Few Shot Part Segmentation Reveals Compositional Logic for Industrial Anomaly Detection

Soopil Kim[1,3], **Sion An**[1], **Philip Chikontwe**[1], **Myeongkyun Kang**[1,3],
Ehsan Adeli[3], **Kilian M. Pohl**[3], **Sang Hyun Park**[1,2]

[1]Robotics and Mechatronics Engineering, DGIST, Daegu, Korea
[2]AI Graduate School, DGIST, Daegu, Korea
[3]Stanford University, Stanford, CA 94305, USA
{soopilkim, shpark13135}@dgist.ac.kr

Abstract

Logical anomalies (*LA*) refer to data violating underlying logical constraints *e.g.,* the quantity, arrangement, or composition of components within an image. Detecting accurately such anomalies requires models to reason about various component types through segmentation. However, curation of pixel-level annotations for semantic segmentation is both time-consuming and expensive. Although there are some prior few-shot or unsupervised co-part segmentation algorithms, they often fail on images with industrial object. These images have components with similar textures and shapes, and a precise differentiation proves challenging. In this study, we introduce a novel component segmentation model for *LA* detection that leverages a few labeled samples and unlabeled images sharing logical constraints. To ensure consistent segmentation across unlabeled images, we employ a histogram matching loss in conjunction with an entropy loss. As segmentation predictions play a crucial role, we propose to enhance both local and global sample validity detection by capturing key aspects from visual semantics via three memory banks: *class histograms*, *component composition embeddings* and *patch-level representations*. For effective LA detection, we propose an adaptive scaling strategy to standardize anomaly scores from different memory banks in inference. Extensive experiments on the public benchmark MVTec LOCO AD reveal our method achieves 98.1% AUROC in LA detection *vs.* 89.6% from competing methods.

Introduction

In industrial images, defects can be categorized into two main types: *structural* and *logical* anomalies (Bergmann et al. 2022). Structural anomalies (*e.g.,* cracks and contamination) occur in localized regions often absent in normal data, whereas logical anomalies refer to data that does not adhere to underlying logical constraints, *e.g.,* component composition and arrangement. Herein, effective detection requires the consideration of long-range dependencies within and across images.

Existing research on anomaly detection (AD) for industrial images has primarily focused on unsupervised approaches that aim to learn the distribution of normal data and detect outliers as anomalies. This has resulted in state-of-the-art models that have reported impressive scores ex-
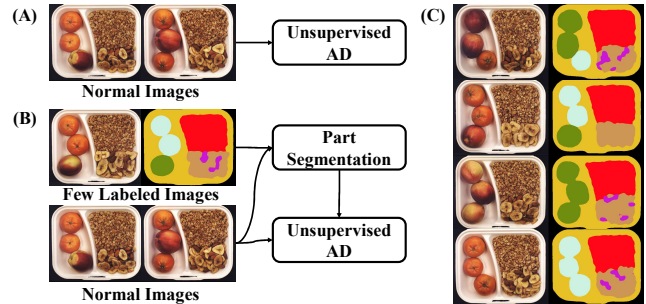
Figure 1: Comparison of approaches at a conceptual level. (A) The anomaly detection (AD) model is directly trained using images. (B) Our proposed method guides part segmentation models using a few labeled samples to accurately segment components and then uses the segments for AD. (C) Examples of logical anomalies show the importance of semantically segmenting components for detection.

ceeding 99% (Roth et al. 2022). This high score can be attributed to the nature of public benchmarks (*e.g.,* MVTec AD (Bergmann et al. 2019) and MTD (Huang et al. 2020)), which predominantly comprises structural anomalies, resulting in models with much lower performance when targeting logical anomalies (Bergmann et al. 2022).

To address logical AD, current methods implicitly consider global dependencies among multiple components for effective detection, as described in Fig. 1A. For example, (Bergmann et al. 2022) proposed a hybrid feature reconstruction model, while (Tzachor, Hoshen et al. 2023) introduced a histogram-based density estimation model. Despite these advancements, performance is constrained by the inability to accurately differentiate various components. For more accurate logical AD, it is essential to semantically segment the product's components, as they often exhibit similar features (*e.g.,* peaches *vs.* mandarins in Fig. 1C). This task is closely related to co-part segmentation (Hung et al. 2019), as normal samples' similar components follow pre-defined logics. However, existing unsupervised methods (Hung et al. 2019; Gao et al. 2021) often fail to precisely segment such components since they cannot distinguish similar features without relying on supervised guidance.

A more effective approach could involve guiding part segmentation using a set of labeled images by employing manufacturers' prior knowledge about the individual elements required for product assembly as shown in Fig. 1B. However, creating pixel-level annotations for numerous training images is a costly and labor-intensive task. While few-shot segmentation methods have made impressive advances to reduce the number of labeled samples (Wang et al. 2023; Hong et al. 2022), they equally fail to segment different parts that have similar textures or shapes. To this end, we introduce a novel part segmentation model tailored to distinguish components in industrial images using few labeled images and several unlabeled images. Specifically, we utilize positional features for prediction and minimize a histogram matching loss for unlabeled images, ensuring each image maintains a consistent number of pixels per class. The combination of different losses enables the model to accurately segment elements across the images.

We integrate accurate part segmentation in our novel AD method, **PSAD** (**P**art **S**egmentation-based **A**nomaly **D**etection). Specifically, PSAD detects local and global dependencies of elements by relying on memory banks for *class histograms*, *component composition embeddings*, and *patch-level representations*. To obtain a unified anomaly score from the different scaled outputs of the memory banks, we propose an adaptive strategy to re-scale anomaly scores using scores from training data. We evaluate the proposed method on a public dataset consisting of both logical and structural AD with five categories. We report results with higher AUROC compared to state-of-the-art not only in logical AD, but also in structural AD. Our contributions can be summarized as follows:

- We propose a novel anomaly detection method PSAD that employs 3 different memory banks by utilizing visual features and semantic segmentation.

- We propose a new part segmentation method that is supervised by a limited number of labeled images with regularization using logical constraints shared across unlabeled images.

- We propose an adaptive scaling method to aggregate anomaly scores with different scales.

- Our method achieves state-of-the-art performance in both logical and structural anomaly detection.

Related Works

Anomaly Detection in Industrial Images: In literature, existing anomaly detection (AD) methods often train models to first learn the distribution of normal data and then detect outliers as anomalies. These methods can be broadly categorized into *reconstruction*, *self-supervision*, and *density estimation-based* models.

Reconstruction-based methods learn to reconstruct normal input samples and determine the anomaly score based on the difference between inputs and the reconstruction (Lee, Lee, and Song 2022; Liu et al. 2023b; Tien et al. 2023). *Self-supervised* methods create synthetic abnormal samples and use them to train a classifier. For instance, CutPaste (Li

et al. 2021) and DRAEM (Zavrtanik et al. 2021) generate abnormal samples for learning abnormality. *Density estimation* methods first extract features from normal samples using pre-trained models and then compare them with test sample features to compute anomaly scores (Roth et al. 2022; Jiang et al. 2022; Hyun et al. 2023). We note that existing methods focus on utilizing local features since most benchmarks mainly contain structural anomalies rather than logical anomalies.

Following the release of the first dataset comprising logical anomalies (Bergmann et al. 2022), several unsupervised methods have been proposed. GCAD (Bergmann et al. 2022) trains local and global models that reconstruct pre-trained image features based on local and global dependencies. SINBAD (Tzachor, Hoshen et al. 2023) extracts a set of orderless elements and randomly projects element features to compute a histogram, with anomaly scores obtained via density estimation. ComAD (Liu et al. 2023a) applied K-Means clustering on pre-trained features to segment multiple components within an image. However, performance was limited because a precise discrimination of different components is challenging.

We observe that product manufacturers are aware of the logical constraints on various components and this prior knowledge can be leveraged for AD. In this paper, we introduce a novel method PSAD using density estimation and semantic segmentation to precisely differentiate components. However, PSAD doesn't demand many labeled images due to our proposed few-shot segmentation method.

When using multiple anomaly scores and aggregating them, previous works simply add the scores (Tsai et al. 2022) or manually set hyper-parameters to scale them (Liu et al. 2023a). However, these approaches may degrade performance when the multiple scores follow different distributions or the hyper-parameters are incorrectly set (Table 3). Even if (Bergmann et al. 2022) attempted to normalize two distinct anomaly scores without defining hyper-parameters, they utilized a validation dataset to determine the statistics of these scores, potentially sacrificing valuable training data. Instead, we propose an adaptive scaling of the scores that solely relies on the training data by treating each sample as a test sample.

Object Part Segmentation: As part segmentation is vital for logical AD, one can train a supervised model for object part segmentation (Chen et al. 2014). However, due to costly pixel labeling, unsupervised models (Sra and Dhillon 2005) that can learn arbitrary segmentation using a collection of unlabeled images are preferable. (Hung et al. 2019) proposed an end-to-end segmentation model with pre-trained CNN features using semantic consistency and geometric concentration losses. Later, (Siarohin et al. 2021) and (Gao et al. 2021) trained a segmentation model using a part-assembly procedure that reconstructs a target image by transforming parts of a source image. Though viable alternatives for industrial image segmentation, learning objectives based on geometric concentration or affine transformations are not generally applicable in industrial images since multiple objects can appear in distant positions (*e.g.,* mandarins in 'breakfast box' and hexagonal nuts in 'screw bag' in MVTec
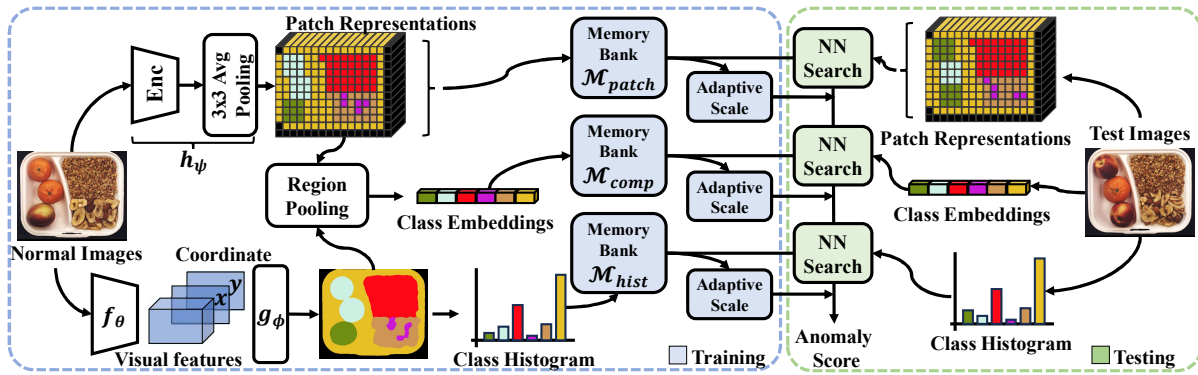
Figure 2: Illustration of PSAD (Part Segmentation-based Anomaly Detection). During training depicted in the blue box, 3 different memory banks are constructed using normal images. The anomaly score of a test image is computed by finding its nearest neighbor (NN search) and adaptive scaling.

LOCO dataset (Bergmann et al. 2022)). In addition, models often under- or over-segment object parts as the labels are arbitrarily optimized. In this paper, we instead propose a new part segmentation model that can segment components in various industrial images using only few labeled samples.

Few Shot Semantic Segmentation: Few-shot segmentation (FSS) has been proposed to overcome the data-hungry nature of deep learning models with different approaches employing generated or augmented images (Mondal, Dolz, and Desrosiers 2018), generative models (Tritrong et al. 2021; Han et al. 2022; Baranchuk et al. 2021), meta-learning (Hong et al. 2022; Kim et al. 2023), transductive inference (Boudiaf et al. 2021), and foundation models (Wang et al. 2023). In general, FSS models employing pre-trained generative models report good part segmentation, especially on several well-aligned images such as face or car. However, generative model training is challenging and requires several samples to guarantee good performance. Note that our method is closely related to the transductive approach RePRI (Boudiaf et al. 2021) that uses a fixed pre-trained backbone and trains a pixel classifier with several regularization losses. During inference, only the classifier (prototype-based) is updated with few samples. While impressive, training is regularized by the initial segmentation which may be often noisy. Thus, we instead update the backbone and the classifier with a histogram matching loss to better utilize logical constraints shared across normal images.

Methods

Problem Setting   Unsupervised anomaly detection (AD) aims to train a model that can identify abnormal data from a set of normal data $\{X^1, ..., X^{N_{train}}\}$ where N_{train} is the number of data and their labels are all assigned as 0 (normal). The model is trained to distinguish between normal and abnormal test data, predicting labels as 0 (normal) or 1 (anomalous).

To detect logical anomalies (LA), accurate part segmentation has to be preceded. In this process, the class of each component is defined by the manufacturer as each normal image contains a predetermined number of specific parts appearing in predefined locations. Consequently, variations in the object's location may lead to different classes, even for the same object (e.g. 'pushpins' and 'splicing connectors' in Fig. 4). Instance segmentation differs from semantic segmentation, as predefined class labels are not assigned to the instances. Since constructing pixel-level annotations of lots of images is labor-intensive, we assume that only a scarce number of labeled images $\{X^{l,i} \in \mathbb{R}^{W \times H \times 3}, Y^{l,i} \in \mathbb{R}^{W \times H \times N_C}\}_{i=1}^{N_L}$ and a substantial set of unlabeled images $\{X^{u,j}\}_{j=1}^{N_U}$ are provided for training. Here, N_L, N_U, and N_C represent the numbers of labeled images, unlabeled images, and classes, respectively. The model is optimized to reduce a combination of supervised losses for X^l and unsupervised losses for X^u.

Overview   Our proposed PSAD (Part Segmentation-based Anomaly Detection) consists of two parts: *semantic part segmentation* and *AD using part segmentation*. For part segmentation, we design a model to distinguish multiple components based on visual and positional features (Fig. 3). A visual feature extractor and a pixel classifier are jointly optimized with a few labeled images and logical constraints shared across numerous unlabeled images. For AD with part segmentation, the segmentation model applied to normal samples is leveraged to construct three distinct memory banks (Fig. 2). In particular, (1) a class histogram memory bank $\mathcal{M}_{hist}$ that records the quantity and arrangement of each component to assess the relative abundance of different components within the images. (2) a component compositions memory bank $\mathcal{M}_{comp}$ that helps determine the validity of various compositions of components to identify anomalies that arise from unexpected or irregular component arrangements. Finally, (3) a memory bank $\mathcal{M}_{patch}$ specifically designed for patch-level features to capture fine-grained details within the image.

From these memory banks, we generate three different anomaly scores, each of which has a different scale and distribution. To effectively compare and combine the scores, we perform adaptive scaling using statistics obtained from the training data to ensure scores can be reliably compared
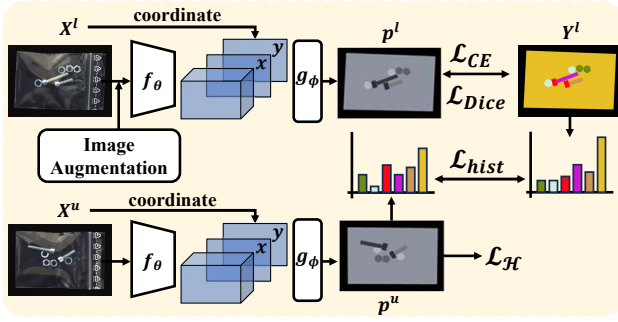
Figure 3: Proposed part segmentation model that predicts segmentation utilizing visual and positional features.

across different scales.

Part Segmentation Using Limited Annotations

The segmentation model consists of a feature extractor f_θ and a pixel classifier g_ϕ, with $f_\theta(X)$ being a feature map having the same size as input X. Since the object's location is important, the pixels' coordinate $c \in \mathbb{R}^{W \times H \times 2}$ and $f_\theta(X)$ are concatenated as input for g_ϕ. During training, given model prediction probability $p = g_\phi(f_\theta(X) \oplus c)$, parameters θ and ϕ are optimized via:

$$\mathcal{L} = \mathcal{L}_{Dice} + \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{\mathcal{H}} + \lambda_3 \mathcal{L}_{hist}, \qquad (1)$$

where each λ is a hyper-parameter. For labeled images X^l, our model relies on cross-entropy loss $\mathcal{L}_{CE}$ and dice similarity loss $\mathcal{L}_{Dice}$. Note that prediction on unlabeled images X^u may be uncertain, especially with limited X^l. A common approach to handle this is by incorporating an entropy loss $\mathcal{L}_{\mathcal{H}}$ to reduce uncertainty (Wang et al. 2022). Nonetheless, minimizing $\mathcal{L}_{\mathcal{H}}$ with only a few labeled images can lead to unexpected training outcomes and potentially degrade accuracy. To mitigate this, we propose a histogram matching loss $\mathcal{L}_{hist}$ to ensure consistency in segmenting each part with an equal number of pixels. We randomly select a label Y^l from $\{Y^{l,i}\}_{i=1}^{N_L}$ and compare the class-level volume with predictions p^u from unlabeled images:

$$\mathcal{L}_{hist} = \frac{1}{N_C} \sum_{n=1}^{N_C} \left\| \frac{1}{WH} \sum_{w,h} Y^l_{w,h,n} - \frac{1}{WH} \sum_{w,h} p^u_{w,h,n} \right\|$$
$$(2)$$

While model parameters are updated to reduce uncertain predictions under the constraints from $\mathcal{L}_{CE}$, $\mathcal{L}_{Dice}$, and $\mathcal{L}_{hist}$, the model also learns consistent segmentation on numerous unlabeled images based on both the visual and positional similarity of each component.

Handling Multiple Types of Products   In industrial image datasets, products may be composed of various subtypes (e.g. 'juice bottle' and 'splicing connector' in the MVTec LOCO AD dataset). In such cases, it is necessary to ensure that X^l and X^u belong to the same product type for the comparison in $\mathcal{L}_{hist}$. To classify X^u without human annotation, we compare unlabeled and labeled images in latent space

and find the nearest labeled image for each unlabeled image. Specifically, we extract global-average-pooled features from the images using a pre-trained encoder before training the segmentation model. Each X^u's type is determined as the type of the nearest labeled images in the latent space. Subsequently, X^l and X^u of the same type can be used together in $\mathcal{L}_{hist}$. As a result, our model can effectively handle datasets that contain multiple types of products.

Anomaly Detection Using Part Segmentation

Our proposed PSAD follows a density estimation approach (Defard et al. 2021) in which normal data features are stored in a memory bank $\mathcal{M} = \{e^k\}_{k=1}^{N_\mathcal{M}}$, with $N_\mathcal{M}$ denoting the number of elements in $\mathcal{M}$. To determine the anomaly score s of a test sample e_{test}, we find its nearest neighbor among elements in $\mathcal{M}$:

$$s = \arg \min_{e \in \mathcal{M}} \|e_{test} - e\|^2 . \qquad (3)$$

Patch-level density-based methods (Roth et al. 2022; Jiang et al. 2022) have proven to be effective in detecting structural anomalies by focusing on local features. However, logical anomalies often arise when multiple components appear together to form a single product or entity.

Class Histogram Memory Bank   The first memory bank $\mathcal{M}_{hist}$ focuses on quantifying the number of components for each class through a class histogram. Given normal images and their corresponding segmentation, we construct a histogram that represents the distribution of pixels among different classes. The histograms are then stored in $\mathcal{M}_{hist}$ and respective anomaly scores are predicted using Eq.(3).

Component Composition Memory Bank   It is worth noting that solely relying on $\mathcal{M}_{hist}$ cannot verify whether the components are combined correctly or not. To address this, we introduce a component composition memory bank $\mathcal{M}_{comp}$ that stores feature compositions of different parts within an image. After a feature map $w = h_\psi(X)$ is extracted using a pre-trained encoder h_ψ, the segmentation map allows us to define a class embedding as an averaged feature vector of pixels belonging to each class. A concatenation of these class embeddings is saved in $\mathcal{M}_{comp}$ to effectively capture visual features of each component and their compositions within the image. The anomaly score is predicted using Eq.(3).

Patch Representation Memory Bank   Finally, we construct memory bank $\mathcal{M}_{patch}$ by storing patch-level representations to detect fine-grained features following established approaches (Defard et al. 2021; Roth et al. 2022): $\mathcal{M}_{patch} = \bigcup_{k=1}^{N_{train}} \{h_\psi(X^k)_l\}_{l=1}^{N_P}$, where $h_\psi(X^k)_l$ is the lth patch representations extracted from X^k and N_P denotes the number of patches. The anomaly score of X_{test} is predicted as: $s = \max_{e \in \{h_\psi(X_{test})_l\}_{l=1}^{N_P}} \min_{e' \in \mathcal{M}_{patch}} \|e - e'\|^2$.

Aggregating Anomaly Scores of Different Scales   Considering the distinct scales and distributions of the memory banks, it is essential to set appropriate hyperparameters for each anomaly score as arbitrarily configuring a

single hyperparameter can negatively impact overall accuracy. To mitigate this, our solution is scaling s based on the anomaly scores of training data in each memory bank. In particular, we derive a set of anomaly scores denoted as $S_{train} = \{s^1, ..., s^{N_\mathcal{M}}\}$ from the training data by treating each data point e^k as a test sample. We then construct the memory bank using all other training samples excluding e^k as follows: $s^k = \min_{e \in \mathcal{M}, e \neq e^k} \left\| e^k - e \right\|^2$. In the context of $\mathcal{M}_{hist}$ and $\mathcal{M}_{comp}$, e stands for a class histogram and a component composition embedding derived from a data sample, respectively. However, for $\mathcal{M}_{patch}$, s^k is defined in a different way as multiple elements are saved from a data sample as: $s^k = \max_{e \in \{h_\psi(X^k)_l\}_{l=1}^{N_P}} \min_{e' \in \mathcal{M}'_{patch}} \left\| e - e' \right\|^2$, where $\mathcal{M}'_{patch} = \bigcup_{m=1, m\neq k}^{N_{train}} \{h_\psi(X^m)_l\}_{l=1}^{N_P}$.

We define a normalized anomaly score considering the statistics of S_{train} as $\hat{s}_\mathcal{M} = s / \max\{s^1, ..., s^{N_\mathcal{M}}\}$. This adaptive scaling approach improves accuracy and robustness in detecting anomalies. The final anomaly score is defined as a sum of three anomaly scores from different memory banks: $s = \hat{s}_{\mathcal{M}_{hist}} + \hat{s}_{\mathcal{M}_{comp}} + \hat{s}_{\mathcal{M}_{patch}}$, facilitating both structural and logical anomaly scoring.

Implementation Details: We use a pre-trained Wide ResNet101 (Zagoruyko and Komodakis 2016) for initializing the parameters of the segmentation model f_θ. Among 4 convolutional blocks in f_θ, features extracted from the first 3 blocks are resized to the size of input X and concatenated to obtain $v \in \mathbb{R}^{W \times H \times (256+512+1024)}$. Labeled images were augmented following (Buslaev et al. 2020). For training, we used an AdamW optimizer with a learning rate 0.001 and batch size of 5 per iteration on an NVIDIA RTX A5000 GPU workstation. The model was first trained for 50 epochs using only $\mathcal{L}_{CE}$ and $\mathcal{L}_{Dice}$. After warming up with the supervised loss, the model is trained using Eq.(1) for additional 50 epochs. As $\mathcal{L}_{Dice}$ is usually larger than the other losses, hyper-parameters λ_1, λ_2, and λ_3 were set as 10. h_ψ has Wide ResNet101 as the visual feature encoder and a 3×3 average pooling operation following the setting of PatchCore (Roth et al. 2022) which is one of the state-of-the-art models proposed for detecting structural anomalies.

Experiments

Experimental Setting: We evaluated our method on MVTec LOCO AD dataset (Bergmann et al. 2022), the only benchmark for detecting logical anomalies to the best of our knowledge. This dataset consists of 5 categories (breakfast box/juice bottle/pushpins/screw bag/splicing connectors). For each category, 351/335/372/360/360 normal images were used for training and 275/330/310/341/312 images for testing following the setting of the comparison methods. Test data is categorized into good, structural anomaly (SA), and logical anomaly (LA). For the segmentation task, we used 5 labeled images since existing FSS models show more stable accuracy on the 5-shot setting. If the products have multiple types (*e.g.,* 3 types within 'juice bottle' and 'splicing connectors'), we created a labeled image for each type, thereby employing a total of 3 labeled

images. State-of-the-art AD methods including PatchCore (Roth et al. 2022), RD4AD (Deng and Li 2022), DRAEM (Zavrtanik et al. 2021), AST (Rudolph et al. 2023), ST (Bergmann et al. 2020), ComAD (Liu et al. 2023a), GCAD (Bergmann et al. 2022), SINBAD (Tzachor, Hoshen et al. 2023), and SLSG (Yang et al. 2023) are used as comparison methods. The models are evaluated on SA and LA detection separately. We resized all images so that the number of pixels on the longer side among width and height is 512. The area under the ROC curve (AUROC) is used as a metric following previous works (Roth et al. 2022).

Comparison with State-of-the-art Methods: Table 1 lists the AUROC for LA and SA detection of methods trained and tested on MVTec LOCO AD dataset. Existing methods designed to focus on local features (e.g., PatchCore, RD4AD, DRAME, AST, ST) have lower scores in LA, as they can not capture the global dependencies among multiple components in the image, despite showing better scores in SA detection. While recent methods for detecting LA (e.g. ComAD, GCAD, SINBAD, SLSG) show improvements over those focusing on local features, they still fail to precisely distinguish between different components within the image. On the other hand, our proposed PSAD shows significant gains over others (i.e. +8.5 avg AUROC score across 5 categories in LA detection). Notably, we achieved a 99.3% AUROC score in the 'screw bag' category, while the highest accuracy among the other methods in this category was only 80.1%. In addition, our proposed method also showed the best average score in SA detection. These results show that using semantic information can be beneficial for detecting both LA and SA. As a result, our proposed method obtained the best AUROC scores in both tasks.

Qualitative Comparison of FSS Methods: In Fig. 4, we show a qualitative comparison of different segmentation models on MVTec LOCO AD dataset. When we applied unsupervised co-part segmentation models SCOPS (Hung et al. 2019) and Part-Assembly (Gao et al. 2021), we obtained arbitrary segmentation results that can not discriminate between different components that are supposed to be segmented into different classes. In some cases, Part-Assembly failed to obtain proper results, as they focus on single-body objects. We also evaluated various state-of-the-art FSS models: a meta-learning-based model VAT (Hong et al. 2022), a foundation model SegGPT (Wang et al. 2023), and a transductive model RePRI (Boudiaf et al. 2021). VAT and RePRI, each with ResNet-101 backbone pre-trained on PASCAL-5^i dataset (Shaban et al. 2017) showed poor performance in most cases. This is due to (1) frozen encoder during training, (2) encoder pre-trained on different domain data, and (3) models relying on inaccurate/noisy initial predictions. Though SegGPT showed relatively good results in some categories such as 'juice bottle', it showed limited performance when multiple components have similar textures but different classes. For example in Fig. 4, SegGPT fails to distinguish the left and right parts of 'splicing connectors' and the short and long bolts in 'screw bag', as they share similar textures. The limitations of existing FSS methods are mainly attributed to training models with existing segmentation datasets that do not necessitate considering position

	Category	PatchCore	RD4AD	DRAEM	ST	AST	GCAD	SINBAD	ComAD	SLSG	**PSAD**
LA	Breakfast Box	74.8	66.7	75.1	68.9	80.0	87.0	96.5	91.1	-	**100.0**
	Juice Bottle	93.9	93.6	97.8	82.9	91.6	**100.0**	96.6	95.0	-	99.1
	Pushpins	63.6	63.6	55.7	59.5	65.1	97.5	83.4	95.7	-	**100.0**
	Screw Bag	57.8	54.1	56.2	55.5	80.1	56.0	78.6	71.9	-	**99.3**
	Splicing Connectors	79.2	75.3	75.2	65.4	81.8	89.7	89.3	**93.3**	-	91.9
	Average (LA)	74.0	70.7	72.0	66.4	79.7	86.0	88.9	89.4	89.6	**98.1**
SA	Breakfast Box	80.1	60.3	85.4	68.4	79.9	80.9	**87.5**	81.6	-	84.9
	Juice Bottle	98.5	95.2	90.8	**99.3**	95.5	98.9	93.1	98.2	-	98.2
	Pushpins	87.9	84.8	81.5	90.3	77.8	74.9	74.2	**91.1**	-	89.8
	Screw Bag	92.0	89.2	85.0	87.0	**95.9**	70.5	92.2	88.5	-	95.7
	Splicing Connectors	88.0	95.9	95.5	**96.8**	89.4	78.3	76.7	94.9	-	89.3
	Average (SA)	89.3	85.1	87.6	88.4	87.7	80.7	84.7	90.9	91.4	**91.6**
	Average	81.7	77.9	79.8	77.4	83.7	83.4	86.8	90.1	90.3	**94.9**

Table 1: Performance comparison of the proposed model PSAD against state-of-the models on MVTec LOCO AD dataset. LA and SA denote logical and structural anomalies, respectively. Boldface represents the best score.

Models	LA	SA
SCOPS (Hung et al. 2019)	82.5	90.2
Part-Assembly (Gao et al. 2021)	80.3	85.6
SegGPT (Wang et al. 2023)	88.7	87.2
VAT (Hong et al. 2022)	79.2	87.8
RePRI (Boudiaf et al. 2021)	83.6	88.4
Ours ($\mathcal{L}_{sup}$)	95.9	89.6
Ours ($\mathcal{L}_{sup} + \mathcal{L}_{\mathcal{H}}$)	96.3	90.0
Ours ($\mathcal{L}_{sup} + \mathcal{L}_{\mathcal{H}} + \mathcal{L}_{hist}$)	**98.1**	**91.6**

Table 2: Average AUROC scores of our proposed PSAD using different FSS models.

$\mathcal{M}_{hist}$	$\mathcal{M}_{comp}$	$\mathcal{M}_{patch}$	AS	LA	SA
✓				94.2	71.1
	✓			90.9	85.4
		✓		73.9	89.3
✓	✓	✓		96.8	87.6
✓	✓	✓	✓	**98.1**	**91.6**

Table 3: Average AUROC scores of our proposed PSAD using different combinations of memory banks. 'AS' stands for adaptive scaling.

or length comparisons. As most meta-learning FSS models are focused on discriminating classes based on texture and shape, they equally show limited accuracy on industrial images which may have multiple similar components belonging to different classes.

When we trained our model using only supervised losses for labeled images, we obtained better results in 'pushpins' and 'splicing connectors' but poor predictions in other categories despite using $\mathcal{L}_{\mathcal{H}}$. For example, it failed to discriminate between the short and long bolts in 'screw bag'. However, when $\mathcal{L}_{hist}$ was employed, accurate segmentation results were obtained on various types of products. This shows that using $\mathcal{L}_{hist}$ with the other loss functions is more beneficial to obtain consistent segmentation by leveraging logical constraints.

Anomaly Detection Using Different Segmentation Models: Based on the segmentation results, we also check

whether accurate segmentation correlates with AD performance. Table 2 shows the average AUROC scores of our methods using different segmentation models. FSS (Seg-GPT, VAT, and RePRI) and unsupervised models (SCOPS and Part-Assembly) showed low performance in LA detection, as they under-perform in segmenting some categories' data such as 'pushpins'. Nevertheless, it is worth noting that their LA detection scores are higher than the score of our baseline PatchCore. This shows that leveraging even imperfect segmentation can be beneficial for LA detection.

On the other hand, our model trained with $\mathcal{L}_{sup}$ (= $\mathcal{L}_{CE} + \mathcal{L}_{Dice}$) showed significantly improved scores even if it is not as accurate as our final model. This shows that our approach which jointly trains the encoder and classifier utilizing image augmentations and positional information is beneficial for segmenting industrial image data. When we use $\mathcal{L}_{\mathcal{H}}$ and $\mathcal{L}_{hist}$ together, we obtained further improved LA detection scores. Interestingly, the detection performance of SA is also enhanced with the utilization of more accurate segmentation results. These findings show the crucial role of accurate segmentation in achieving precise AD.

Effect of Various Memory Banks And Adaptive Scaling: Table 3 shows AD performance using different combinations of memory banks. When each memory bank is employed alone, $\mathcal{M}_{hist}$ and $\mathcal{M}_{comp}$ showed good performance on LA detection but poor performance on SA detection, whereas $\mathcal{M}_{patch}$ showed the opposite case. When we add these anomaly scores from 3 different memory banks without any scaling strategy, we obtained a better LA detection score and a degraded SA detection score. It is mainly attributed to varying scales of anomaly scores depending on memory banks. When we apply adaptive scaling, we obtained the best scores in both LA and SA detection.

Figure 5 illustrates histograms of anomaly scores obtained from various memory banks and the unified anomaly scores after adaptive scaling. It shows that the scale of anomaly scores varies across memory banks, implying the importance of scaling scores before aggregation. Notably, the anomaly scores from the patch representation memory bank are poor due to its reliance on local features. Nevertheless, after normalizing each score and integrating them into a uni-
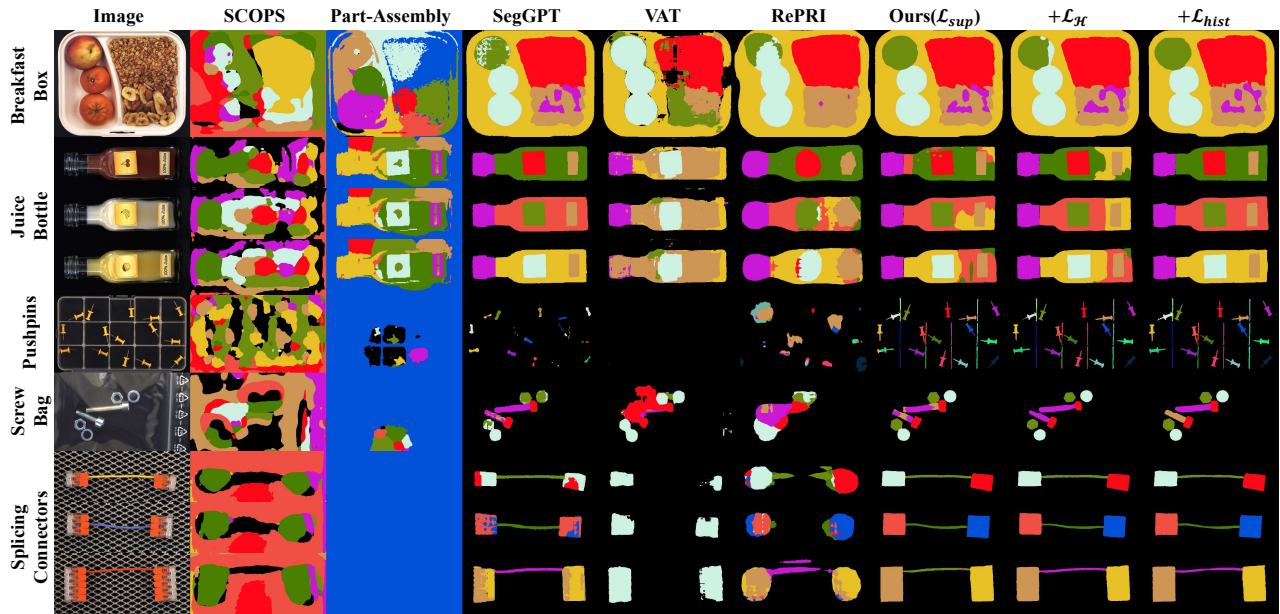
Figure 4: Qualitative comparison of FSS models. $\mathcal{L}_{sup}$ ($= \mathcal{L}_{CE} + \mathcal{L}_{Dice}$) denotes a supervised loss for labeled images. For the unsupervised methods, such as SCOPS and Part-Assembly, we arbitrarily set the number of parts as 10.
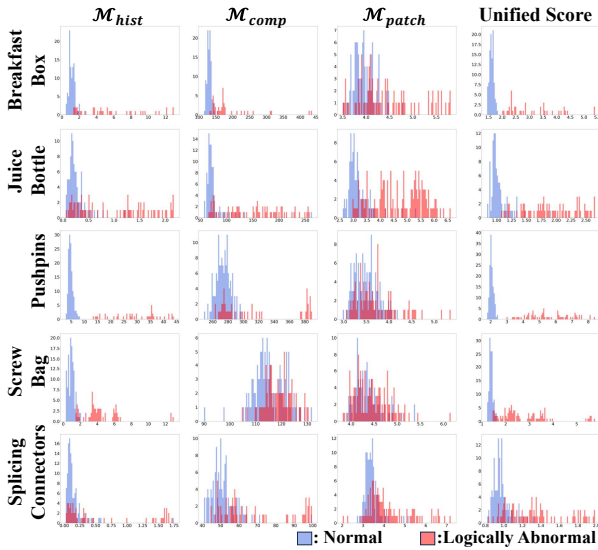

Figure 5: Histogram visualizations of anomaly scores from different memory banks and the unified anomaly scores.

$N_{\mathcal{M}}$	100%	50%	25%	12.5%
Avg AUROC	97.4	97.1	96.6	96.2

Table 4: AUROC in LA detection of our proposed PSAD using different numbers of normal images N_{train}. In this experiment, a combination of $\mathcal{M}_{hist}$ and $\mathcal{M}_{comp}$ is used.

still outperforms the other methods with the reduced dataset. This finding underscores the significance of accurate segmentation maps in enabling precise LA detection even with limited data.

fied score, a clear discrimination between normal and abnormal samples was observed. Overall, these findings highlight the significance of adaptive scaling in improving the effectiveness of AD using multiple memory banks.

Logical Anomaly Detection Using Less Training Samples: Table 4 lists the AUROC of LA detection using varying numbers of unlabeled images. Despite a slight decrease in the average AUROC scores using less data, our approach

Conclusion

In this paper, we incorporate part segmentation into anomaly detection (AD) to detect logical and structural anomalies. To avoid constructing a large training dataset for segmentation, we propose a new segmentation model that utilizes a few labeled images and logical constraints shared across normal images. We also propose a novel AD method that involves constructing 3 distinct memory banks based on the segmentation. To generate a unified anomaly score from varying scales of anomaly scores, we introduce an adaptive scaling strategy. By doing so, our model could detect LA and SA, and yields substantial improvements with minimal effort required from users. As future few-shot segmentation models evolve to require fewer labeled images and produce better results, our AD model will achieve more enhanced performance with less effort from users.

Acknowledgments

References

Baranchuk, D.; Rubachev, I.; Voynov, A.; Khrulkov, V.; and Babenko, A. 2021. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*.

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

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9592–9600.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2020. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4183–4192.

Boudiaf, M.; Kervadec, H.; Masud, Z. I.; Piantanida, P.; Ben Ayed, I.; and Dolz, J. 2021. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13979–13988.

Buslaev, A.; Iglovikov, V. I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; and Kalinin, A. A. 2020. Albumentations: fast and flexible image augmentations. *Information*, 11(2): 125.

Chen, X.; Mottaghi, R.; Liu, X.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1971–1978.

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

Deng, H.; and Li, X. 2022. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9737–9746.

Gao, Q.; Wang, B.; Liu, L.; and Chen, B. 2021. Unsupervised co-part segmentation through assembly. In *International Conference on Machine Learning*, 3576–3586. PMLR.

Han, M.; Zheng, H.; Wang, C.; Luo, Y.; Hu, H.; and Du, B. 2022. Leveraging GAN priors for few-shot part segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1339–1347.

Hong, S.; Cho, S.; Nam, J.; Lin, S.; and Kim, S. 2022. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, 108–126. Springer.

Huang, Y.; et al. 2020. Surface defect saliency of magnetic tile. *The Visual Computer*, 36: 85–96.

Hung, W.-C.; Jampani, V.; Liu, S.; Molchanov, P.; Yang, M.-H.; and Kautz, J. 2019. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 869–878.

Hyun, J.; Kim, S.; Jeon, G.; Kim, S. H.; Bae, K.; and Kang, B. J. 2023. ReConPatch: Contrastive Patch Representation Learning for Industrial Anomaly Detection. *arXiv preprint arXiv:2305.16713*.

Jiang, X.; Liu, J.; Wang, J.; Nie, Q.; Wu, K.; Liu, Y.; Wang, C.; and Zheng, F. 2022. Softpatch: Unsupervised anomaly detection with noisy data. *Advances in Neural Information Processing Systems*, 35: 15433–15445.

Kim, S.; Chikontwe, P.; An, S.; and Park, S. H. 2023. Uncertainty-aware semi-supervised few shot segmentation. *Pattern Recognition*, 137: 109292.

Lee, S.; Lee, S.; and Song, B. C. 2022. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10: 78446–78454.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9664–9674.

Liu, T.; Li, B.; Du, X.; Jiang, B.; Jin, X.; Jin, L.; and Zhao, Z. 2023a. Component-aware anomaly detection framework for adjustable and logical industrial visual inspection. *arXiv preprint arXiv:2305.08509*.

Liu, W.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2023b. Diversity-measurable anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12147–12156.

Mondal, A. K.; Dolz, J.; and Desrosiers, C. 2018. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. *arXiv preprint arXiv:1810.12241*.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; and Wandt, B. 2023. Asymmetric student-teacher networks for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2592–2602.

Shaban, A.; Bansal, S.; Liu, Z.; Essa, I.; and Boots, B. 2017. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*.

Siarohin, A.; Roy, S.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2021. Motion-supervised co-part segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 9650–9657. IEEE.

Sra, S.; and Dhillon, I. 2005. Generalized nonnegative matrix approximations with Bregman divergences. *Advances in neural information processing systems*, 18.

Tien, T. D.; Nguyen, A. T.; Tran, N. H.; Huy, T. D.; Duong, S.; Nguyen, C. D. T.; and Truong, S. Q. 2023. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24511–24520.

Tritrong, N.; et al. 2021. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4475–4485.

Tsai, C.-C.; et al. 2022. Multi-scale patch-based representation learning for image anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3992–4000.

Tzachor, N. C.; Hoshen, Y.; et al. 2023. Set features for fine-grained anomaly detection. *arXiv preprint arXiv:2302.12245*.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*.

Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4248–4257.

Yang, M.; Liu, J.; Yang, Z.; and Wu, Z. 2023. SLSG: Industrial Image Anomaly Detection by Learning Better Feature Embeddings and One-Class Classification. *arXiv preprint arXiv:2305.00398*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zavrtanik, V.; et al. 2021. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330–8339.