Filter-Former For Anomaly Detection

Participant Information

- Name(s): Chen Jing
- Affiliation(s): BOE Technology Group Co., Ltd., Beijing, china
- Contact Information: chenjing8854@boe.com.cn
- Track: Track I - Adapt & Detect

Abstract

This report presents our approach to robust anomaly detection using the Augmented INP-Former[1] model. Aligning reference images with test images is often complicated by variations in appearance and positioning, which limits the accuracy of detection. Most anomalies manifest as local variations, meaning that valuable normal information remains within anomalous images. We argue that this information is useful and may be more aligned with the anomalies, as both the anomalies and the normal information originate from the same image. Therefore, INP-Former, which directly extracts Intrinsic Normal Prototypes (INPs) from the test image, rather than relying on external normality from the training set. Building on INF-Former, we introduce a branch to extract image features and fuse them with the features extracted as INPs to filter out normal backgrounds and enhance anomalous regions. We evaluated the performance of Filter Former on the test_private and test_private_mixed test datasets of the MVTec-AD 2[7] dataset, achieving average pixel-level F1 scores of 0.4443 and 0.3968.

Introduction

- Background

Visual anomaly detection is crucial in industrial inspection and medical screening[2], where identifying defects and abnormalities can significantly impact quality control and safety. Traditional methods often rely on comparing test images to normal references from a training set, which can be limited by misalignment and intra-class variations. The latest advancements in VAD have been primarily driven by deep learning, specifically through supervised learning [3, 4], unsupervised learning [5, 6, 8], and semi-supervised learning methods [9]. Although powerful, supervised methods typically require large labeled datasets, which are often difficult to obtain in sufficient quantities. As a result, unsupervised learning has gained increasing popularity. These models learn from data representing the system's normal operation, aiming to capture the intrinsic feature distribution of normal samples. Any deviations from this distribution are identified as potential anomalies. However, when confronted with variations in real-world production data that were not present during training, these models face significant challenges in terms of robustness. Factors such as changes in camera specifications, lighting conditions, or gradual wear of mechanical components can cause shifts in data distribution, known as domain shifts. These shifts can severely impact the performance of anomaly detection systems, leading to false positives or missed detections. Therefore, improving VAD algorithms with the capability to handle various changes has become a key challenge in the development of VAD technologies.

- Challenge Description

Challenge require to develop anomaly detection models based on the one-class training paradigm, using only normal images for training and assessing the models' anomaly detection capabilities on validation and test sets that contain a mix of normal and abnormal images. The MVTec-AD 2 dataset includes 8 new challenging scenarios with varying lighting conditions to reflect real-world distribution shifts. The ground truth of the official test set is not publicly available, emphasizing the unsupervised nature of industrial anomaly detection. Model evaluation is based on pixel-level F1 scores (SegF1), which require selecting a single threshold for continuous anomaly maps. The final model ranking is calculated as the average rank of the model's average SegF1 on the private test set (with the same lighting conditions as the training images) and the private mixed test set (with both seen and unseen lighting conditions in the training images) across all 8 object categories of MVTec-AD 2. Submissions are made through the MVTec Benchmark Server, which also serves as the official leaderboard for the MVTec-AD 2 dataset.

Methodology

- Model Design

Approach: In our pursuit of robust anomaly detection, we initiate the model selection process by evaluating the most promising architectures. We assess their respective strengths and weaknesses in relation to robust anomaly detection, which guides us in choosing the initial architecture for further model development. Recently, various strategies aiming to tackle industrial anomaly detection have emerged, We have experimented with methods that have recently been published and have achieved state-of-the-art performance metrics in the industry. We have selected the EfficientAD[10], which employs a classic student-teacher structure, as well as the recently published methods DECO-DIFF[11] and INP-Former[1] from 2025, as our candidate algorithms for experimental validation. We compared the performance of these three methods on some of the categories in the MVTec-AD 2 dataset. The comparison of these three model’s performance is shown in Figure 1.

Our preliminary findings show that INP-Fomer[1] generally outperforms the others, except in the wallplugs and sheet_metal category. The student-teacher approach can be more complex, especially in the design of the auto-encoder to handle a variety of augmentations. DECO-Diff[11] involves a VAE encoder and masked forward diffusion before making predictions. INP-Former also includes a similar internal structure for image encoding and decoding. Based on a comprehensive consideration of the effects, we've decided to use the INP-Former as our base model, to address this challenge. We've also added a feature prediction block to this network and implemented an augmented training strategy. Therefore, we've named our model filter INP-Former (Filter-Former).

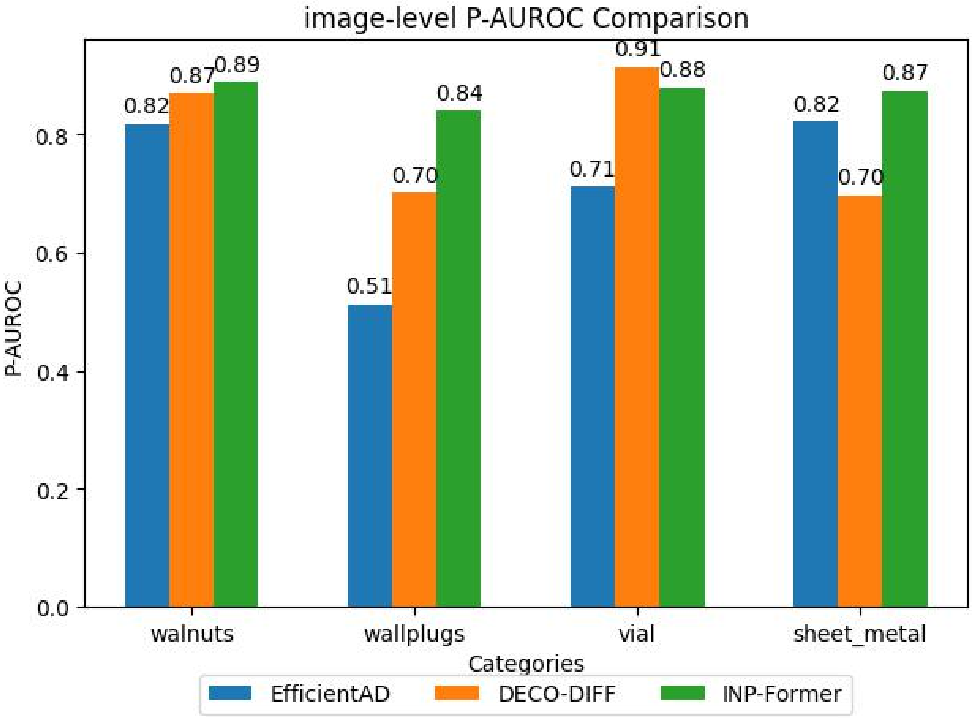

Figure 1 comparison of Models Performance under Augmented MVTec-AD 2 Test set

- Architecture

The architecture of our proposed method, Filter-Former, is illustrated in Figure 2. The blue modules represent the original structure of INP-Former, while the green modules highlight our improvements. The detection process of the model is described as follows:

The input image is first fed into a pre-trained encoder module for feature extraction, which is composed of a Vision Transformer (ViT)[12]. The ViT efficiently extracts image features and generates a list of encoder features (encoder_list) through its multi-layer outputs. In the original INP-Former structure, a set of INP parameters is introduced to learn the distribution within samples. These INP parameters are combined with the encoder_list through attention computation to extract the intrinsic normal prototypes. Subsequently, the encoder_list is processed by a Multi-Layer Perceptron (MLP) and, together with the normal prototypes extracted by INP, is fed into a Decoder to reconstruct the normal image features. During training, the model is updated only on normal samples. Therefore, INPs are unknown in the test abnormal samples, the model has weaker guidance capabilities, which enables the detection of anomalous regions. However, in some datasets, after more training epochs, the features output by the Decoder may become too similar to those output by the Encoder, leading to the potential omission of defects.

To address this issue, we introduce a Diff Predictor module. This module performs self-attention feature extraction on the output of the Encoder and predicts the differences in the original structure. The primary objective of the Diff Predictor is to capture the discrepancies between training and test images, which may indicate anomalous regions. By comparing the Encoder output and the Diff Predictor output, we can calculate the feature differences, which are assessed by a Feature Comparison module to determine whether these differences correspond to anomalous regions.

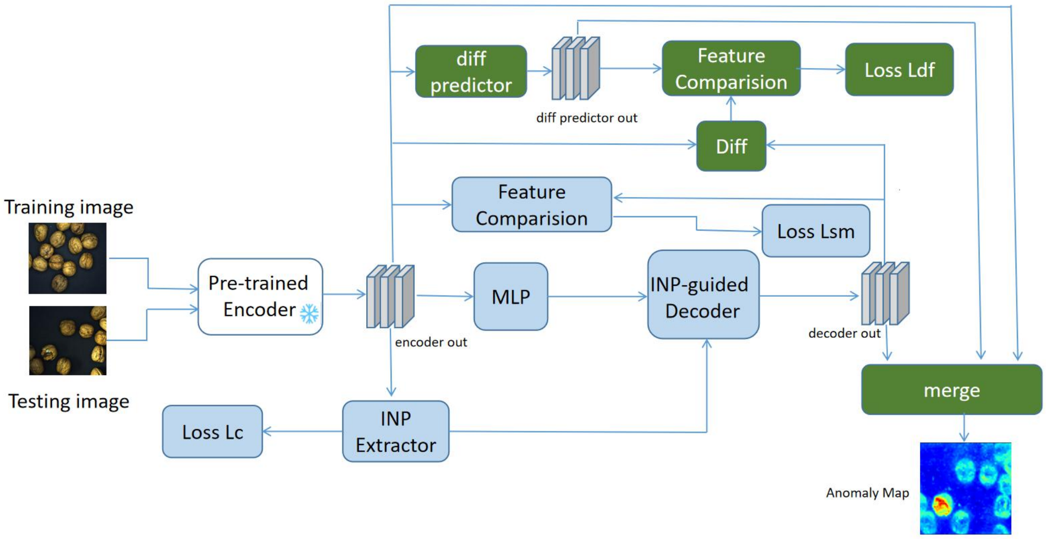

Figure 2 Architecture of Filter-Former：INP-Former(blue blocks) and Our Enhanced Modules(green blocks)

To further optimize the model's performance, we reserve two loss functions of INP-Former: Loss Lc and loss Lsm (Soft Mining Loss). Lsm is designed to handle samples that are difficult to optimize, i.e., abnormal pixels that the model struggles to distinguish during training. By assigning higher weights to these challenging samples, the model can focus more on these difficult regions, thereby enhancing overall detection performance.

We introduce the loss Ldf to measure the discrepancy between the encoder features and decoder features of normal samples. This guides the model to more accurately filter out normal background, thereby enhancing the detection of anomalous regions.

During the prediction phase, we process the Encoder output through the Diff Predictor module and then compute the cosine similarity between the processed output and the Decoder output to generate the final anomaly map.

- Training

In our research, we have designed the training process to ensure that the model can effectively learn the features of normal samples and detect anomalies. The following details the training procedure of our model, including the specifics of one-class training, as well as preprocessing steps and data augmentation techniques.

Preprocessing and Data Augmentation

Initially, we performed preprocessing on the input images to meet the model's input requirements. Specifically, we resized the images to 896x896 pixels, a step that ensures all input images have a uniform size, thereby simplifying the computations during the model training process.Regarding data augmentation, we employed a variety of techniques to increase the diversity of the training data and enhance the model's generalization capabilities. These techniques included random cropping, color jittering, rotation, flipping, and directional lighting adjustments. Through these augmentation methods, we were able to simulate different image capture conditions and viewpoint changes encountered in real environments, thus making the model more robust.

One-Class Training

We proceed with one-class training, which is a crucial step in unsupervised anomaly detection. In one-class training, Each category is trained on a single NVIDIA A100 GPU. The learning rate decays from 0.001 to 0.0001 using a cosine annealing schedule and using the AdamW optimizer. the model is trained by only normal samples. For this, we selected 'dinov2reg_vit_large_14', a large-scale model based on Vision Transformer, capable of providing rich feature representations. We set the total number of training epochs to 500. We also set the batch size to 4, a decision made after considering the trade-off between memory usage and computational efficiency. Additionally, we aligned the dimensionality of the INP module with the number of heads in 'dinov2reg_vit_large_14', a choice made after balancing model performance and computational complexity.

Model Saving

We trained the model and saved the best-performing models under different fusion modes on the test public dataset, as well as the final model. The corresponding merge methods for each class were also saved to the storage path in JSON files. our training process included key steps such as image preprocessing, data augmentation, one-class training, and model saving. Through these steps, our model was able to effectively learn the features of normal samples and accurately detect anomaly regions during testing.

Dataset & Evaluation

- Dataset Utilization

We utilizate the MVTec-AD 2 Dataset for training and evaluation. This dataset cover a wide range of industrial objects and textures, providing a comprehensive testbed for our model. Data augmentation techniques, including random cropping, color jittering, rotation, flipping, and directional lighting adjustments, which were applied to simulate real-world variations.

- Evaluation Criteria

For the competition evaluation requirements, we employ pixel-level F1 scores (SegF1) as our evaluation metric. This approach ensures a balanced consideration of precision and recall in the models' anomaly detection performance. Moreover, it necessitates the selection of a single threshold for the typically continuous anomaly maps—a challenge that is often overlooked in the scientific community but is essential for deployment in real-world applications.

Results

- Performance Metrics

The performance of our model and the original INP-Former on the MVTec AD 2 test dataset is presented in table 1 and table 2, respectively. Both models utilize the same values for the shared hyperparameters.

Category	Pixel-F1 Max	
	test_private	test_private_mixed
Can	7.84	0.78
Fabric	49.05	26.23
Fruit Jelly	53.19	53.65
Rice	60.32	59.85
Sheet Metal	50.09	48.57
Vial	39.08	36.68
Wallplugs	9.23	1.12
Walnuts	67.18	55.49
Overall Mean	41.32	35.3

Table 1 Performance evaluation of INP-Former on MVTec-AD 2 Test set

Category	Pixel-F1 Max	
	test_private	test_private_mixed
Can	8.65	1.99
Fabric	50.44	38.7
Fruit Jelly	54.6	54.17
Rice	61.19	59.79
Sheet Metal	54.38	50.6
Vial	46.74	43.13
Wallplugs	11.08	4.47
Walnuts	68.38	64.61
Overall Mean	44.43	39.68

Table 2 Performance evaluation of Filter Former on MVTec-AD 2 Test set

- Comparison

Compared with the INP-Former methods, our approach achieves a 3.11-point improvement in pixel-level F1 max on the private test set and a 4.38-point improvement in pixel-level F1 max on the private mixed test set of the MVTec-AD 2 dataset.

Discussion

- Challenges & Solutions

One of the main challenges was ensuring the robustness of the model to real-world variations. We addressed this by incorporating extensive data augmentations during training and using the INP Coherence Loss to maintain the integrity of the extracted INPs. Another challenge was handling logical anomalies that closely resemble the background. We plan to address this in future work by combining INPs with pre-stored prototypes.

- Model Robustness & Adaptability

One of the main challenges is to ensure the robustness of the model against real-world variations. We addressed this issue by incorporating extensive data augmentation during training and using the INP consistency loss to maintain the integrity of the extracted INPs. Another challenge is dealing with logical anomalies that are very similar to the background, which are currently relatively weak in detection capabilities.

- Future Work

In future work, we plan to draw on the form of methods similar to PatchCore[] and integrate them with INP to enhance the model's ability to handle logical anomalies. Moreover, we aim to optimize the model to address the issue of decreasing accuracy for certain categories as the number of training steps increases.

Conclusion

In summary, we introduce Filter Former, a robust anomaly detection model that employs image augmentation and is trained in a fully unsupervised manner. It demonstrates strong pixel-level detection performance on the MVTec AD 2 dataset. The model's ability to directly extract INPs from test images enhances detection accuracy and robustness, positioning it as a promising solution for real-world applications. Our meticulously designed training data augmentation achieves pixel-level F1 Max performance of 0.4443 on the test private set and 0.3968 on the test private mixed set, highlighting the potential robustness of our model in addressing real-world anomaly detection challenges. For future work, we plan to further investigate the category-specific limitations of our model and make necessary adjustments to our detection framework.

References

[1] Luo W, Cao Y, Yao H, et al. Exploring Intrinsic Normal Prototypes within a Single Image for Universal Anomaly Detection[J]. arXiv preprint arXiv:2503.02424, 2025.

[2]Jiang X, Li J, Deng H, et al. Mmad: The first-ever comprehensive benchmark for multimodal large language models in industrial anomaly detection[J]. arXiv preprint arXiv:2410.09453, 2024.

[3] Tian Y, Ye Q, Doermann D. Yolov12: Attention-centric real-time object detectors[J]. arXiv preprint arXiv:2502.12524, 2025.

[4] Yuan L, Chen Y, Wang T, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 558-567.

[5] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9650-9660.

[6] Roth K, Pemula L, Zepeda J, et al. Towards total recall in industrial anomaly detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 14318-14328.

[7] Heckler-Kram L, Neudeck J H, Scheler U, et al. The MVTec AD 2 Dataset: Advanced Scenarios for Unsupervised Anomaly Detection[J]. arXiv preprint arXiv:2503.21622, 2025.

[8] Jeong J, Zou Y, Kim T, et al. Winclip: Zero-/few-shot anomaly classification and segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 19606-19616.

[9] Villa-Pérez, Miryam Elizabeth, et al. "Semi-supervised anomaly detection algorithms: A comparative summary and future research directions." Knowledge-Based Systems 218 (2021): 106878.

[10] Batzner K, Heckler L, König R. Efficientad: Accurate visual anomaly detection at millisecond-level latencies[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 128-138.

[11] Beizaee F, Lodygensky G A, Desrosiers C, et al. Correcting Deviations from Normality: A Reformulated Diffusion Model for Multi-Class Unsupervised Anomaly Detection[J]. arXiv preprint arXiv:2503.19357, 2025.

[12] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.