MIAD: A Maintenance Inspection Dataset for Unsupervised Anomaly Detection

Tianpeng Bao[1] Jiadong Chen[1] Wei Li[1] Xiang Wang[1] Jingjing Fei[1] Liwei Wu[1]
Rui Zhao[1,3] Ye Zheng[2*]
SenseTime Research[1] JD.com, Inc.[2]
Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China[3]
{baotianpeng, liwei1, wangxiang, feijingjing1, wuliwei, zhaorui}@sensetime.com
chenjiadong1@senseauto.com zhengye12@jd.com

Abstract

Visual anomaly detection plays a crucial role in not only manufacturing inspection to find defects of products during manufacturing processes, but also maintenance inspection to keep equipment in optimum working condition particularly outdoors. Due to the scarcity of the defective samples, unsupervised anomaly detection has attracted great attention in recent years. However, existing datasets for unsupervised anomaly detection are biased towards manufacturing inspection, not considering maintenance inspection which is usually conducted under outdoor uncontrolled environment such as varying camera viewpoints, messy background and degradation of object surface after long-term working. We focus on outdoor maintenance inspection and contribute a comprehensive Maintenance Inspection Anomaly Detection (MIAD) dataset which contains more than 100K high-resolution color images in various outdoor industrial scenarios. This dataset is generated by a 3D graphics software and covers both surface and logical anomalies with pixel-precise ground truth. Extensive evaluations of representative algorithms for unsupervised anomaly detection are conducted, and we expect MIAD and corresponding experimental results can inspire research community in outdoor unsupervised anomaly detection tasks. Worthwhile and related future work can be spawned from our new dataset.

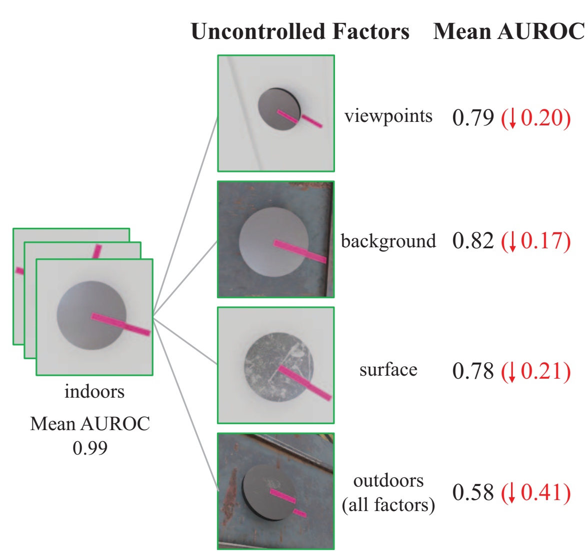

Figure 1. Images for maintenance inspection are usually captured outdoors and suffer from varying camera viewpoints, messy background and degradation of object surface after working for a long time, which makes the distribution of non-defective samples more complicated. Some state-of-the-art methods can achieve AUROC of 0.99 on the toy dataset for witness marks with indoor setting. But when we vary each outdoor uncontrolled factor, we observe obvious drops in performance. We show the mean AUROC across state-of-the-art methods on MVTec AD including PatchCore [25], Reverse Distillation [11], FastFlow [38] and DRAEM [39]. The detailed metrics are shown in Table 6.

1. Introduction

Anomaly detection plays a crucial role in not only manufacturing inspection but also maintenance inspection. Manufacturing inspection is intended to find defect of products during manufacturing process in many industrial fields, such as electronics [13], metals [31], fabrics [28] and food [5]. Maintenance inspection is intended to find whether equipment or product is in optimum working condition after leaving the factory particularly outdoors, such as inspection of power transmission lines [29], photovoltaic power station [21], wind power plant [34] and overhead catenary system [6]. Among various monitoring approaches including ultrasonic and X-ray, optical inspection is the most basic, low-cost and frequently used type and is considered non-contact and non-destructive [13, 34].

Due to the scarcity of defective samples and abundance of non-defective samples, research community as well as industrial community has paid increased attention to unsuper-

*Corresponding author.

vised setting of anomaly detection, i.e. algorithms should be trained solely on non-defective images and be tested on both non-defective and diverse defective images.

The problem of unsupervised anomaly detection for manufacturing inspection [8] has been thorough researched, which benefits from plenty of emerging datasets especially MVTec AD [2]. MVTec AD is the first comprehensive dataset for unsupervised anomaly detection in industrial inspection which provides pixel-accurate ground truth regions and allows to evaluate algorithms at both image level and pixel level. Since MVTec AD is collected indoors under controlled environments, images of the same category are roughly aligned at pixel level with a clean background. Based on this prior knowledge, PatchCore [25] achieves state-of-the-art performance with image-level AUROC of more than 0.99. However, when we try to apply PatchCore and other representative methods to maintenance inspection in the outdoors, we observe significant drops in performance. Besides, there are only hundreds of images in the training dataset per category, which is enough for indoor scenarios but is not sufficient for outdoor scenarios due to the complex distribution of non-defective samples. In a word, existing datasets including MVTec AD are not suitable for research on outdoor maintenance inspection due to **pixel-level alignment** and **limited training dataset size**.

The imaging equipment for maintenance inspection are usually carried outdoors by unmanned aerial vehicles (UAV) or track inspection vehicles [6] under uncontrolled environment. As shown in Figure 1, images for outdoor maintenance inspection suffer from varying camera viewpoints, messy background and the degradation of object surface (e.g. dust or rust), and it is reasonable to doubt whether these uncontrolled factors break the prior assumption about pixel-level alignment. In order to verify this doubt, we synthesize a toy dataset of witness mark by control variable technology. Experiments on this toy dataset show that uncontrolled viewpoints, background, and surface individually have a negative effect on the accuracy of the state-of-the-art algorithms, and the combination of these uncontrolled factors results in a more challenging task.

Therefore, we build and release a maintenance inspection dataset, named MIAD*, for unsupervised anomaly detection. Considering the accessibility of various industrial scenarios and high-cost labeling of pixel-precise ground truth, we make use of BlenderProc [12], a 3D graphics software, to build the 3D scenes and automatically generate 2D color images with pixel-precise ground truth. The dataset mimics seven outdoor scenarios including photovoltaic module, wind turbine blade, nut and bolt, which is inspired by real-world maintenance inspection. MIAD focuses on the three dominant factors mentioned above, ignoring uncontrolled light, weather, and other factors that may affect imaging quality. Moreover, the area under the receiver operating characteristic curve (AUROC) is utilized as a image-level metric by existing datastes, which treats the performance at the high false positive rate (FPR) and the low FPR equally. We adopt an additional metric to focus on the performance at the low false alarm rate which is pursued by industrial community.

In summary, we make the following contributions:

- We present MIAD, a novel maintenance inspection dataset for unsupervised anomaly detection in uncontrolled environments (varying camera viewpoints, messy background, and object surface degradation), which breaks the pixel-level alignment assumption used in existing manufacturing inspection datasets. It consists of more than 100K high-resolution color images with pixel-level annotation, an order of magnitude larger than existing datasets.
- We perform comprehensive evaluations on MIAD with representative unsupervised anomaly detection methods. Experiments show that all of the evaluated methods perform poorly (mean AUROC 0.730-0.850) when compared to their performance on MVTec AD (mean AUROC 0.980-0.996), indicating that the accuracy of the existing method is still far from practical application, and there is still a long way to go in applying unsupervised anomaly detection to outdoor maintenance.

2. Related Work

The progress of research in industrial anomaly detection has been strongly driven by the development of datasets. Here, we briefly review related datasets and methods for unsupervised anomaly detection.

2.1. Datasets for Unsupervised Anomaly Detection

2.1.1 Datasets in Manufacturing Inspection

AITEX [28] is a public dataset for textile inspection, which consists of 245 images of different fabric textures obtained from a real production plant. **BTAD** [22] contains a total of 2830 real-world images of 3 industrial products showcasing body and surface defects. **DAGM** [35] is a synthetic dataset for defect detection on textured surfaces. It consists of multiple categories generated by a different texture model and defect model, each consisting of 1000 normal samples and abnormal images with annotated defect. **ELPV** [10] is collected based on electroluminescence (EL) imaging of polycrystalline photovoltaic (PV) modules. It contains 2624 8-bit grayscale images of functional and defective solar cells. All images are the 300x300 pixels with varying degree of degradations extracted from 44 different solar modules. **KolektorSDD** [30] is proposed for surface-defect detection for an industrial semi-finished product where the num-

*https://miad-2022.github.io/

Datasets	#Category	#Train	#Test(good+defective)	Scene & Env	Real/Synthetic
AITEX	1	0	140+105	Manufacturing & Indoor	Real
BTAD	3	1799	1031(total)	Manufacturing &Indoor	Real
DAGM	10	6900	4600(total)	Manufacturing &Indoor	Synthetic
ELPV	2	1968	656(total)	Manufacturing & Indoor	Real
KolektorSDD	1	133	266(total)	Manufacturing & Indoor	Real
MTD	2	448	896(total)	Manufacturing & Indoor	Real
MVTec AD	15	3629	467+1258	Manufacturing & Indoor	Real
MVTec 3D AD	10	2656	248+948	Manufacturing & Indoor	Real
MVTec LOCO AD	5	1772	575+993	Manufacturing & Indoor	Real
CPLID	1	1186	770(total)	Maintenance & Outdoor	Real & Synthetic
MIAD	7	**70000**	**17500+17500**	Maintenance & Outdoor	Synthetic

Table 1. Comparison of datasets for unsupervised anomaly detection. Our MIAD dataset focus on outdoor maintenance inspection and is an order of magnitude larger than existing datasets.

ber of defective items available for the training is limited. **MTD** [16] contains 1344 images, with the cropped ROIs of 6 kind of magnetic tile and pixel-level labels. To simulate the manufacturing process in real assembly line, this dataset collects images under multiple illumination conditions for each given magnetic tile. **MVTec AD** [2] is a widely used unsupervised anomaly detection dataset, which contains 5354 high-resolution normal and anomalous images of 15 different real-world products. Most of recent unsupervised anomaly detection methods are driven from this dataset. **MVTec 3D AD** [3] was proposed recently to encourage research into 3D anomaly detection and segmentation. It contains over 4000 high-resolution 3D scans of industrially manufactured products across 10 categories. Each sample is represented by an organized point cloud and a corresponding RGB image with a one-to-one mapping between the pixels in the point cloud and those in the RGB image. **MVTec LOCO AD** (MVTec Logical Constraints Anomaly Detection) [1] includes both structural and logical anomalies with 3644 images from five different categories inspired by real-world industrial inspection scenarios.

The details of aforementioned datasets are summarized in Table 1. Compared with our MIAD dataset, these datasets are all collected indoors and does not take into consideration uncontrolled environment such as random camera viewpoints, messy background, degradation of object surface (rust or dust), etc. In addition, their datasets have small scales, so it may be unable to fully exploit the potential of unsupervised anomaly detection algorithms and to adequately validate them.

2.1.2 Datasets in Maintenance Inspection

Most of the above datasets are collected from indoor scenarios under controlled environments. We also survey some outdoor datasets which are rarely mentioned by the unsupervised anomaly detection community.

Xiaoxia Li et al. [21] collected two typical PV module visible defects: snail trail and dust shading by the UAV inspection system. The railway catenary dataset used in [6] consists of the catenary support device images captured from an approximately 100-km line along a high-speed railway, in which 2000 catenary support devices and 40000 fasteners exist. **CPLID** [32] is proposed to detect the defect of grid insulator, which contains 1956 high resolution images. It is captured by an UAV with a DJI M200 camera at a resolution of 4608 × 3456 pixels and stored in BMP format.

However, the PV module [21] and the railway catenary[6] datasets for outdoor inspection are not publicly available, and thus they cannot be used to verify any proposed results or make comparisons between different methods. In CPLID [32], the majority of abnormal images are produced as a result of data augmentation, such as adding noise to normal images, which is insufficient for representing outlier samples. Furthermore, these outdoor datasets are limited to a single industry, resulting in a lack of diversity. In this paper, we present an outdoor maintenance inspection dataset that includes more than 100K annotated images and covers a variety of outdoor industrial maintenance inspection scenarios to support relevant research more effectively.

2.2. Methods for Unsupervised Anomaly Detection

Following the survey [8], we classify the unsupervised anomaly detection algorithms into four classes: reconstruction-based, representation-based, normalizing flow-based and data augmentation-based methods. We will select one or two representative methods from each class to make an analysis on our MIAD dataset in Section 4.

2.2.1 Reconstruction-based Methods

Reconstruction-based methods are designed to score anomaly by reconstruction error. A neural network is trained only on the normal images to generate high-fidelity input reconstructions and, it is not able to reconstruct abnormal images correctly in inference. The greater probability of anomaly there is, the higher anomaly scores of a pixel will be. Typical algorithms mainly include auto-encoders [18, 27, 4], variational auto-encoders [17] and generative adversarial networks [14]. Recently the teacher-student framework [33, 36, 11] is adopted for unsupervised anomaly detection. In the training phase, only the student network is trained to reconstruct the multi-layer feature of the teacher network whose frozen parameters are pre-trained on the ImageNet. In the testing phase, both teacher and student networks are fed with a sample and the corresponding anomaly map is computed based on the difference of the two multi-layer features. Especially, the student network of Reverse Distillation [11] takes teacher model's one-class embedding rather than raw images as input to reverse data flow in knowledge distillation, which achieves promising results on the MVTec AD benchmark due to heterogeneity of the teacher and student networks. UniAD [37] accomplishes anomaly detection for multiple classes with a unified framework by designing improvements base on the transformers.

2.2.2 Representation-based Methods

Representation-base methods extract discriminative features from normal images and further build the normal distribution. Anomaly score results are obtained by measuring the distance between the test images with the distribution of normal images. Typical methods include SPADE [7], PaDiM [9], FYD [40], PatchCore [25], etc. However, these methods are CPU/GPU memory-consuming when there are thousands of training samples. Their training time complexities scale linearly with the dataset size, and require large memory allocation for gallery features.

CFA [19] consists of a learnable patch descriptor to learn embed representations and a scalable memory bank independent of the size of the dataset. It adopts transfer learning to increase the normal feature density so that abnormal features can be clearly distinguished by applying patch descriptor and memory bank to a pre-trained CNN.

2.2.3 Normalizing Flow-based Methods

Normalizing Flows [24] are intended to learn transformations between data distributions and well-defined standard normal distribution. Recently, some work began to use it for unsupervised anomaly detection and localization. In these methods, normal image features are embedded into standard normal distribution and the probability is used to identify and locate anomalies. DifferNet [26] achieved good image level anomaly detection performance by using it to estimate the precise likelihood of test images. Unfortunately, this work failed to obtain the exact anomaly localization results since they flattened the outputs of feature extractor. CFLOW-AD [15] proposes to use hard code position embedding to leverage the distribution learned by Normalizing Flows. Furthermore, FastFlow [38] implements efficient 2D normalizing flows and use it as the anomaly estimator, which can learns to transform the visual input as a tractable distribution at train stage and obtain the probability to identify anomalies at inference stage.

2.2.4 Data Augmentation-based Methods

The core idea of data augmentation-based methods is to simulate the anomalies through augmentation, which is data-dependent and hand-crafted. DRAEM [39] learns a joint representation of an anomalous image and its anomaly-free reconstruction, while simultaneously learning a decision boundary between normal and anomalous examples. CutPaste [20] uses a simple data augmentation strategy that cuts an image patch and pastes at a random location of a large image to train the feature extraction model.

3. Benchmark

In this section, we will first describe the details of our MIAD dataset. And then we will define the metrics for anomaly classification and segmentation performance.

3.1. Dataset Description

The MIAD dataset consists of seven outdoor maintenance inspection scenarios. All scenarios are inspired by real-world maintenance inspection including metal welding for oil transmission pipeline, photovoltaic modules in photovoltaic power station [21], wind turbine blades in wind power plant [34], the catenary dropper in overhead catenary system [6]. The electrical insulator, nut and bolt, and witness mark are widely used in many industrial fields such as power transmission lines [29] and catenary support devices [6]. Note that outdoor maintenance inspection is usually conducted during the day, but in some special cases it has to be conducted during the night. For example, the track inspection vehicles for catenary support device only work at night since the railway is occupied by transporting passengers during the day. The example images of each scenario are illustrated in Figure 2.

There are a total of 105000 color images with 512×512 pixel high-resolution. The training set comprises 70000 images without any defect, 10000 of each scenario. The test set comprises 35000 images, 2500 non-defective and 2500 defective images of each scenario. Further information about each category is summarized in Table 2.

Different from the existing datasets for manufacturing inspection which mainly collected in controlled environment, the MIAD dataset focus on the impact of uncon-

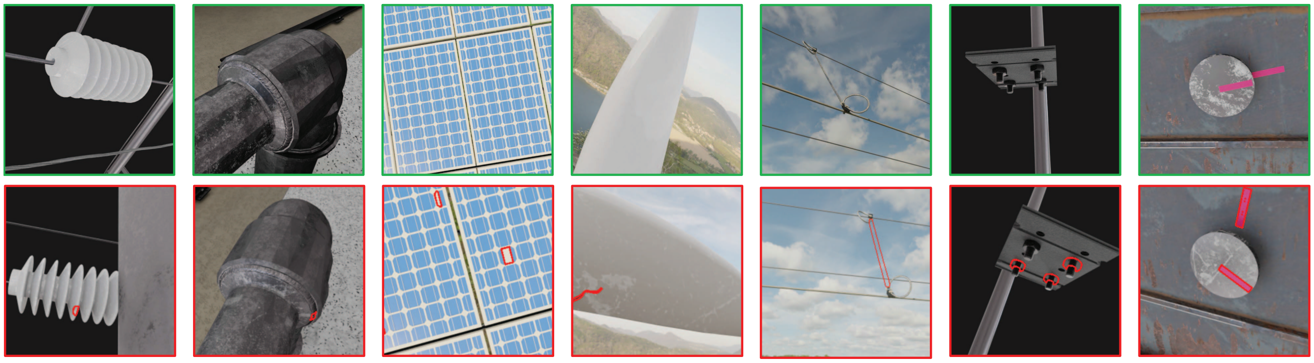

Figure 2. Example images of the MIAD dataset for seven maintenance inspection scenarios. The test set for each scenario comprises non-defective (top row) and defective images (bottom row). The first four scenarios contain surface anomalies, and the rest contain logical anomalies. Pixel-precise annotations are provided for all anomalies.

Anomaly	Scenarios	Uncontrolled Factors	Day/Night	#Defect Types	#Defect Pixel Ratio
Surface	Electrical Insulator	UV, UB, US	Night	1	0.04%
	Metal Welding	UV, US	Day	2	0.10%
	Photovoltaic Module	UV	Day	3	0.11%
	Wind Turbine	UV, UB, US	Day	1	0.05%
Logical	Catenary Dropper	UV, UB	Day	3	0.37%
	Nut and Bolt	UV, UB	Night	3	0.36%
	Witness Mark	UV, UB, US	Day	1	0.70%

Table 2. Overview of the MIAD dataset. In the third column, UV, UB and US stand for Uncontrolled Viewpoints, Uncontrolled Background and Uncontrolled Surface, respectively. The defect pixel ratio equals the number of defective pixels divided by the number of all pixels.

trolled environment for maintenance inspection. Images acquisition in maintenance inspection is usually conducted by a UAV system, which results in varying camera viewpoints (uncontrolled viewpoints). Messy background (uncontrolled background) is also inevitable when imaging in an open environment. Moreover, degradation of object surface (uncontrolled surface) after long-term outdoor working makes the distribution of non-defective objects more complex. For convenience, we utilize UV, UB, US to denote uncontrolled viewpoints, uncontrolled background and uncontrolled surface in this paper, respectively.

We utilize BlenderProc[12], a procedural Blender pipeline for photorealistic rendering, to generate our MIAD dataset. A basic pipeline for rendering with BlenderProc is shown as the top flow in Fig 3. After some basic initialization of the blender project (e.g. configures computing device, creates a camera), a 3D scene is loaded and a camera is set with a predefined position and pose inside this scene. Then the RGB and semantic segmentation images can be rendered simultaneously based on the physical based rendering [23] technology. In order to simulate the uncontrolled surface, background and viewpoints in maintenance inspection, three modules are implemented. The bottom flow in Fig 3 depicts the modified pipeline and the details will be describes in the following three subsections.

3.1.1 Uncontrolled Surface

In the outdoors, the surface of objects may deteriorate (such as dust or rust) due to bad weather or other environment, resulting in random irregular textures. To simulate the uncontrolled surface, we randomly change materials of the object of interest by a mix shader which can mix the original shader with another one like dust. The original and additional materials can been found in blenderco textures library[†]. Plenty of similar texture files are downloaded and they are randomly split into training set and test set to prevent the data leakage.

3.1.2 Uncontrolled Background

To simulate the uncontrolled background during the day, e.g. the wind turbine scenario in our MIAD dataset, we select reasonable images from the Internet[‡] for each scenario and then randomly change the background image in BlenderProc. As for uncontrolled background during the night, e.g. the nut and bolt scenario in our MIAD dataset, reasonable background images are almost the same (pure black), we randomly add background objects to simulate the messy background.

[†] https://blenderco.cn/category/tiet
[‡] https://polyhaven.com

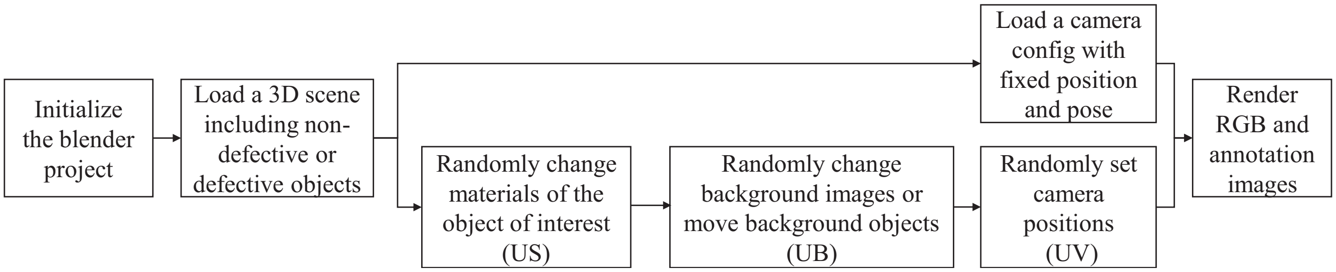

Figure 3. The basic pipeline (top flow) and a modified pipeline (bottom flow) for rendering photorealistic data by BlenderProc.

3.1.3 Uncontrolled Viewpoints

The images varies with the uncontrolled camera viewpoints which has 6 degree of freedom. In order to capture the object of interest, the 3D pose is constrained based on that the camera always look at a point of interest on the object, while the 3D position can been randomly set. For convenient, we sample the camera position in spherical coordinate rather than Cartesian coordinate. In other word, the camera position can be denoted as (r, θ, ϕ), where r is the radius, θ is the azimuth angle and ϕ is the elevation angle. r affects the size of objects, while θ and ϕ jointly affects the visible region of objects. In accordance with the size of the objects and the size of defect regions, different scenarios necessitate different sampling ranges of r which is shared by training and test set to make the dataset follow similar distribution. To prevent the data leakage, the θ and ϕ for training images satisfy

$$\theta, \phi \in \cup_{n=0}^{179} [2n, 2n+1), \tag{1}$$

and the θ and ϕ for test images satisfy

$$\theta, \phi \in \cup_{n=0}^{179} [2n+1, 2n+2), \tag{2}$$

where θ and ϕ are both in degrees.

3.2. Evaluation Metrics

All evaluated algorithms should provide a one-channel anomaly map, in which large values indicate that a certain pixel belongs to an anomalous region. And the maximum score over all pixels in a given anomaly map is regarded as the image-level anomaly score.

To assess the image-level anomaly classification performance, we adopt AUROC as an evaluation metric following MVTec AD [2]. Although AUROC is a independent of thresholds, a threshold must be determined to make a binary decision when algorithms are applied in industrial scenarios. Moreover, AUROC treats the performance at the high FPR and the low FPR equally, but industrial community pursues better performance only at the low FPR because lower false alarm rate brings lower cost for manually checking the report from algorithms. FPR is usually required no larger than 1%, so we propose to use recall at 1% FPR as an additional image-level metric to measure whether a method is applicable in industry.

In order to assess the anomaly segmentation capability, a naive way is to calculate the pixel-level AUROC by regarding the classification result of each pixel as a sample. However, the defect pixel ratio of all scenarios in Table 2 is less than 0.01 and this means pixel-level AUROC can easily surpass 0.99 if all pixels are predicted as non-defective, which indicates that pixel-level AUROC is not a proper metric to reflect different performance of evaluated methods, especially in outdoor inspection. Following MVTec AD [2], a normalized per-region overlap (PRO) between segmentation and ground truth is calculated and the area under the PRO curve (AUPRO) is adopted as a pixel-level metric.

4. Experiments

In this section, we conduct a thorough evaluation of multiple state-of-the-art methods for unsupervised anomaly detection on our dataset to serve as a baseline for future methods. The strengths and weaknesses of each method are discussed on the various surface and logical anomalies. Moreover, we analysis the impact of three uncontrolled factors of outdoors and the impact of training dataset size in a quantitative way.

4.1. Evaluated Methods

We select Reverse Distillation [11], PatchCore [25], FastFlow [38] and DRAEM [39] as representative of reconstruction-, representation-, normalizing flow- and data augmentation-based methods, respectively. The naive method L2 Auto-Encoder as described by [4] is also adopted as a baseline. Since the above methods train separate models for different scenarios, we also adopt UniAD [37] which shares the same parameters for multiple scenarios. Detailed information of each algorithm including the input size, data augmentation and neural network can be found in Appendix.

4.2. Overall Results

Evaluation results for the classification of anomalous images are given per methods and dataset categories in Tables

Scenarios	RD	PatchCore	FastFlow	DRAEM	AE	UniAD
Electrical Insulator	0.68 0.01	0.55 0.01	0.54 0.01	**0.81** **0.24**	0.49 0.01	0.50 0.01
Metal Welding	0.91 0.01	**0.95** 0.22	0.78 0.02	**0.95** **0.37**	0.47 0.01	0.57 0.01
Photovoltaic Module	0.85 0.11	0.53 0.01	0.97 0.74	**1.00** **0.85**	0.68 0.03	0.64 0.05
Wind Turbine	0.91 0.04	0.80 0.06	**0.97** **0.67**	0.88 0.45	0.48 0.01	0.85 0.12
Catenary Dropper	**0.96** 0.22	0.93 **0.60**	0.95 **0.60**	0.95 0.48	0.61 0.05	0.81 0.07
Nut and Bolt	**0.89** **0.20**	0.67 0.05	0.55 0.01	0.84 0.02	0.50 0.01	0.52 0.01
Witness Mark	0.57 **0.03**	**0.71** 0.02	0.51 0.01	0.52 0.01	0.49 0.01	0.59 0.01
Mean	0.82 0.09	0.73 0.14	0.75 0.29	**0.85** **0.35**	0.53 0.02	0.64 0.04

Table 3. Results of the evaluated methods when applied to the classification of anomalous images. For each method, the AUROC (top row) and Recall@1%FPR (bottom row) are given. RD stands for Reverse Distillation and AE stands for Auto-Encoder.

Scenarios	RD	PatchCore	FastFlow	DRAEM	AE	UniAD
Electrical Insulator	0.90	0.64	0.74	**0.98**	0.51	0.64
Metal Welding	0.68	0.47	**0.87**	0.64	0.60	0.76
Photovoltaic Module	0.92	0.52	0.59	**0.98**	0.85	0.69
Wind Turbine	0.95	0.73	**0.96**	0.92	0.45	0.89
Catenary Dropper	0.76	0.86	**0.87**	0.53	0.59	**0.87**
Nut and Bolt	**0.94**	0.89	0.66	0.72	0.48	0.74
Witness Mark	0.10	0.65	0.45	0.48	0.56	**0.79**
Mean	0.75	0.68	0.73	0.75	0.58	**0.77**

Table 4. Results of the evaluated methods when applied to the segmentation of anomalous regions. For each method, the AUPRO is given. RD stands for Reverse Distillation and AE stands for Auto-Encoder.

3. From the view of the metric AUROC, all of the evaluated methods perform poorly with highest mean AUROC 0.85 when compared to their performance on MVTec AD (mean AUROC 0.980-0.996), proving that outdoor inspection is more challenging than indoor inspection for all existing algorithms due to uncontrolled environment. A higher AUROC is not always means a higher recall at 1% FPR. For example Reverse Distillation achieves highest AUROC with 0.96 on the catenary dropper subset but obtains Recall@1%FPR with 0.22 which is near three times smaller than than the highest score 0.60. From the view of the metric Recall@1%FPR, existing methods achieve highest recall 0.35, indicating that existing methods are far from practical application for outdoor inspection.

Factors	RD	PatchCore	FastFlow	DRAEM	Mean
baseline	0.85/0.53	0.80/0.40	0.83/0.44	0.90/0.44	0.85/0.45
+UV	0.89/0.37	0.64/0.02	0.70/0.18	0.78/0.23	0.75/0.20
+UB	0.73/0.17	0.71/0.29	0.71/0.41	0.81/0.19	0.74/0.27
+US	0.81/0.05	0.76/0.07	0.79/0.16	0.89/0.21	0.81/0.12
+UV+US+UB	0.68/0.01	0.55/0.01	0.54/0.01	0.81/0.24	0.65/0.07

Table 5. Impact of uncontrolled factors under outdoor environment on the electrical insulator. For each method, the image-level AUROC (the first number) and Recall@1%FPR (the second number) are given. RD stands for Reverse Distillation and AE stands for Auto-Encoder.

From Tables 3, we can also observe that none of the methods manages to consistently perform well across all surface and logical anomaly. DRAEM is more preferable for surface anomaly while Reverse Distillation is more preferable for logical anomaly.

Evaluation results for the segmentation of anomalous regions are given for all methods and dataset categories in Tables 4. We can observe that a high AUPRO does not necessarily coincide with a high AUROC or a high Recall@1%FPR.

Qualitative results of each method are shown in Figure 4, majority of results are failure cases due to uncontrolled viewpoints, background and surface.

Since the Auto-Encoder and UniAD performs poorly on MIAD, we will conduct further experiments only with the other four methods.

4.3. Impact of Outdoor Uncontrolled Factors

In this section, we will study the impact of the three outdoor uncontrolled factors. As described in Table 2, not all scenarios cover these three factors, so we select the electrical insulator and the witness mark as the representative of surface anomaly and logical anomaly, respectively. We generate two baseline datasets under controlled environment, and individually add UV, UB, US to the baseline. At the end, we add these three factors simultaneously, which is under the setting with our MIAD. The results are shown in Table 5 and 6. UV, UB and US will bring significant drops in performance separately to most methods except PatchCore on the witness mark. The three factors seems to have equal impact from the view of AUROC. However, from the view of Recall@1%FPR, US plays a dominant role for the surface anomaly. The combination of these factors results in greater drops without exception.

4.4. Impact of Training Dataset Size

According to Table 1, the training dataset size per category of existing datasests for unsupervised anomaly detection is not greater than 1200. The limited training dataset

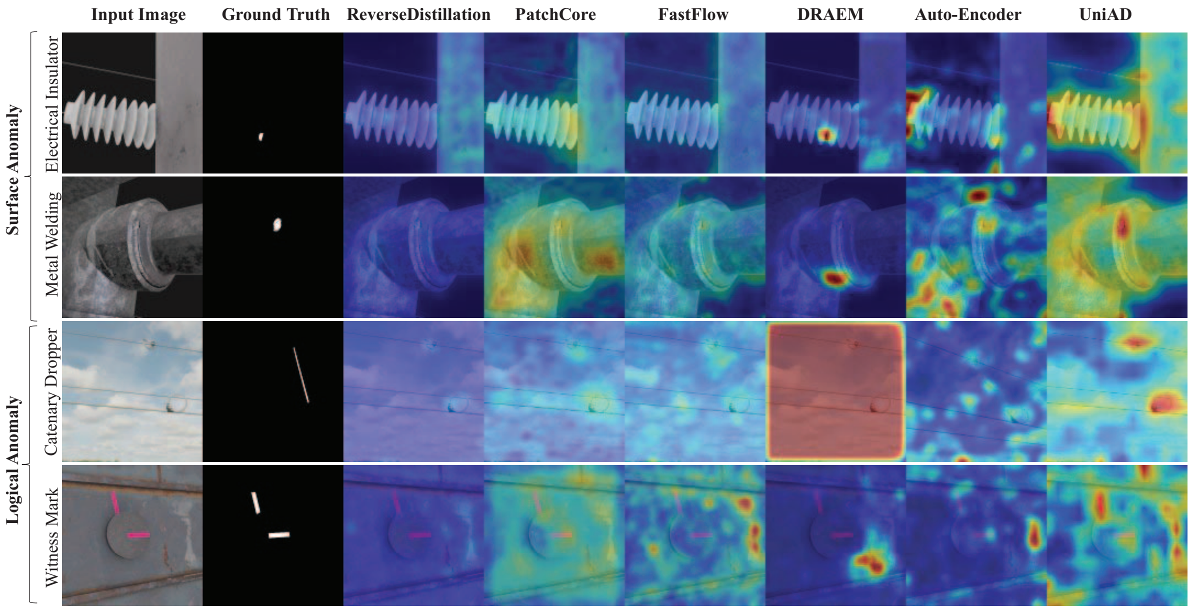

Figure 4. Qualitative anomaly segmentation results for each evaluated method on the MIAD dataset.

Factors	RD	PatchCore	FastFlow	DRAEM	Mean
baseline	0.95/0.42	1.00/1.00	1.00/1.00	1.00/1.00	0.99/0.86
+UV	0.61/0.04	1.00/1.00	0.88/0.06	0.66/0.08	0.79/0.30
+UB	0.57/0.03	0.93/0.39	0.92/0.28	0.86/0.27	0.82/0.24
+US	0.64/0.04	1.00/0.99	0.99/0.65	0.47/0.02	0.78/0.43
+UV+US+UB	0.57/0.03	0.71/0.02	0.51/0.01	0.52/0.01	0.58/0.02

Table 6. Impact of uncontrolled factors under outdoor environment on the witness mark. For each method, the image-level AUROC (the first number) and Recall@1%FPR (the second number) are given. RD stands for Reverse Distillation and AE stands for Auto-Encoder.

Size	RD	PatchCore	FastFlow	DRAEM
1000(10%)	0.56/0.35	0.54/0.67	0.51/0.51	0.63/0.51
2000(20%)	0.62/0.44	0.55/0.65	0.53/0.51	0.67/0.51
4000(40%)	0.65/0.48	-/-	0.53/0.50	0.67/0.51
8000(80%)	0.67/0.48	-/-	0.54/0.51	0.78/0.52
10000(100%)	0.68/**0.57**	-/-	0.53/0.51	**0.81**/0.52

Table 7. Results of different training dataset size. For each algorithm, the image-level AUROC on the electrical insulator and witness mark is given. Due to GPU memory limitation, PatchCore is failed to run when training data size is larger than 2000.

size is partly due to costly collection, but the main reason is that hundreds of samples is enough to stands for the distribution of roughly aligned non-defective samples. We believe an order of magnitude larger training data size is essential for outdoor anomaly detection to describe the complex non-defective objects. In order to verify this thought, we sample training data size from 1000 to 10000 on the electrical insulator which is a represent for surface anomaly and the witness mark which is a represent for logical anomaly. As shown in Table 7, Reverse Distillation enjoys the growth of the training data size, while PatchCore and FastFlow obtain a little benefit after the training dataset size is larger than 2000. DRAEM, which is more preferable for surface anomaly, benefit a lot from the extensive training data on the electrical insulator but fail on the witness mark. Furthermore, PatchCore is failed to run when training data size is larger than 2000 due to GPU memory limitation. How to exploit the bonus of massive non-defective samples in the outdoors is an interesting research direction.

5. Conclusion

We introduce the MIAD dataset, a novel dataset for unsupervised anomaly detection in various maintenance inspections. This dataset is intended for tasks under uncontrolled environments including uncontrolled viewpoints, uncontrolled backgrounds and uncontrolled surfaces. It provides researchers with sufficient images, which are an order of magnitude larger than existing datasets, to facilitate algorithms to learn the complex distribution of outdoor non-defective samples. Some representative methods including Reverse Distillation, PatchCore, FastFlow, DRAEM, Auto-Encoders and UniAD are evaluated on the MIAD dataset, and the significant drops in performance demonstrate that the maintenance inspection is more challenging than manufacturing inspection. We expect that MIAD can attract more attention of the research community on the outdoor maintenance inspection and that worthwhile future work can be spawned from the proposed dataset.

References

[1] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *IJCV*, 2022. 3

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019. 2, 3, 6

[3] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The MVTec 3D-AD dataset for unsupervised 3D anomaly detection and localization. *arXiv preprint arXiv:2112.09045*, 2021. 3

[4] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2019. 4, 6

[5] Tadhg Brosnan and Da-Wen Sun. Improving quality inspection of food products by computer vision—a review. *Journal of food engineering*, 2004. 1

[6] Junwen Chen, Zhigang Liu, Hongrui Wang, Alfredo Núñez, and Zhiwei Han. Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network. *IEEE Transactions on Instrumentation and Measurement*, 2017. 1, 2, 3, 4

[7] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. In *arXiv preprint arXiv:2005.02357*, 2020. 4

[8] Yajie Cui, Zhaoxiang Liu, and Shiguo Lian. A survey on unsupervised industrial anomaly detection algorithms. In *arXiv preprint arXiv:2204.11161*, 2022. 2, 3

[9] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *ICPR*, 2021. 4

[10] Sergiu Deitsch, Vincent Christlein, Stephan Berger, Claudia Buerhop-Lutz, Andreas Maier, Florian Gallwitz, and Christian Riess. Automatic classification of defective photovoltaic module cells in electroluminescence images. *Solar Energy*, 185:455–468, June 2019. 2

[11] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, 2022. 1, 4, 6

[12] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019. 2, 5

[13] Abd Al Rahman M. Abu Ebayyeh and Alireza Mousavi. A review and analysis of automatic optical inspection and quality monitoring methods in electronics industry. *IEEE Access*, 2020. 1

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 4

[15] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. *arXiv preprint arXiv:2107.12571*, 2021. 4

[16] Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 36(1):85–96, 2020. 3

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 4

[18] Yann LeCun. Generalization and network design strategies. *Connectionism in perspective*, 1989. 4

[19] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *arXiv preprint arXiv:2206.04325*, 2022. 4

[20] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 4

[21] Xiaoxia Li, Qiang Yang, Zhebo Chen, and Wenjun Yan. Visible defects detection based on uav-based inspection in large-scale photovoltaic systems. *IET Renewable Power Generation*, 2017. 1, 3, 4

[22] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021. 2

[23] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. 5

[24] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *PMLR*, 2015. 4

[25] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022. 1, 2, 4, 6

[26] Marco Rudolph, Bastian Wandt, and Bodo Rosenha. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021. 4

[27] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Structuring autoencoders. In *ICCV*, 2019. 4

[28] Javier Silvestre-Blanes, Teresa Albero-Albero, Ignacio Miralles, Rubén Pérez-Llorens, and Jorge Moreno. A public fabric database for defect detection methods and results. *Autex Research Journal*, 2019. 1, 2

[29] A. Sriram and T. D. Sudhakar. Technology revolution in the inspection of power transmission lines-a literature review. In *2021 7th International Conference on Electrical Energy Systems (ICEES)*, pages 256–262. IEEE, 2021. 1, 4

[30] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, May 2019. 2

[31] Xian Tao, Dapeng Zhang, Wenzhi Ma, Xilong Liu, and De Xu. Automatic metallic surface defect detection and recognition with convolutional neural networks. *Applied Sciences*, 2018. 1

[32] Xian Tao, Dapeng Zhang, Zihao Wang, Xilong Liu, Hongyan Zhang, and De Xu. Detection of power line insulator defects using aerial images analyzed with convolutional neural networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 50(4):1486–1498, 2018. 3

[33] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for unsupervised anomaly detection. In *BMVC*, 2021. 4

[34] Wenjie Wang, Yu Xue, Chengkuan He, and Yongnian Zhao. Review of the typical damage and damage-detection methods of large wind turbine blades. *Energies*, 2022. 1, 4

[35] Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM Symposium*, 2007. 2

[36] Shinji Yamada and Kazuhiro Hotta. Reconstruction student with attention for student-teacher pyramid matching. In *arXiv preprint arXiv:2111.15376*, 2021. 4

[37] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *NeurIPS*, 2022. 4, 6

[38] 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*, 2021. 1, 4, 6

[39] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM – A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 1, 4, 6

[40] Ye Zheng, Xiang Wang, Rui Deng, Tianpeng Bao, Rui Zhao, and Liwei Wu. Focus your distribution: Coarse-to-fine non-contrastive learning for anomaly detection and localization. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 4