Sam2Sim: Leveraging Segment Anything Model for Environment-Adaptive Anomaly Detection through Component-Aware SimpleNet

Leng Seng Hak, *Kim Hyungwon
Chungbuk National Univ.

{leng_senghak, *hwkim}@chungbuk.ac.kr

Abstract

Anomaly detection is an essential essence in many industrial applications, where it is used to identify defects or deviating patterns. In most cases, however, anomaly samples are rare in the training data, which puts a serious limitation on the ability of supervised learning methods to cope with anomalies. Traditional methods address this challenge by synthesizing anomalies or unsupervised learning based on reconstruction. Although such types of methods have shown some promising results, they usually presuppose static conditions and cannot adapt under dynamic conditions. The paper proposes a brand-new ensemble model, Sam2Sim, that makes use of foreground extraction based on the Segment Anything Model 2 and anomaly detection using the SimpleNet architecture. Filtering attention on foreground objects by removal of background noises confers robustness to environmental changes and low false positives in Sam2Sim. The Sam2Sim is evaluated on the MVTec dataset with the insertion of a pseudo-background mechanism mimicking real-world scenarios during testing. Experiments conducted on the results have improved image-level anomaly detection and localization in the dynamic background setting. This obviously shows the great potential of Sam2Sim in the reliable and efficient detection of anomalies in dynamic industrial environments.

Keywords: Anomaly detection, SimpleNet, SAM2, Foreground Extraction

Ⅰ. Introduction

Anomaly detection is the study and identification of defects or deviations from the normal pattern of occurrences in samples. One of the most important challenges is that the training data often has a shortage of anomalous samples, making it difficult to apply in supervised learning. To address this issue, researchers have suggested various strategies. One is the synthesis of anomalous patterns during training [1], [2], [3], which forces the model to learn the identification process of these anomalous characteristics - a technique known as similarity-based anomaly detection. Another strategy is unsupervised learning based on a reconstruction-based anomaly detection model [4], [5]. The underlying hypothesis is that a reconstruction-based model can only restore the samples of normality seen during the training period, and it will not be able to reconstruct the anomalous part because it is missing in the training data. Previous works have often assumed a prior static environment hypothesis, though the obtained empirical results have not been explicitly presented. Recently, some component-aware schemes have shown promising performances with a logical set of constraints [6], [7], but the application in object-based detection tasks becomes difficult under dynamic environments. When it comes to adapting models accurately, real-life conditions such as climatic variations can pose challenges to this situation.

In this paper, we propose an ensemble model that focuses on the foreground via background removal, leveraging the Segment Anything Model 2 (SAM 2) [8], in conjunction with anomaly identification and localization using the SimpleNet [3] architecture, we called it Sam2Sim. The model is evaluated on MVTec dataset. A pseudo-background insertion mechanism is further used during the inference stage to measure model performance in scenarios closer to reality. The remainder of this paper is organized as follows: Section II discusses our proposed Sam2Sim method in detail followed by Section III which describes our experimental outcomes from this study. Final Section IV provides a brief conclusion and remarks.

Ⅱ. Methodology

The Sam2Sim methodology can be referred to as a two-step process for image analysis. As illustrates in Figure 1, the first stage is the background removal block based on the SAM2 model. This block segregates the foreground from the background and removes noise that interferes with the processing of the image. The SAM 2 block runs on pre-trained weights and does not undergo backpropagation in the training or inference phase.

The proposed methodology first categorizes the input image as either object-type or texture-type. In the case of object-type images, a foreground extractor is applied to remove the background. The principle involved in this method is that texture-based methods are more resistant to changes in the environment than object-type images. The second step of the methodology is represented through the SimpleNet model, a four-component architecture integrated with the Sam2Sim approach. Particularly, SimpleNet includes the following modules: a Feature Extractor, a Feature Adaptors, a generator of Anomalous Features, and a Discriminator. The same multi-stage neural network architecture that takes the pre-processed image, after background removal, further in the pipeline for analysis and processing.

Feature Extractor: The feature extractor extracts local features from images. For any given image in the training or test set, a set of features is extracted from the different hierarchical levels by a pre-trained network. As the pre-trained network has already been biased to the dataset on which it was trained, only a subset of all possible hierarchical levels will be selected for the target dataset. Here, the neighbourhood is defined as a patch of size around the location, and the local features are obtained through an aggregate function aggregating the features within the neighbourhood. Then, all feature maps are linearly resized into the same size and channel-wisely concatenated to get the final feature map.

Feature Adaptors: Feature adaptor is used to adapt the features learned during training to the target domain for the distribution shift between industrial images and the dataset pre-trained for the backbone model. Specifically, the Feature Adaptor projects the local feature to an adapted one by some simple neural blocks, like a fully connected layer or a multi-layer perceptron (MLP).

Anomalous Feature Generator: The Sam2Sim method trains the Discriminator using anomalous features generated from the addition of Gaussian noise to normal features. This makes an efficient distinction compared to the synthesis of defect images since the compact feature space allows distinguishing anomalous from normal features by the Discriminator.

Discriminator: The Sam2Sim Discriminator is a 2-layer MLP that estimates the normality score for each spatial location, using both normal and anomalous features during training.

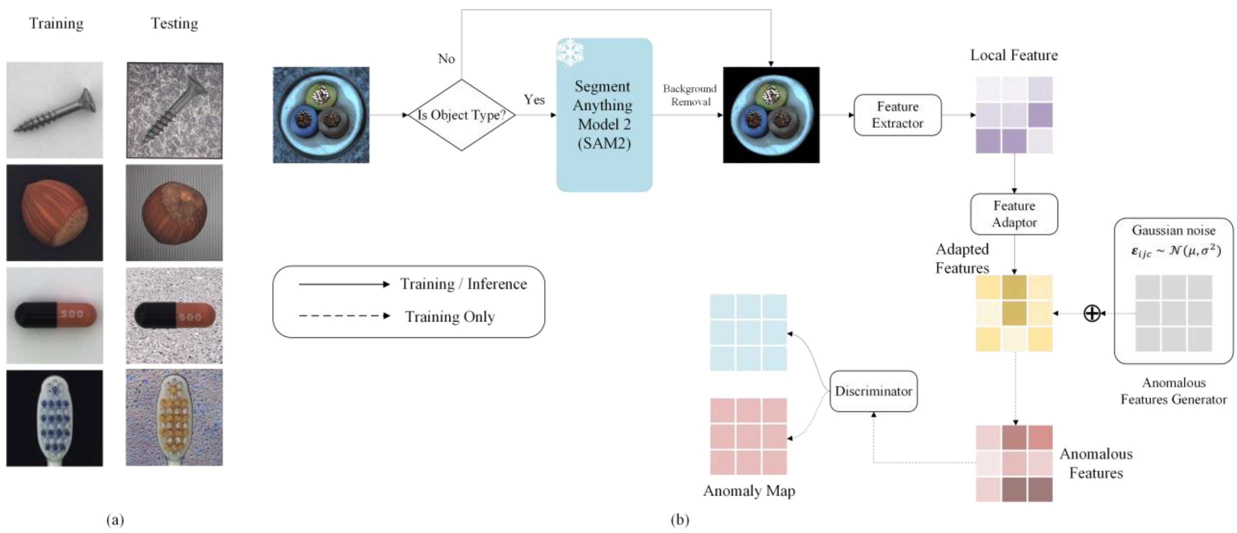

Figure 1. (a) The image during the training and the pseudo background during the testing for dynamics background during testing (b) Proposed Sam2Sim methodology, the model will check if it is object type or texture type to decide whether to undergo the background removal process before input to SimpleNet model and perform anomaly detection

In the process of training, Sam2Sim uses a similar architecture of SimpleNet and exploits identical backpropagation [3]. The loss function used is the L1 loss, also known as the absolute loss or Manhattan distance. The training objective is to minimize the average of these truncated L1 losses across all training samples and spatial locations. This acts as an incentive for the Discriminator component to be truthful about normal and anomalous features but penalizes it from being very sure of its predictions. The image-level anomaly detection performance is evaluated using the standard Area Under the Receiver Operator Curve (I-AUROC) metric and the pixel-wise AUROC (P-AUROC) will be used for evaluating the anomaly localization.

III. Results and Discussion

The experiments were conducted primarily on the MVTec AD benchmark, comprising 5,354 images across 15 categories (5 texture and 10 object classes). The SAM2 tiny model was employed for background removal, with ImageNet-pretrained backbones and a WideResNet50 architecture. Anomaly feature generation involved adding Gaussian noise to normal features, while the discriminator utilized linear, batch norm, leaky ReLU, and linear layers. Training was performed using the Adam optimizer with specific learning rates and weight decay over 160 epochs. Background removal was applied to all object classes except "transistor," where the background's role in anomaly detection necessitated its inclusion.

Table 1. Comparison of generic SimpleNet with static background during training and inference, dynamic background settings

Model	SimpleNet	SimpleNet	Sam2Sim
Type	Static Background	Dynamics Background	
Carpet	0.99518/0.97867	0.99318/0.97625	0.99318/0.97625
Grid	0.99331/0.98692	0.99666/0.98761	0.99666/0.98761
Leather	1.0/0.99161	1.0/0.99209	1.0/0.99209
Tile	0.99783/0.95617	0.99892/0.94347	0.99892/0.94347
Wood	1.0/0.93695	1.0/0.93731	1.0/0.93731
Avg. Text.	0.99726/0.97006	0.99775/0.96735	0.99775/0.96735
Bottle	1.0/0.98047	0.54127/0.54343	**1.0/0.98061**
Cable	1.0/0.97438	0.61694/0.79009	**0.99269/0.97238**
Capsule	0.97806/0.98908	0.58037/0.32428	**0.96490/0.98983**
Hazelnut	0.99928/0.97274	0.49036/0.39007	**0.96393/0.97515**
Metal Nut	1.0/0.98476	0.47019/0.46334	**1.0/0.98648**
Pill	0.98717/0.98295	0.58783/0.43707	**0.97654/0.98761**
Screw	0.980733/0.992592	0.57614/0.36167	**0.93810/0.97960**
Toothbrush	1.0/0.984590	0.57500/0.40274	**0.96667/0.98133**
Transistor	1.0/0.96627	1.00000/0.96531	**1.0/0.96531**
Zipper	0.99894/0.98747	0.35032/0.67711	**0.97925/0.98486**
Avg. Obj	0.99442/0.9815	0.57884/0.53551	**0.97821/0.98032**
Average	0.99536/ 0.97772	0.71847/0.67945	**0.98472/0.97599**

dropped drastically to just an average I-AUROC of 0.57884 for objects. In contrast, Sam2Sim was strong against dynamic backgrounds and kept an average, very high I-AUROC of 0.98032 for objects. The findings underscore the significant impact of dynamic backgrounds on model performance. Dynamic environments can lead to a substantial degradation in accuracy, potentially resulting in complete model failure.

IV. Conclusion

In conclusion, Sam2Sim uses foreground extraction and robust anomaly identification to ensure greater robustness compared to traditional methods, especially dynamic environments with changing backgrounds. Experiments conducted on the MVTec dataset showed the effectiveness of the model in image-level anomaly detection and localization, hence proving its potential for the reliable and efficient detection of anomalies in an real-world industrial setting.

ACKNOWLEDGMENT

This work was supported by Regional Leading Research Center (RLRC) of the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No.2022R1A5A8026986) and supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-01304, Development of Self-Learnable Mobile Recursive Neural Network Processor Technology). It was also supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Communication Technology Research Center support program (IITP-2024-2020-0-01462) supervised by the IITP (Institute or Information & communications Technology Planning & Evaluation).

REFERENCES

[1] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, ‘CutPaste: Self-Supervised Learning for Anomaly Detection and Localization’, 2021, *arXiv*. doi: 10.48550/ARXIV.2104.04015.

[2] V. Zavrtanik, M. Kristan, and D. Skočaj, ‘DRAEM -- A discriminatively trained reconstruction embedding for surface anomaly detection’, 2021, *arXiv*. doi: 10.48550/ARXIV.2108.07610.

[3] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, ‘SimpleNet: A Simple Network for Image Anomaly Detection and Localization’, 2023, *arXiv*. doi: 10.48550/ARXIV.2303.15140.

[4] J. Yang, Y. Shi, and Z. Qi, ‘DFR: Deep Feature Reconstruction for Unsupervised Anomaly Segmentation’, 2020,

doi: 10.48550/ARXIV.2012.07122.

[5] K. Batzner, L. Heckler, and R. König, ‘EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies’, 2023, doi: 10.48550/ARXIV.2303.14535.

[6] S. Kim *et al.*, ‘Few Shot Part Segmentation Reveals Compositional Logic for Industrial Anomaly Detection’, 2023, *arXiv*. doi: 10.48550/ARXIV.2312.13783.

[7] T. Liu *et al.*, ‘Component-aware anomaly detection framework for adjustable and logical industrial visual inspection’, 2023, *arXiv*. doi: 10.48550/ARXIV.2305.08509.

[8] N. Ravi *et al.*, ‘SAM 2: Segment Anything in Images and Videos’, 2024, *arXiv*. doi: 10.48550/ARXIV.2408.00714