in the combined anomaly map $\boldsymbol{M}_i$. We follow Batzner et al. (2024) in mapping the destination values to 0 and 0.1 as they result in colour maps suitable for 0-to-1 colour scales. It should be noted that the choice of mapping the quantile values has no impact on the AU-ROC scores as it depends solely on the ranking of the scores.

4 Experimental Evaluation

In this section, we answer the following three questions: (i) How does `ULSAD` perform as compared to the SOTA methods? (ii) How effective is the local and global branch for the detection of *structural* and *logical* anomalies? (iii) How does each component in `ULSAD` impact the overall performance?

4.1 Setup

Benchmark Datasets. We evaluate our method on the following five IAD benchmarking datasets:

[1] MVTec AD (Bergmann et al., 2019). It consists of images from industrial manufacturing across 15 categories comprised of 10 objects and 5 textures. In totality, it contains 3,629 normal images for training. For evaluation, 1,258 anomalous images with varying pixel level defects and 467 normal images.

[2] MVTec-Loco (Bergmann et al., 2022). An extension of MVTec dataset, it encompasses both local structural anomalies and logical anomalies violating long-range dependencies. It consists of 5 categories, with 1,772 normal images for training and 304 normal images for validation. It also contains 1568 images, either normal or anomalous, for evaluation.

[3] BTAD (Mishra et al., 2021). It includes 3 categories with 1,799 normal images for training. It also consists of 290 anomalous images and 451 normal images for testing.

[4] MPDD (Jezek et al., 2021). It focuses on metal part fabrication defects. The images are captured in variable spatial orientation, position, and distance of multiple objects concerning the camera at different light intensities and with a non-homogeneous background. It consists of 6 classes of metal parts with 888 training images. For evaluation, the dataset has 176 normal and 282 anomalous images.

[5] VisA (Zou et al., 2022). It contains 10,821 high-resolution images (9,621 normal and 1,200 anomalous images) across 12 different categories. The anomalous images contain different types of anomalies such as

Table 1: Average Detection Performance in AUROC (%). Style: **best** and second best

Method	BTAD	MPDD	MVTec	MVTec-LOCO	VisA
PatchCore (Roth et al., 2022)	93.27	93.27	**98.75**	81.49	91.48
CFLOW (Gudovskiy et al., 2021)	93.57	87.11	94.47	73.62	93.57
DRÆM (Zavrtanik et al., 2021)	73.42	74.14	75.26	62.35	73.42
EfficientAD (Batzner et al., 2024)	88.26	85.42	98.23	80.62	91.21
FastFlow (Yu et al., 2021)	91.68	65.03	90.72	71.00	91.68
PaDiM (Defard et al., 2021)	93.20	68.48	91.25	68.38	93.20
Reverse Distillation (Deng & Li, 2022)	83.87	79.62	79.65	61.56	83.87
DFR (Yang et al., 2020)	94.60	79.75	93.54	72.87	**94.60**
ULSAD (Ours)	**96.17** ± 0.45	**95.73** ± 0.45	97.65 ± 0.38	**84.1** ± 0.86	92.46 ± 0.45

Table 2: Average Segmentation Performance in AUROC (%) and AUPRO (%). Style: **best** and second best

Method	BTAD	MPDD	MVTec	MVTec-LOCO	VisA
PatchCore (Roth et al., 2022)	96.85 | 71.48	**98.07** | 90.84	**97.71** | 91.15	75.77 | 69.09	97.93 | 85.12
CFLOW (Gudovskiy et al., 2021)	96.60 | 73.11	97.42 | 88.56	97.17 | 90.14	76.99 | 66.93	98.04 | 85.29
DRÆM (Zavrtanik et al., 2021)	59.04 | 22.48	86.96 | 70.04	75.01 | 49.72	63.69 | 40.06	71.31 | 54.68
EfficientAD (Batzner et al., 2024)	82.13 | 54.37	97.03 | 90.44	96.29 | 90.11	70.36 | 66.96	97.51 | 84.45
FastFlow (Yu et al., 2021)	96.15 | 75.27	93.60 | 76.89	96.44 | 88.79	75.55 | 53.04	97.32 | 81.70
PaDiM (Defard et al., 2021)	97.07 | 77.80	94.51 | 81.18	96.79 | 91.17	71.32 | 67.97	97.09 | 80.80
Reverse Distillation (Deng & Li, 2022)	**97.85** | **81.47**	97.83 | 91.86	97.25 | **93.12**	68.55 | 66.28	**98.68** | **91.77**
DFR (Yang et al., 2020)	97.62 | 59.06	97.33 | 90.46	94.93 | 89.42	61.72 | 69.78	97.90 | 91.72
ULSAD (Ours)	96.73 | 75.41 ± 0.51 | ± 3.95	97.45 | **92.02** ± 0.99 | ± 2.64	97.61 | 91.67 ± 0.64 | ± 1.36	**80.06** | **73.73** ± 0.20 | ± 0.35	98.24 | 87.12 ± 0.20 | ± 0.89

scratches, bent, cracks, missing parts or misplacements. For each type of defect, there are 15-20 images, and an image can depict multiple defects.

Evaluation metrics. We measure the image-level anomaly detection performance via the area under the receiver operator curve (AUROC) based on the assigned anomaly score. To measure the anomaly localization performance, we use pixel-level AUROC and area under per region overlap curve (AUPRO). Furthermore, following prior works (Roth et al., 2022; Gudovskiy et al., 2021; Bergmann et al., 2019), we compute the average metrics over all the categories for each of the benchmark datasets. Moreover, for ULSAD, we report all the results over 5 runs with different random seeds.

Baselines. We compare our method with existing state-of-the-art unsupervised AD methods, namely PatchCore (Roth et al., 2022), PaDim (Defard et al., 2021), CFLOW (Gudovskiy et al., 2021), FastFLOW (Yu et al., 2021), DRÆM (Zavrtanik et al., 2021), Reverse Distillation (Deng & Li, 2022) and DFR (Yang et al., 2020). In this study, we only consider baselines that are capable of both anomaly detection and localization.

Implementation details. ULSAD is implemented in PyTorch (Paszke et al., 2019). For the baselines, we followed the implementation in Anomalib (Akcay et al., 2022), a widely used AD library for benchmarking. In ULSAD, we used a Wide-ResNet50-2 pre-trained on ImageNet (Zagoruyko & Komodakis, 2016) similar to PathCore (Roth et al., 2022). We used a convolutional autoencoder-like network in the global branch ϕ and the feature prediction network ψ. It consists of convolution layers with LeakyReLU activation in the encoder and deconvolution layers in the decoder. The exact architecture is provided in the Appendix A. Unless otherwise stated, for all the experiments, we considered an image size of 256×256. We train ULSAD over 200 epochs for each category using an Adam optimizer with a learning rate of 0.0002 and a weight decay of 0.00002. For the baselines, we used the hyperparameters mentioned in the respective paper.

4.2 Evaluation Results

We summarize the anomaly detection performance of ULSAD in Table 1 and the localization performance in Table 2. On the BTAD dataset, we improve over the DFR by approximately 2% in detection. Inspecting the images from the dataset, we hypothesize that the difference stems from the use of a global branch in ULSAD

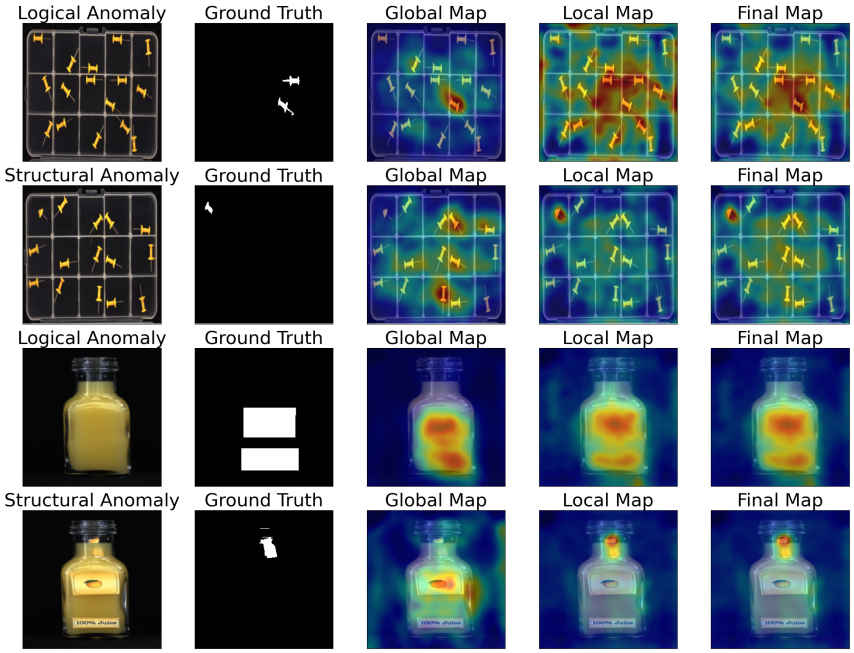


Figure 6: Example of anomaly maps obtained from global and local branches along with the combined map.

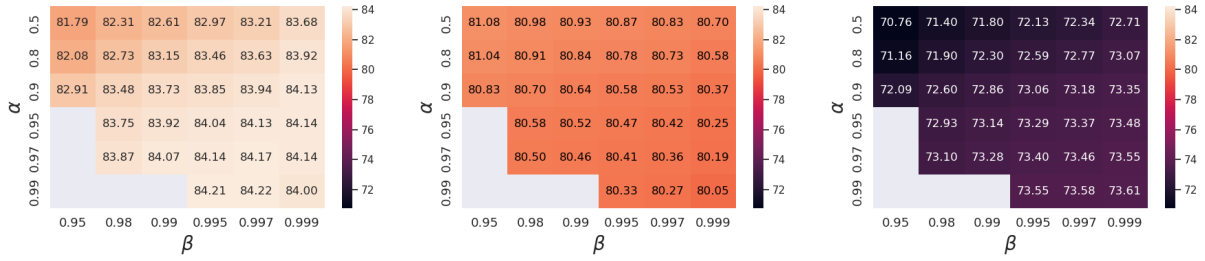
as the structural imperfections are not limited to small regions. For localization, reverse distillation performs better owing to the use of multi-scale anomaly maps. We can observe similar improvements over DFR on MVTec dataset. Although PatchCore provides superior performance on MVTec, it should be noted that even without using a memory bank, `ULSAD` provides comparable results. Then, we focus on more challenging datasets such as MPDD and MVTecLOCO. While MPDD contains varying external conditions such as lighting, background and camera angles, MVTecLOCO contains both logical and structural anomalies. We can observe improvements over DFR ($\sim 12-16\%$) in both datasets. This highlights the effectiveness of our method. We visualize the anomaly maps for samples from the "pushpin" and "juice bottle" categories in Figure 6. We provide anomaly maps obtained from different methods in Figure 9. It can be seen that while the global branch is more suited to the detection of logical anomalies, the local branch is capable of detecting localized structural anomalies. Extended versions of Table 1 and 2 are provided in the Appendix B.

4.3 Ablation study

Analysis of main components. We investigate the impact of the key components in `ULSAD` by sequentially adding them. For the ablations, we report the detection and localization performance on the MVTecLOCO dataset in Table 3. In the initial set of experiments, we set both λ_l and λ_g to 0. This results in the use of only the difference in values while computing the global and local objectives. The first row of the table corresponds to the use of only the local branch. We observe a slight improvement in both detection and localization performance when considering the global branch in addition to the local branch. For the sake of completeness, we also consider here a variant of the global loss $\mathcal{L}_{\text{pg}}$ where we compute the ℓ_2 distance between the feature maps instead of computing the self- and cross-attention maps. We refer to the alternative in the table as $\mathcal{L}_{\text{pg}}^d$. It can be seen

Table 3: Ablation of the main components of `ULSAD`.

Local Branch	**Global Branch**				**Performance (%)**
λ_l	λ_g	$\mathcal{L}_{\text{lg}}$	$\mathcal{L}_{\text{pg}}^d$	$\mathcal{L}_{\text{pg}}$	IAUROC | PAUROC | PAUPRO
0.0	-	-	-	-	77.67 | 75.17 | 73.37
0.0	0.0	-	✓	-	77.69 | 79.77 | 75.26
0.0	0.0	-	-	✓	71.67 | 73.92 | 67.22
0.0	0.0	✓	✓	-	81.40 | 82.12 | 77.47
0.0	0.0	✓	-	✓	81.08 | 81.97 | 76.45
0.5	-	-	-	-	79.14 | 76.57 | 73.41
0.5	0.5	-	✓	-	80.50 | 81.85 | 77.35
0.5	0.5	-	-	✓	74.51 | 76.59 | 69.01
0.5	0.5	✓	✓	-	82.19 | 81.25 | 75.50
0.5	0.5	✓	-	✓	84.15 | 79.78 | 73.68
0.5	0.5	✓	-	✓	**84.10** | **80.06** | **73.73** ± 0.86 | ± 0.20 | ± 0.35



(a) Image AUROC (b) Pixel AUROC (c) Pixel AUPRO

Figure 7: Ablation of α and β.

that when using the cross-attention map without considering the difference in direction, the performance drops compared to using $\mathcal{L}^d_{\text{pg}}$. On incorporation of the loss term that relates the global and local branches $\mathcal{L}^d_{\text{lg}}$, we see a significant boost in the overall performance. Further, the difference in performance between the two variants of the global loss becomes negligible. Considering the difference of direction in both the global and local objectives, we observe improvements over just using the difference in values, as seen in the last five rows of the table. Overall, the best performance is obtained by incorporating both the local-global loss term and the attention-based global objective.

Effect of normalization. We analyze the impact of the quantile-based normalization on the performance metrics by considering multiple values for α and β. The results are shown in the Figure 7. It can be seen that the final performance is robust to the choice of α and β. For our experiments, we set $\alpha = 0.9$ and $\beta = 0.995$.

Effect of backbone. We investigate the impact of using different pre-trained backbones in `ULSAD` in Figure 8. It is observed that the overall best performance is obtained by using a Wide-ResNet101-2 architecture in both detection and localization. More specifically, for detection, Wide-ResNet variants are more effective than the ResNet architectures, whereas, for localization performance measured using Pixel AUROC, the deeper networks such as ResNet152 and Wide-ResNet101-2 seem to have precedence over their shallower counterparts. Overall, we can see that performance is robust to the choice of pre-trained model architecture. In our experiments, we utilize a Wide-ResNet50-2 architecture which is used by most of our baselines for fair comparison.

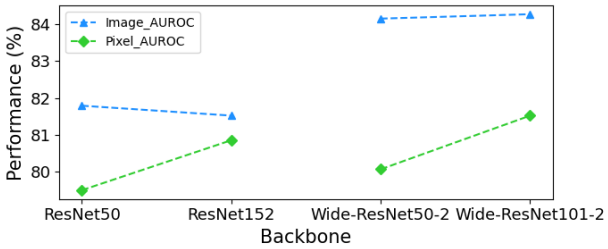


Figure 8: Ablation study of the backbone network

5 Conclusion

Our study focuses on Deep Feature Reconstruction (DFR), which offers a memory- and compute-efficient way of detecting structural anomalies. We develop `ULSAD`, a unified framework for detecting structural and logical anomalies, which utilizes a dual-branch architecture by extending DFR. Specifically, we improve the training objective of DFR to enhance the capability to detect structural anomalies, which we refer to as the local branch. Furthermore, the bottleneck used in FRN at the local branch acts as an implicit regularizer, preventing the over-expressiveness of the model without using additional computationally expensive methods of regularization. Additionally, we introduce an attention-based loss using an autoencoder-like network for detecting logical anomalies in the global branch. Through extensive experimentation across five benchmarking image AD datasets, we demonstrate that `ULSAD` can provide competitive performance in the detection and localization of both structural and logical anomalies compared to eight state-of-the-art methods. Importantly, it performs well even when compared to memory-intensive retrieval-based methods such as PatchCore (Roth et al., 2022). Lastly, ablations demonstrate the impact of various components used in `ULSAD` and the influence of the pre-trained backbone θ on the overall performance.

Reproducibility Statement. We provide extensive descriptions of implementation details (in Section 4.1) and algorithm (in Algorithm 1) to help readers reproduce our results. Every measure is taken to ensure fairness in our comparisons by adopting the most commonly adopted evaluation settings in the anomaly detection literature. More specifically, we use the Anomalib library for our experiments, whenever applicable, for comparing their performance with `ULSAD`. For methods such as DFR (Yang et al., 2020), which is not implemented in Anomalib, we refer to the original codebase provided by the authors. We will publicly release the code upon acceptance.

Ethics Statement. We have read the TMLR Ethics Guidelines (https://jmlr.org/tmlr/ethics.html) and ensured that this work adheres to it. All benchmark datasets and pre-trained model checkpoints are publicly available and not directly subject to ethical concerns.

References

Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*, pp. 420–434. Springer, 2001.

Mohiuddin Ahmed, Abdun Naser Mahmood, and Md Rafiqul Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2 2016. ISSN 0167-739X. doi: 10.1016/J.FUTURE.2015.01.001.

Samet Akcay, Dick Ameln, Ashwin Vaidya, Barath Lakshmanan, Nilesh Ahuja, and Utku Genc. Anomalib: A Deep Learning Library for Anomaly Detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 1706–1710. IEEE, 10 2022. ISBN 978-1-6654-9620-9. doi: 10.1109/ICIP46576.2022.9897283. URL https://ieeexplore.ieee.org/document/9897283/.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Kilian Batzner, Lars Heckler, and Rebecca König. EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 128–138, 1 2024. URL https://arxiv.org/abs/2303.14535v2.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTEC ad-A comprehensive real-world dataset for unsupervised anomaly detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9584–9592, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00982.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 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, 2020. URL www.mvtec.com.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond Dents and Scratches: Logical Constraints in Unsupervised Anomaly Detection and Localization. *International Journal of Computer Vision*, 130(4):947–969, 2022. ISSN 1573-1405. doi: 10.1007/s11263-022-01578-9. URL https://doi.org/10.1007/s11263-022-01578-9.

Heejeong Choi, Donghwa Kim, Jounghee Kim, Jina Kim, and Pilsung Kang. Explainable anomaly detection framework for predictive maintenance in manufacturing systems. *Applied Soft Computing*, 125:109147, 8 2022. ISSN 1568-4946. doi: 10.1016/J.ASOC.2022.109147.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: a Patch Distribution Modeling Framework for Anomaly Detection and Localization. In *International Conference on Pattern Recognition*, volume 12664 LNCS, pp. 475–489. Springer Science and Business Media Deutschland GmbH, 11 2021. ISBN 9783030687984. doi: 10.48550/arxiv.2011.08785. URL https://arxiv.org/abs/2011.08785v1.

Hanqiu Deng and Xingyu Li. 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, 2022.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection – a survey. *ACM Comput. Surv.*, 54(7), jul 2021. ISSN 0360-0300. doi: 10.1145/3464423. URL https://doi.org/10.1145/3464423.

Han Gao, Huiyuan Luo, Fei Shen, and Zhengtao Zhang. Towards Total Online Unsupervised Anomaly Detection and Localization in Industrial Vision, 5 2023. URL https://arxiv.org/abs/2305.15652v1.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton Van Den Hengel. Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:1705–1714, 4 2019. ISSN 15505499. doi: 10.48550/arxiv.1904.02639. URL https://arxiv.org/abs/1904.02639v2.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows. *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, pp. 1819–1828, 7 2021. doi: 10.1109/WACV51458.2022.00188. URL https://arxiv.org/abs/2107.12571v1.

Waleed Hilal, S. Andrew Gadsden, and John Yawney. Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. *Expert Systems with Applications*, 193:116429, 5 2022. ISSN 0957-4174. doi: 10.1016/J.ESWA.2021.116429.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: Evaluating current methods in complex conditions. *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*, 2021-October:66–71, 2021. ISSN 2157023X. doi: 10.1109/ICUMT54235.2021.9631567.

Chieh-Hsin Lai, Dongmian Zou, and Gilad Lerman. Robust Subspace Recovery Layer for Unsupervised Anomaly Detection. *arXiv preprint*, 3 2019. doi: 10.48550/arxiv.1904.00152. URL https://arxiv.org/abs/1904.00152v2.

Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. CFA: Coupled-hypersphere-based Feature Adaptation for Target-Oriented Anomaly Localization. *IEEE Access*, 10:78446–78454, 6 2022. ISSN 21693536. doi: 10.1109/ACCESS.2022.3193699. URL https://arxiv.org/abs/2206.04325v1.

Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1):104–135, 2024.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization. *IEEE International Symposium on Industrial Electronics*, 2021-June, 4 2021. doi: 10.1109/ISIE45552.2021.9576231. URL http://arxiv.org/abs/2104.10036http://dx.doi.org/10.1109/ISIE45552.2021.9576231.

Duc Tam Nguyen, Zhongyu Lou, Michael Klar, and Thomas Brox. Anomaly Detection With Multiple-Hypotheses Predictions. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4800–4809. PMLR, 7 2019. URL https://proceedings.mlr.press/v97/nguyen19b.html.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2), 7 2020. doi: 10.1145/3439950. URL http://arxiv.org/abs/2007.02500http://dx.doi.org/10.1145/3439950.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury Google, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf Xamla, Edward Yang, Zach Devito, Martin Raison Nabla, Alykhan Tejani, Sasank Chilamkurthy, Qure Ai, Benoit Steiner, Lu Fang Facebook, Junjie Bai Facebook, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32, 2019.

Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. OCGAN: One-class novelty detection using gans with constrained latent representations. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:2893–2901, 6 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00301.

Tal Reiss, Niv Cohen, Eliahu Horwitz, Ron Abutbul, and Yedid Hoshen. Anomaly Detection Requires Better Representations. In *European Conference on Computer Vision*, pp. 56–68, 10 2022. ISBN 9783031250682. doi: 10.1007/978-3-031-25069-9{_}4. URL https://arxiv.org/abs/2210.10773v1.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards Total Recall in Industrial Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022. doi: 10.48550/arxiv.2106.08265. URL https://arxiv.org/abs/2106.08265v2.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep One-Class Classification. In *International Conference on Machine Learning*, pp. 4393–4402. PMLR, 2018. URL https://proceedings.mlr.press/v80/ruff18a.html.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Gregoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus Robert Muller. A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE*, 109(5):756–795, 5 2021. ISSN 15582256. doi: 10.1109/JPROC.2021.3052449.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 12 2015. ISSN 15731405. doi: 10.1007/S11263-015-0816-Y/FIGURES/16. URL https://link.springer.com/article/10.1007/s11263-015-0816-y.

Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially Learned One-Class Classifier for Novelty Detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3379–3388, 12 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00356.

Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 14897–14907, 11 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.01466. URL https://arxiv.org/abs/2011.11108v1.

Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=aRtjVZvbpK.

Wuqin Tang, Qiang Yang, Kuixiang Xiong, and Wenjun Yan. Deep learning based automatic defect identification of photovoltaic module using electroluminescence images. *Solar Energy*, 201:453–460, 5 2020. ISSN 0038-092X. doi: 10.1016/J.SOLENER.2020.03.049.

Robert Tibshirani and Trevor Hastie. Outlier sums for differential gene expression analysis. *Biostatistics*, 8 (1):2–8, 1 2007. ISSN 1465-4644. doi: 10.1093/BIOSTATISTICS/KXL005. URL https://academic.oup.com/biostatistics/article/8/1/2/252461.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1365–1374, 2019.

Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning Semantic Context from Normal Samples for Unsupervised Anomaly Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3110–3118, 5 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16420.

Jie Yang, Yong Shi, and Zhiquan Qi. DFR: Deep Feature Reconstruction for Unsupervised Anomaly Segmentation. *Neurocomputing*, 424:9–22, 12 2020. doi: 10.1016/j.neucom.2020.11.018. URL http://arxiv.org/abs/2012.07122http://dx.doi.org/10.1016/j.neucom.2020.11.018.

Jinsung Yoon, Kihyuk Sohn, Chun-Liang Li, Sercan O. Arik, and Tomas Pfister. SPADE: Semi-supervised Anomaly Detection under Distribution Mismatch. *Transactions on Machine Learning Research*, 11 2023. URL http://arxiv.org/abs/2212.00173.

Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. FastFlow: Unsupervised Anomaly Detection and Localization via 2D Normalizing Flows. *arXiv preprint arXiv:2111.07677*, 11 2021. doi: 10.48550/arxiv.2111.07677. URL https://arxiv.org/abs/2111.07677v2.

Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. *British Machine Vision Conference 2016, BMVC 2016*, 2016-September:1–87, 5 2016. doi: 10.5244/C.30.87. URL https://arxiv.org/abs/1605.07146v4.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sks9_ajex.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRÆM - A discriminatively trained reconstruction embedding for surface anomaly detection. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8310–8319, 8 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.00822. URL https://arxiv.org/abs/2108.07610v2.

Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient GAN-Based Anomaly Detection. *arXiv preprint*, 2 2018. doi: 10.48550/arxiv.1802.06222. URL https://arxiv.org/abs/1802.06222v2.

Jie Zhang, Masanori Suganuma, and Takayuki Okatani. Contextual Affinity Distillation for Image Anomaly Detection. *arXiv preprint arXiv:2307.03101*, 7 2023. URL https://arxiv.org/abs/2307.03101v1.

Kang Zhou, Yuting Xiao, Jianlong Yang, Jun Cheng, Wen Liu, Weixin Luo, Zaiwang Gu, Jiang Liu, and Shenghua Gao. Encoding Structure-Texture Relation with P-Net for Anomaly Detection in Retinal Images. In *European conference on computer vision*, volume 12365 LNCS, pp. 360–377. Springer, 2020. doi: 10.1007/978-3-030-58565-5{_}22. URL https://link.springer.com/10.1007/978-3-030-58565-5_22.

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

A Implementation Details

ULSAD is implemented in PyTorch (Paszke et al., 2019). Specifically, we used the Anomalib (Akcay et al., 2022) library by incorporating our code within it. It helps us have a fair comparison as we use the implementations of baselines from Anomalib. Moreover, we used a single NVIDIA A4000 GPU for all the experiments. The architecture of FRN and global autoencoder-like model is provided in Table 4 and 5, respectively.

Table 4: Feature Reconstruction Network of ULSAD.

	Layer Name	Stride	Kernel Size	Number of Kernels	Padding	Activation
Encoder	Conv-1	2	3×3	768	1	ReLU
	BatchNorm-1	-	-	-	-	-
	Conv-2	2	3×3	1536	1	ReLU
	BatchNorm-2	-	-	-	-	-
	Conv-3	1	3×3	1536	1	ReLU
	BatchNorm-3	-	-	-	-	-
Decoder	ConvTranspose-1	2	4×4	768	1	ReLU
	BatchNorm-4	-	-	-	-	-
	ConvTranspose-2	2	4×4	384	1	ReLU
	BatchNorm-5	-	-	-	-	-
	ConvTranspose-3	1	5×5	384	1	ReLU
	BatchNorm-6	-	-	-	-	-

Table 5: Global Autoencoder of ULSAD.

	Layer Name	Stride	Kernel Size	Number of Kernels	Padding	Activation
Encoder	Conv-1	2	4×4	32	1	ReLU
	BatchNorm-1	-	-	-	-	-
	Conv-2	2	4×4	32	1	ReLU
	BatchNorm-2	-	-	-	-	-
	Conv-3	2	4×4	64	1	ReLU
	BatchNorm-3	-	-	-	-	-
	Conv-4	2	4×4	64	1	ReLU
	BatchNorm-4	-	-	-	-	-
	Conv-5	2	4×4	64	1	ReLU
	BatchNorm-5	-	-	-	-	-
	Conv-6	1	8×8	64	1	ReLU
	BatchNorm-6	-	-	-	-	-
Decoder	Interpolate-1 (31, mode= "bilinear")	-	-	-	-	-
	Conv-1	1	4×4	64	2	ReLU
	BatchNorm-1	-	-	-	-	-
	Interpolate-2 (8, mode= "bilinear")	-	-	-	-	-
	Conv-2	1	4×4	64	2	ReLU
	BatchNorm-2	-	-	-	-	-
	Interpolate-3 (16, mode= "bilinear")	-	-	-	-	-
	Conv-3	1	4×4	64	2	ReLU
	BatchNorm-3	-	-	-	-	-
	Interpolate-4 (32, mode= "bilinear")	-	-	-	-	-
	Conv-4	1	4×4	64	2	ReLU
	BatchNorm-4	-	-	-	-	-
	Interpolate-5 (64, mode= "bilinear")	-	-	-	-	-
	Conv-5	1	4×4	64	2	ReLU
	BatchNorm-5	-	-	-	-	-
	Interpolate-6 (32, mode= "bilinear")	-	-	-	-	-
	Conv-6	1	3×3	64	1	ReLU
	BatchNorm-6	-	-	-	-	-
	Conv-7	1	3×3	384	1	ReLU

B Extended Results

Extended versions of the Table 1 and 2 are provided in Tables 6-20. It shows the performance of `ULSAD` per category of the benchmark datasets for anomaly detection and localization. Additionally, we provide a visual comparison of the generated anomaly maps using the MVTecLOCO dataset in Figure 9.

Table 6: Anomaly detection based on Image AUROC on MVTec dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	`ULSAD` (Ours)
bottle	100.0	94.6	98.57	99.37	100.0	98.10	94.92	100.0	100.0 ± 0.00
cable	93.46	74.29	89.30	86.94	98.54	95.41	79.54	94.61	97.92 ± 0.18
capsule	91.74	65.46	86.04	88.43	97.93	81.01	96.01	95.57	94.61 ± 0.28
carpet	93.26	57.70	98.76	97.31	97.91	57.95	97.59	99.70	98.50 ± 0.17
grid	93.57	76.02	99.08	84.04	97.24	93.07	94.57	100.0	92.67 ± 1.20
hazelnut	100.0	84.43	81.57	86.07	100.0	99.82	100.0	93.39	99.93 ± 0.05
leather	99.97	79.31	100.0	99.66	100.0	42.39	99.46	99.92	100.0 ± 0.00
metal_nut	99.76	45.06	94.53	96.92	99.61	67.20	93.06	99.34	98.88 ± 0.07
pill	90.73	44.65	87.53	88.52	94.35	54.66	92.06	99.08	96.17 ± 0.39
screw	88.30	89.38	66.00	75.24	98.26	94.57	93.69	98.09	95.20 ± 0.15
tile	100.0	90.15	95.42	95.49	98.67	97.37	92.97	99.85	99.99 ± 0.02
toothbrush	83.33	80.28	79.44	93.61	100.0	84.72	100.0	100.0	100.0 ± 0.00
transistor	91.50	88.37	94.42	92.29	100.0	83.29	80.54	96.57	97.65 ± 0.57
wood	98.33	90.96	97.54	98.33	99.30	53.16	98.77	98.13	98.81 ± 0.23
zipper	93.07	68.17	92.54	86.48	99.47	92.04	89.97	99.22	94.36 ± 0.13
Mean	94.47	75.26	90.72	91.25	<u>98.75</u>	79.65	93.54	98.23	97.65 ± 0.38

Table 7: Anomaly segmentation performance based on Pixel AUROC on MVTec dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	`ULSAD` (Ours)
bottle	98.58	76.53	97.8	98.3	97.98	98.31	90.83	98.31	96.21 ± 2.21
cable	96.1	66.59	95.71	96.81	98.03	96.37	91.37	98.5	97.71 ± 0.06
capsule	98.71	86.96	98.37	98.67	98.77	98.96	98.46	98.33	98.95 ± 0.03
carpet	98.57	71.95	98.27	98.68	98.67	99.05	98.46	94.83	99.18 ± 0.06
grid	97.49	53.56	98.39	92.82	97.86	99.01	97.41	96.02	95.47 ± 1.09
hazelnut	98.64	84.66	94.79	97.85	98.43	98.91	98.53	96.15	98.81 ± 0.03
leather	99.42	63.32	99.62	99.30	98.87	99.17	99.33	97.5	98.68 ± 0.01
metal_nut	97.97	80.25	97.01	96.71	98.51	97.68	93.02	98.07	97.62 ± 0.03
pill	97.83	77.17	96.38	95.03	97.53	96.96	96.86	98.63	96.67 ± 0.09
screw	97.64	83.38	89.87	97.89	99.19	99.43	99.07	98.50	99.33 ± 0.01
tile	96.68	85.75	93.14	92.42	94.86	95.47	90.82	91.61	95.78 ± 0.05
toothbrush	98.16	90.70	97.50	98.83	98.67	98.99	98.49	96.0	98.42 ± 0.02
transistor	89.91	63.23	96.45	96.85	96.84	86.77	79.11	94.77	98.89 ± 0.05
wood	94.70	71.73	95.71	93.83	93.31	95.06	95.36	90.85	95.20 ± 0.31
zipper	97.08	69.31	97.62	97.82	98.06	98.54	96.85	96.21	97.24 ± 0.07
Mean	97.17	75.01m	96.44	96.79	<u>97.71</u>	97.25	94.93	96.29	97.61 ± 0.64

Table 8: Anomaly segmentation performance based on Pixel AUPRO on MVTec dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
bottle	94.19	50.05	92.0	95.11	92.28	95.12	83.14	93.84	90.16 ± 3.24
cable	85.85	28.58	86.65	89.65	90.77	90.32	83.09	92.53	88.63 ± 0.48
capsule	90.47	81.11	90.15	92.62	92.4	93.93	96.33	91.09	93.77 ± 0.16
carpet	92.64	48.64	94.63	95.59	92.7	96.41	95.47	90.99	96.39 ± 0.24
grid	90.52	17.71	93.95	82.52	89.46	96.39	91.15	93.14	83.3 ± 3.96
hazelnut	96.12	76.19	93.92	92.95	94.44	96.92	97.17	83.25	94.87 ± 0.3
leather	98.39	52.1	99.06	97.91	96.33	97.97	98.34	97.32	97.44 ± 0.01
metal_nut	88.97	35.79	85.89	90.45	91.9	94.4	87.01	92.97	91.58 ± 0.19
pill	93.67	64.26	91.0	93.88	93.92	94.76	95.86	95.93	94.5 ± 0.08
screw	90.25	53.22	68.6	92.14	95.39	97.05	95.96	96.04	96.45 ± 0.1
tile	91.49	58.48	81.01	78.32	79.64	88.4	79.36	83.54	87.82 ± 0.15
toothbrush	81.05	54.02	80.62	93.52	86.48	92.23	92.93	88.61	86.28 ± 0.46
transistor	78.75	51.37	88.92	89.04	94.06	75.05	64.25	82.82	91.61 ± 0.68
wood	90.5	45.29	93.26	91.39	85.08	92.69	92.48	76.16	91.34 ± 0.29
zipper	89.3	28.98	92.12	92.48	92.43	95.18	88.74	93.48	90.89 ± 0.35
Mean	90.14	49.72	88.79	91.17	91.15	<u>93.12</u>	89.42	90.11	91.67 ± 1.36

Table 9: Anomaly detection performance based on Image AUROC on MVTecLOCO dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
breakfast_box	71.86	70.26	74.04	63.66	85.24	52.69	65.46	74.80	83.54 ± 0.23
juice_bottle	81.70	62.55	78.03	88.42	92.51	76.28	86.81	98.89	97.12 ± 0.10
pushpins	73.43	51.32	61.20	61.30	75.54	50.72	72.68	80.58	86.85 ± 0.94
screw_bag	65.48	59.39	68.04	60.14	69.90	65.15	63.55	67.42	70.71 ± 1.49
splicing_connectors	75.63	68.25	73.71	68.40	84.24	62.95	75.87	81.39	82.30 ± 0.72
Mean	73.62	62.35	71.00	68.38	81.49	61.56	72.87	80.62	84.1 ± 0.86

Table 10: Anomaly segmentation performance based on Pixel AUROC on MVTecLOCO dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
breakfast_box	89.6	63.61	82.73	87.35	88.53	85.78	76.25	80.76	89.14 ± 0.11
juice_bottle	91.37	80.71	86.33	91.99	90.54	90.41	87.06	88.40	89.07 ± 0.10
pushpins	70.66	54.74	82.94	40.72	67.67	41.42	29.42	59.96	75.64 ± 0.36
screw_bag	69.94	65.23	58.07	65.35	62.40	67.33	59.74	61.64	71.35 ± 0.12
splicing_connectors	63.40	54.16	67.69	71.20	69.71	57.82	56.14	61.02	75.10 ± 0.20
Mean	76.99	63.69	75.55	71.32	75.77	68.55	61.72	70.36	80.06 ± 0.20

Table 11: Anomaly segmentation performance based on Pixel AUPRO on MVTecLOCO dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
breakfast_box	67.27	36.11	63.8	74.28	73.08	69.67	63.56	58.44	71.36 ± 0.38
juice_bottle	80.75	51.51	77.90	88.78	85.42	84.95	82.88	86.51	87.72 ± 0.09
pushpins	61.09	24.68	50.62	52.71	63.52	53.52	59.12	59.25	68.34 ± 0.55
screw_bag	54.39	31.27	38.1	61.42	56.12	59.66	71.66	62.45	66.52 ± 0.33
splicing_connectors	71.15	56.72	34.77	62.64	67.29	63.62	71.67	68.14	74.70 ± 0.25
Mean	66.93	40.06	53.04	67.97	69.09	66.28	69.78	66.96	73.73 ± 0.35

Table 12: Anomaly detection performance based on Image AUROC on MPDD dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
tubes	99.64	61.28	89.36	56.48	87.50	89.67	94.47	95.28	93.39 ± 0.57
metal_plate	97.42	80.05	86.92	42.69	99.72	91.87	68.31	100.0	93.81 ± 0.28
connector	99.52	83.33	52.38	86.07	100.0	93.10	100.0	50.00	96.00 ± 0.82
bracket_white	79.89	84.00	50.78	80.33	89.67	83.67	54.55	96.48	100.0 ± 0.00
bracket_black	96.48	65.36	62.83	66.69	86.97	50.73	72.63	85.45	93.09 ± 0.32
bracket_brown	49.70	70.81	47.89	78.66	95.78	68.70	88.54	85.32	98.08 ± 0.14
Mean	87.11	74.14	65.03	68.48	93.27	79.62	79.75	85.42	95.73 ± 0.45

Table 13: Anomaly segmentation performance based on Pixel AUROC on MPDD dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
tubes	99.15	76.87	98.44	91.35	98.45	99.08	98.62	98.98	98.55 ± 0.10
metal_plate	98.56	96.23	92.99	91.67	98.30	97.50	93.59	96.52	96.77 ± 0.09
connector	97.38	90.01	92.69	97.93	99.11	98.55	98.68	99.32	99.40 ± 0.18
bracket_white	96.74	86.64	90.23	97.21	97.90	98.13	96.63	97.71	98.73 ± 0.1
bracket_black	97.68	95.89	94.39	93.79	97.52	96.40	98.42	97.17	97.43 ± 0.21
bracket_brown	95.04	76.11	92.84	95.13	97.15	97.34	98.06	92.46	93.83 ± 2.41
Mean	97.42	86.96	93.60	94.51	98.07	97.83	97.33	97.03	97.45 ± 0.99

Table 14: Anomaly segmentation performance based on Pixel AUPRO on MPDD dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
tubes	96.76	44.84	94.85	71.53	93.83	95.98	95.20	96.27	94.33 ± 0.33
metal_plate	91.53	82.83	74.62	75.47	92.50	92.00	83.99	83.59	90.07 ± 0.22
connector	91.32	72.06	76.98	92.74	96.89	95.29	95.60	97.77	97.98 ± 0.60
bracket_white	78.66	69.13	49.65	81.16	83.13	84.71	77.02	93.27	95.33 ± 0.37
bracket_black	89.48	93.12	79.65	83.51	93.65	89.14	95.57	89.98	90.17 ± 0.67
bracket_brown	83.62	58.29	85.62	82.69	85.07	94.04	95.37	81.77	84.24 ± 6.38
Mean	88.56	70.04	76.89	81.18	90.84	91.86	90.46	90.44	92.02 ± 2.64

Table 15: Anomaly detection performance based on Image AUROC on BTAD dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
01	98.64	80.17	94.46	99.51	98.09	92.23	99.51	94.15	**100.0** ± 0.00
02	82.12	65.23	84.27	82.17	81.73	61.73	85.68	75.42	**88.5** ± 0.78
03	99.95	74.87	96.3	97.92	**100.0**	97.65	98.62	95.22	**100.0** ± 0.00
Mean	93.57	73.42	91.68	93.20	93.27	83.87	94.60	88.26	**96.17** ± 0.45

Table 16: Anomaly segmentation performance based on Pixel AUROC on BTAD dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
01	95.44	59.11	93.05	96.54	95.94	96.98	96.93	64.59	95.86 ± 0.03
02	94.81	69.29	96.16	95.11	95.18	96.83	96.77	85.67	94.76 ± 0.88
03	99.55	48.73	99.25	99.56	99.44	99.74	99.12	96.12	99.55 ± 0.02
Mean	96.60	59.04	96.15	97.07	96.85	**97.85**	97.62	82.13	96.73 ± 0.51

Table 17: Anomaly segmentation performance based on AUPRO on BTAD dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
01	66.79	21.57	60.83	75.76	64.34	79.45	83.77	29.75	72.88 ± 0.12
02	54.32	27.64	67.98	59.19	52.36	66.05	65.58	44.37	55.16 ± 6.83
03	98.21	18.24	96.99	98.45	97.76	98.92	27.83	88.98	98.18 ± 0.08
Mean	73.11	22.48	75.27	77.80	71.48	**81.47**	59.06	54.37	75.41 ± 3.95

Table 18: Anomaly detection performance based on Image AUROC on VisA dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
candle	94.38	79.43	93.18	86.19	98.59	85.54	89.65	80.52	87.11 ± 0.29
capsules	69.9	72.77	81.05	61.72	69.92	87.37	76.75	63.73	79.61 ± 0.72
cashew	94.7	95.5	87.78	90.94	96.29	85.38	93.80	96.11	94.72 ± 0.16
chewinggum	99.02	83.68	95.18	98.20	99.29	81.92	99.22	98.27	99.49 ± 0.12
fryum	92.98	70.46	92.60	85.06	93.5	77.94	96.58	95.70	95.86 ± 0.14
macaroni1	92.72	72.8	82.48	78.62	91.50	82.06	95.14	95.23	90.66 ± 0.76
macaroni2	63.44	47.85	69.75	70.05	71.36	81.75	86.25	83.82	82.84 ± 1.05
pcb1	91.06	72.27	88.07	87.59	95.08	92.60	97.57	93.78	92.92 ± 0.11
pcb2	79.95	91.17	86.47	83.20	92.46	87.57	91.55	94.95	93.67 ± 0.18
pcb3	82.23	81.29	81.47	72.79	92.46	90.87	97.27	95.92	93.62 ± 0.16
pcb4	96.29	90.44	95.68	95.67	99.20	96.17	97.62	97.89	99.43 ± 0.03
pipe_fryum	96.54	75.32	96.16	89.28	98.07	85.68	98.36	98.59	99.61 ± 0.11
Mean	87.77	77.75	87.49	83.28	91.48	86.24	85.18	91.21	92.46 ± 0.45

Table 19: Anomaly segmentation performance based on Pixel AUROC on VisA dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
candle	98.75	83.1	97.24	97.68	98.92	99.11	98.41	89.93	97.77 ± 0.03
capsules	96.88	62.39	97.13	90.60	97.62	99.56	99.13	96.93	98.31 ± 0.31
cashew	99.25	74.17	98.57	97.45	98.88	97.23	95.63	98.85	99.49 ± 0.02
chewinggum	99.02	84.11	98.83	98.82	98.72	99.37	99.16	98.69	98.10 ± 0.41
fryum	97.08	85.7	93.20	96.20	94.30	96.33	95.45	96.52	97.38 ± 0.19
macaroni1	98.71	63.95	98.60	97.85	98.13	99.48	99.73	99.59	99.00 ± 0.13
macaroni2	97.35	79.02	94.65	95.40	96.79	99.33	99.43	98.84	98.20 ± 0.28
pcb1	99.05	27.98	99.29	98.67	99.47	99.65	99.30	98.98	99.61 ± 0.01
pcb2	96.40	59.49	97.12	98.12	97.72	98.28	96.13	98.37	98.03 ± 0.09
pcb3	97.23	76.43	97.04	98.06	98.13	98.98	97.99	98.91	98.45 ± 0.05
pcb4	97.97	83.42	97.51	97.00	97.83	98.29	96.58	95.49	95.22 ± 0.27
pipe_fryum	98.79	75.99	98.72	99.19	98.68	98.6	97.97	98.99	99.29 ± 0.03
Mean	98.04	71.31	97.32	97.09	97.93	**98.68**	97.90	97.51	98.24 ± 0.20

Table 20: Anomaly segmentation performance based on AUPRO on VisA dataset.

Category	CFLOW (2021)	DRÆM (2021)	FastFlow (2021	PaDiM (2021)	PatchCore (2022)	R4AD (2022)	DFR (2020)	EffAD (2024)	ULSAD (Ours)
candle	92.7	80.29	91.65	92.77	94.08	95.30	95.56	77.31	92.49 ± 0.15
capsules	74.64	34.4	81.8	48.42	68.88	92.20	92.09	83.8	82.76 ± 1.18
cashew	93.0	48.33	85.54	82.36	88.01	91.81	90.51	91.57	91.85 ± 1.15
chewinggum	89.58	62.66	84.69	84.33	83.86	88.57	85.52	74.87	84.34 ± 1.0
fryum	85.62	71.94	72.39	75.54	78.25	84.8	92.08	82.93	85.47 ± 0.66
macaroni1	89.46	63.37	91.89	88.55	91.74	95.53	97.59	96.06	92.8 ± 0.74
macaroni2	78.74	56.69	71.94	75.76	87.49	94.01	94.23	89.74	88.29 ± 1.89
pcb1	87.24	27.43	85.89	86.39	89.07	95.0	93.55	90.53	90.22 ± 0.28
pcb2	77.83	33.99	77.99	83.68	83.00	89.17	87.26	90.43	84.53 ± 0.53
pcb3	75.03	71.9	71.33	81.37	79.69	90.89	92.48	92.08	85.86 ± 0.38
pcb4	86.53	73.26	83.41	82.47	84.91	89.17	84.38	75.25	73.37 ± 0.81
pipe_fryum	93.14	31.86	81.89	87.99	92.42	94.76	95.46	68.77	93.49 ± 0.08
Mean	85.29	54.68	81.70	80.80	85.12	<u>91.77</u>	91.72	84.45	87.12 ± 0.89

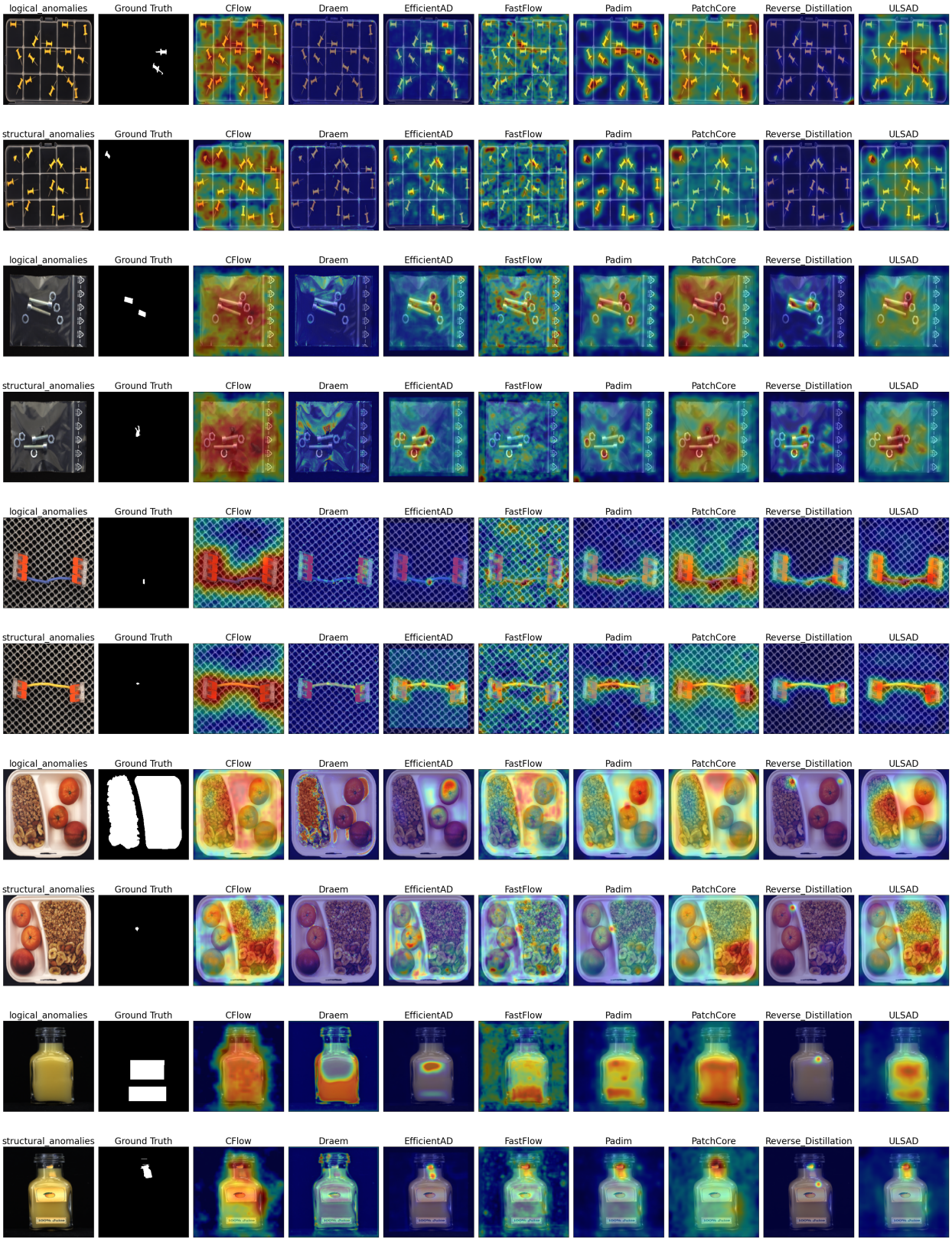


Figure 9: Visualization of anomaly maps on anomalous images from MVTecLOCO dataset.

C Extended Ablations

In this section, we provide additional ablations on the local branch in Table 21 and the total architecture in Table 22. Lastly, in Table 23 we provide the per-category results for the ablation on the pretrained backbone which is summarized in Figure 8.

Table 21: Abalations for local branch. Style: I-AUROC | P-AUROC | P-AUPRO.

category	$\lambda_l = 0$	$\lambda_l = 0.01$	$\lambda_l = 0.5$	$\lambda_l = 0.9$	$\lambda_l = 1.0$
breakfast_box	78.64 | 88.28 | 74.22	79.2 | 88.51 | 74.35	77.86 | 87.79 | 71.27	77.95 | 86.89 | 67.06	79.44 | 86.96 | 65.36
juice_bottle	97.82 | 92.14 | 89.24	97.76 | 92.23 | 89.38	97.56 | 88.78 | 88.16	97.36 | 84.39 | 84.63	97.08 | 83.61 | 83.47
pushpins	72.4 | 69.81 | 65.69	72.77 | 69.84 | 65.68	79.92 | 74.49 | 69.03	76.98 | 74.35 | 63.17	76.53 | 73.69 | 65.18
screw_bag	66.42 | 66.6 | 64.39	67.18 | 68.47 | 65.92	68.06 | 69.13 | 66.22	67.56 | 69.33 | 63.61	66.34 | 69.31 | 62.09
splicing_connectors	73.05 | 59.04 | 73.3	72.84 | 59.15 | 73.29	72.29 | 62.66 | 72.39	72.79 | 64.33 | 70.74	72.36 | 64.5 | 70.57
Mean	77.67 | 75.17 | 73.37	77.95 | 75.64 | 73.72	79.14 | 76.57 | 73.41	78.53 | 75.86 | 69.84	78.35 | 75.61 | 69.33

Table 22: Abalations for total architecture. Style: I-AUROC | P-AUROC | P-AUPRO.

category	$\mathcal{L}_{pg}^{d}$	$\mathcal{L}_{pg}^{d}$; $\mathcal{L}_{lg}$	$\mathcal{L}_{pg}$	$\mathcal{L}_{pg}$; $\mathcal{L}_{lg}$
$\lambda_l = \lambda_g = 0.0$				
breakfast_box	77.29 | 90.41 | 77.16	82.82 | 89.85 | 76.92	66.01 | 87.36 | 67.5	85.08 | 90.19 | 75.36
juice_bottle	96.48 | 92.01 | 88.82	97.93 | 91.82 | 89.26	91.2 | 91.98 | 85.38	97.29 | 92.0 | 89.18
pushpins	70.89 | 80.89 | 70.67	78.66 | 88.09 | 79.11	74.67 | 77.37 | 58.75	74.61 | 85.86 | 76.33
screw_bag	65.48 | 65.87 | 64.04	63.02 | 68.13 | 65.51	61.14 | 59.51 | 57.75	66.93 | 68.67 | 65.35
splicing_connectors	78.29 | 69.68 | 75.61	84.55 | 72.71 | 76.54	65.31 | 53.4 | 66.74	81.5 | 73.11 | 76.01
Mean	77.69 | 79.77 | 75.26	81.4 | 82.12 | 77.47	71.67 | 73.92 | 67.22	81.08 | 81.97 | 76.45
$\lambda_l = \lambda_g = 0.5$				
breakfast_box	79.22 | 90.87 | 78.32	82.45 | 88.49 | 71.03	71.36 | 87.64 | 67.38	83.36 | 89.34 | 72.36
juice_bottle	96.3 | 91.17 | 88.89	98.08 | 87.06 | 88.05	91.14 | 91.84 | 85.92	97.46 | 88.81 | 87.71
pushpins	79.86 | 84.89 | 77.97	82.46 | 87.62 | 80.49	78.64 | 81.14 | 65.12	88.07 | 74.22 | 66.45
screw_bag	66.58 | 68.83 | 66.11	65.11 | 70.04 | 62.81	62.09 | 67.32 | 62.92	70.6 | 71.58 | 67.01
splicing_connectors	80.53 | 73.47 | 75.44	82.85 | 73.03 | 75.13	69.33 | 55.02 | 63.69	81.27 | 74.94 | 74.89
Mean	80.5 | 81.85 | 77.35	82.19 | 81.25 | 75.5	74.51 | 76.59 | 69.01	84.15 | 79.78 | 73.68

Table 23: Abalations for backbone on MvTec-LOCO. Style: I-AUROC | P-AUROC | P-AUPRO.

Class	ResNet50	ResNet152	Wide-ResNet50-2	Wide-ResNet100-2
breakfast_box	82.41 | 89.74 | 73.09 | 61.38	85.11 | 91.15 | 72.0 | 60.06	84.46 | 89.18 | 72.21	82.37 | 89.25 | 74.02 | 62.96
juice_bottle	96.9 | 92.23 | 89.38 | 61.31	97.64 | 91.66 | 89.78 | 61.89	97.11 | 88.9 | 87.94	98.74 | 92.87 | 91.07 | 62.28
pushpins	81.79 | 79.49 | 75.15 | 41.26	73.28 | 76.49 | 63.71 | 23.29	85.48 | 75.46 | 67.82	82.48 | 76.55 | 70.67 | 24.1
screw_bag	66.46 | 69.14 | 67.01 | 21.61	68.06 | 68.63 | 66.97 | 21.2	71.14 | 71.57 | 66.82	73.1 | 68.99 | 66.88 | 21.6
splicing_connectors	80.88 | 72.76 | 74.39 | 34.45	83.53 | 76.31 | 77.24 | 35.9	82.59 | 75.21 | 75.05	84.71 | 79.95 | 78.65 | 42.88
Mean	81.79 | 79.49 | 75.15 | 41.26	81.52 | 80.85 | 73.94 | 40.47	84.16 | 80.06 | 73.97	84.28 | 81.52 | 76.26 | 42.76