Image anomaly detection for heterogeneous datasets

Abstract

Unsupervised anomaly detection (AD) is a critical task in various domains, from manufacturing to infrastructure monitoring. To advance this field, we introduce two novel datasets: CARS-AD and ROADS-AD, designed to challenge existing AD methods with their diverse and heterogeneous image content. CARS-AD comprises real images of cars with various defects, while ROADS-AD contains images of roads from multiple countries, each presenting unique challenges in anomaly detection. These datasets provide ground truth pixel-wise annotations, enabling detailed evaluation and benchmarking of AD algorithms.

We evaluate state-of-the-art AD methods on both datasets, using the AUROC metric to assess detection and localization performance. Our results reveal significant room for improvement, underscoring the complexity of the datasets and the need for robust AD techniques. Notably, Csflow and U-Flow demonstrate superior performance on the CARS-AD Dataset, leveraging their ability to process multi-scale features effectively. Conversely, Reverse Distillation excels in anomaly localization on the ROADS-AD Dataset, showcasing the importance of nuanced approaches for diverse anomaly types.

Our findings underscore the importance of addressing the challenges posed by heterogeneous datasets in unsupervised AD. We hope that the introduction of CARS-AD and ROADS-AD will inspire further research in this field, driving the development of innovative AD methods capable of handling real-world anomalies with greater accuracy and reliability.

1. Introduction

Anomaly detection (AD) addresses the challenge of identifying anomalous samples. In the context of image processing, AD aims to detect samples that deviate from the distribution encompassing most normal and healthy samples. A key challenge in AD is to accurately model this distribution of normality [9, 10, 34]. Additionally, anomalies can often be localized, a process referred to as "Anomaly Localization" or "Anomaly Segmentation."

While AD is fundamentally a classification task that seeks to distinguish between normal and anomalous samples, most supervised methods struggle due to a common issue in AD applications: the scarcity of anomalous samples. Typically, there is an abundance of healthy samples but a limited number of anomalous ones. This scarcity makes the task more challenging, although even a small number of anomalous samples can help alleviate the problem [34].

Image-based AD has diverse applications, from medical image analysis [15,55], to industrial manufacturing [5,32], and video surveillance [6,24]. These methods significantly reduce the need for manual inspection, which can be time-consuming and prone to human error, especially in industrial settings where detecting small defects is critical.

The availability of benchmark datasets such as ImageNet [12], CIFAR-10 [27], MNIST [13], and more recent AD-specific datasets like MVTec-AD [5] and BTAD [32] has led to significant advancements in image-based AD. Deep learning methods leverage the repetitive patterns and homogeneity of non-anomalous samples in AD datasets to extract features that represent normality [3, 31, 41]. Autoencoder-based solutions (AE) train on non-anomalous samples to reconstruct input images, with failures in reconstruction indicating anomalous samples [30, 44, 49, 51, 53].

However, homogeneity is not guaranteed in all applications, and extracting representations of normality from heterogeneous datasets is significantly more challenging. The combination of scarce anomalous samples and high variability among training samples presents a substantial challenge in image-based AD.

In this paper, we focus on image-based AD for heterogeneous datasets, a largely unexplored area in computer vision, and to the best of our knowledge, with no existing datasets specifically designed for this purpose. Our main contributions are:

- We introduce two benchmark datasets comprising heterogeneous images for AD, featuring various types of anomalies in cars and roads. These anomalies are classified based on their appearance, nature, and severity. The datasets include ground truth annotations for both anomaly detection (image-level classification) and anomaly localization (pixel-level segmentation).
- We provide a comprehensive evaluation of the performance of state-of-the-art AD methods on our datasets.

This evaluation reveals the unique challenges posed by heterogeneous datasets compared to more homogeneous datasets like MVTec [5]. The analysis highlights the need for robust AD techniques capable of handling diverse anomaly types and complex image structures.

- We contribute to advancing the field of unsupervised anomaly detection by addressing the limitations of existing datasets and providing a new benchmark for evaluating AD algorithms in real-world scenarios. Through the introduction of CARS-AD and ROADS-AD, we aim to inspire further research and innovation in AD, ultimately leading to the development of more effective anomaly detection solutions.

2. RELATED WORKS

2.1. Learning Methods for AD

Fully-supervised Learning When anomalous samples and their corresponding ground truth annotations are available in the datasets, deep convolutional neural networks (DCNNs) can be trained in a fully-supervised manner. However, due to the unique characteristics of most AD applications—particularly the scarcity of anomalous training samples, which are often nonexistent—fully-supervised methods are typically impractical. Although generating synthetic anomalous samples for fully-supervised learning is possible, deep learning models tend to learn specific cues from synthetic data, leading to poor performance during inference on real datasets [48]. Fully-supervised methods must also address the problem of class imbalance, as the number of non-anomalous samples usually far exceeds that of anomalous samples. This necessitates sophisticated techniques to balance the learning process [8, 35].

Unsupervised Learning In unsupervised settings, no labeled data is available, and thus no ground truth is provided for either anomalous or non-anomalous training samples. Some methods address unsupervised AD by operating in a feature space, using clustering or feature matching techniques [16, 17, 33, 45]. Unsupervised methods rely heavily on the capability of feature extractors—often variations of architectures like VGG [39], ResNet [18], or Inception [42]—to map input data into a feature space where similar images are clustered together and are distant from clusters of dissimilar images.

Semi-supervised Learning For most AD applications, it is significantly easier to collect labels for normal data, while anomalous data and its corresponding labels are difficult to obtain. In this setup, semi-supervised learning techniques can be employed, where the goal is to find a representation of normality based on normal training data [11, 46, 53]. These architectures learn a model for non-anomalous data, which is later used to classify samples that do not fit the model as anomalous. A common approach is to use autoencoders (AEs) trained on non-anomalous data, which learn to reconstruct normal samples with low error rates but fail to reconstruct anomalous samples accurately [20, 21, 23, 41].

2.2. Anomaly Detection Datasets

AD benchmarks were initially established using general image classification datasets in a one-vs-rest fashion [14]. Datasets such as ImageNet [12], CIFAR-10 [27], and MNIST [13] are commonly used as one-vs-rest AD benchmarks, where one class is arbitrarily designated as "normal" and other classes as "abnormal." These datasets can be extended with images from external sources to increase the number of anomalous samples, a method known as outlier exposure (OE) introduced by Ruff *et al*. [19]. However, as noted by Bergmann *et al*. [5], the one-vs-rest approach results in a setup where the distribution of anomalous samples is significantly different from that of normal samples, posing challenges in applications where anomalous and normal samples are more similar.

Ruff *et al*. [38] demonstrated that standard binary cross-entropy classifiers outperform unsupervised methods on one-vs-rest AD benchmarks using the OE setup, due to the rich information content in the multiscale structure of natural images like those in ImageNet and CIFAR-10. Consequently, Ruff *et al*. [38] called for more research on challenging datasets like MVTec-AD to gain groundbreaking insights in AD.

MVTec-AD and BTAD, in contrast, are specifically designed for AD applications. MVTec-AD comprises images of real-world industrial components categorized into different types of industrial parts (e.g., "Transistor", "Screw"). Each category includes "good" (normal) and "defective" (anomalous) samples, with defective samples further subcategorized based on the nature of the defects (e.g., "crack", "hole"). BTAD includes real-world industrial anomaly images of three different industrial products.

While MVTec-AD introduced a more interesting challenge in AD, where the distributions of non-anomalous and anomalous samples are closer, the homogeneity of MVTec-AD data does not realistically approximate other AD applications dealing with larger objects in less controlled environments, such as cars and roads. The categories in MVTec-AD correspond to industrial objects that are very similar in appearance, as shown in Figure Fig. 1, which is common in industrial production lines where each component is replicated at a large scale. This high intra-class similarity, quantified using the structural similarity index measure (SSIM), facilitates the extraction of features that represent "normality" with a characteristic distribution. We computed the SSIM score for MVTec and compared the results to CARS-

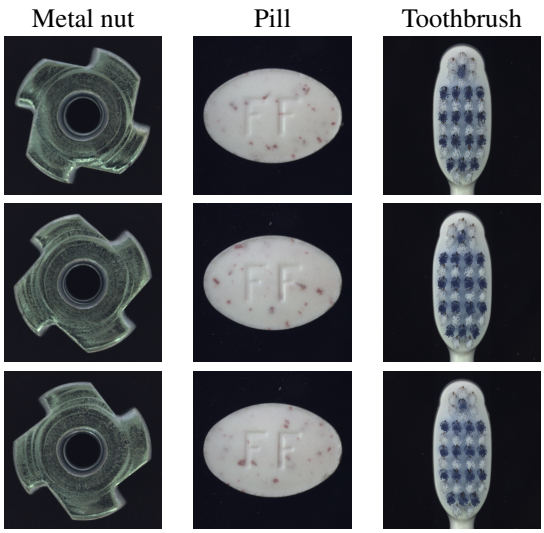

Figure 1. MVTec-AD dataset overview. Each column shows three different objects corresponding to three different categories of the MVTec-AD dataset, namely "Metal nut", "Pill", and "Toothbrush". Images in the same category and their corresponding objects are visually very similar, which is a common trait in real-world industrial applications.

Table 1. Performance of State-of-the-art methods for the MVTec-AD benchmark as measured by the image-level area under the Receiver Operating Characteristic curve (AUROC). The results are calculated as an average over all the categories in MVTec-AD dataset.

Method	AUROC (%)
GANomaly [1]	78.0
DRÆM [53]	98.00
DSR [54]	98.20
Reverse Distillation [11]	98.50
Csflow [37]	98.70
U-Flow [43]	98.87
EfficientAD [4]	99.10
CFA [28]	99.30
Fastflow [52]	99.40

AD and ROADS-AD datasets as shown in Figure Fig. 2. Our datasets, CARS-AD Dataset and ROADS-AD Dataset, provide heterogeneous samples introducing new challenges to AD tasks.

State-of-the-art methods have already achieved highly accurate results on the MVTec-AD benchmark, as shown in Table Tab. 1 with AUROC values exceeding 99%. These results highlight the need for new challenges to further advance AD research.

Motivated by this, we introduce two new benchmarks for AD, consisting of heterogeneous images with larger intra-class distances. In other words, images within any given

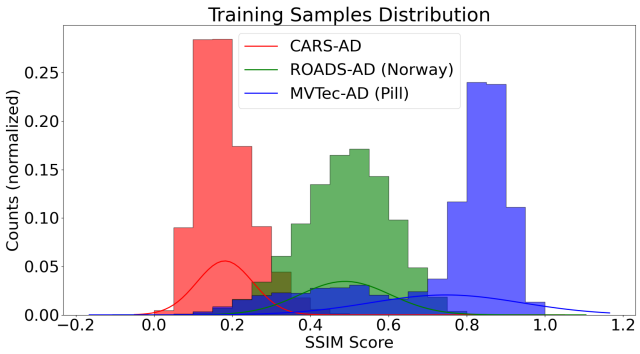

Figure 2. Histogram distributions of training samples for MVTec-AD (Pill), CARS-AD Dataset, and ROADS-AD Dataset (Norway) with respect to the structural similarity index measure (SSIM). Our datasets, CARS-AD and ROADS-AD Dataset contain more dissimilar training samples, thus introducing a new challenge that we regard as "Anomaly Detection for Heterogeneous Datasets"

category are more dissimilar, with less repetitive patterns in color, texture, and shape. Our goal is to motivate the community to develop new methods and techniques to tackle the challenges introduced by these new benchmarks.

3. Datasets Description

3.1. CARS-AD Dataset

The CARS-AD dataset consists of real images of several types of cars, each image depicting a single car. The dataset is split into a train set and a test set. The train set contains 4301 images without defects, while the test set contains 3955 images with and without defects. The defects in the test set are categorized into four main types: "broken_part", "dent", "scratch", and "wind_screen". Some samples may contain multiple defects, and their labels follow the pattern "label1_label2", for example, a car with scratches and a broken windscreen is labeled as "scratch_wind_screen". Healthy samples are labeled as "good".

Unlike MVTec-AD, where most defects are relatively small, the defects in the CARS-AD dataset can cover large areas of the image. While for the human eye it may be easy to spot different kinds of defects in cars regardless of their type, color, and shape, such variability poses significant challenges for deep anomaly detection methods.

The images in the CARS-AD dataset were collected from multiple sources, including the Standford Car Dataset [26], CarDD Dataset [47], Internal and External Parts Dataset [50], and Car Parts and Car Damages Dataset [22]. The cars were extracted from the background using Grounding DINO [29] for localization and Segment Anything [25] for segmentation, as shown in Figure Fig. 4. Additionally, a Gaussian filter was applied to each image before segmentation to reduce noise while preserving the

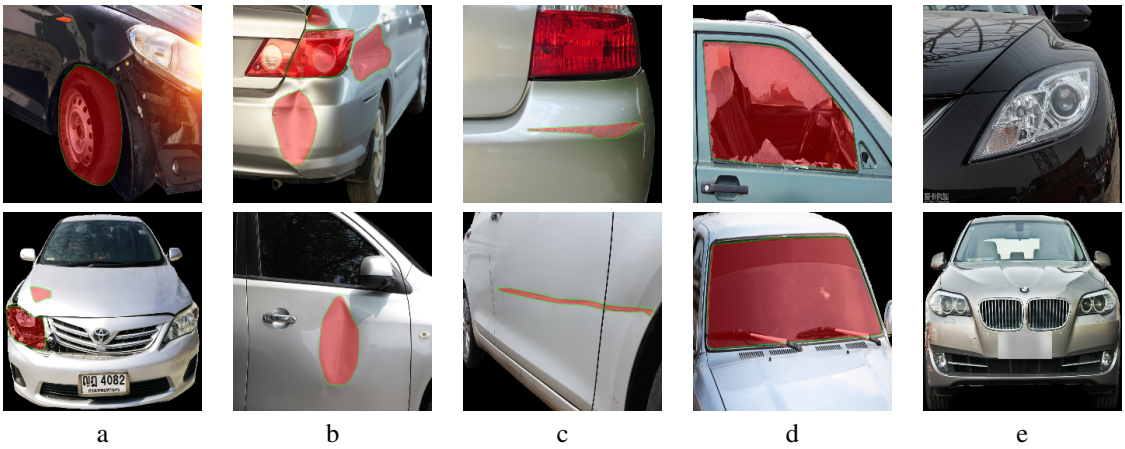

Figure 3. Overview of the CARS-AD Dataset. The dataset provides images of cars with different types of defects as well as healthy samples. The defects are categorized into four main categories: (a) "broken_part", (b) "dent", (c) "scratch", (d) "wind_screen" and column (e) shows instances of healthy samples.

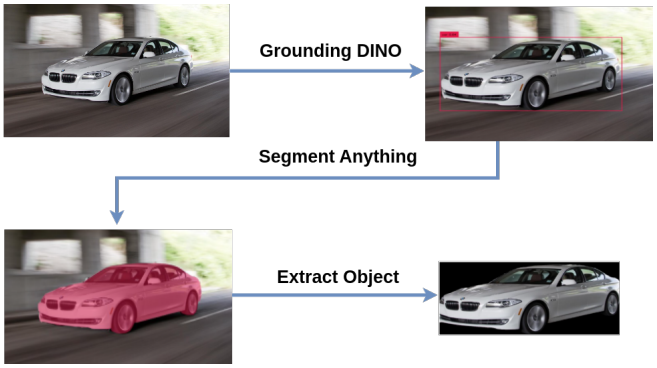

Figure 4. The objects in our datasets were extracted from the background using Grounding DINO for bounding box object detection and Segment Anything for pixel-wise segmentation. In this particular figure, we used "car" as text prompt input to Grounding DINO, then, we used the bounding box coordinates as input to Segment Anything

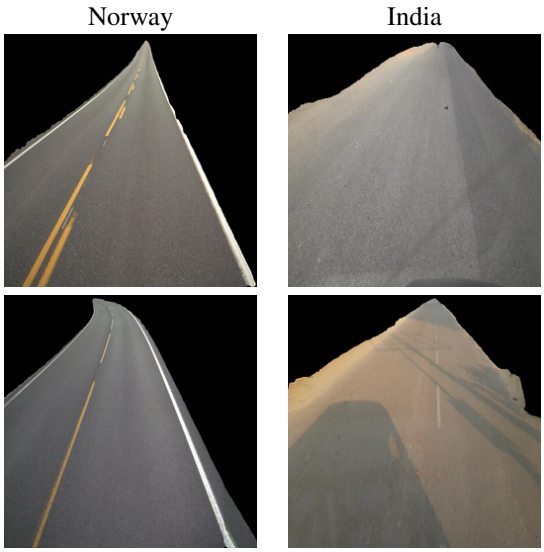

Figure 5. Images of roads in Norway on the left column and roads in India on the right column. The roads in different countries show differences that originate from multiple factors such as climate, terrain, infrastructure development, and cultural preferences.

edges of the objects. In cases where Segment Anything returned multiple masks for the same object, the masks were merged and a polygon was fitted around the object using the OpenCV contours function to fill gaps [7]. Samples that did not produce high-quality masks were discarded. The anomalous samples in the test set were taken from the RDD dataset, which provides bounding box annotations for the anomalous regions. These bounding boxes were used as inputs to Segment Anything to obtain ground truth masks for the anomalous regions. An overview of the CARS-AD dataset is shown in Figure Fig. 3.

3.2. ROADS-AD Dataset

The ROADS-AD dataset consists of real images of roads from multiple countries, specifically China, the Czech Republic, India, Japan, Norway, and the U.S.A. This dataset targets anomaly detection applications by providing both healthy samples of roads in good condition and anomalous samples with defects such as cracks, holes, or surface discontinuities.

This dataset introduces several interesting challenges. Different countries build and maintain their roads differ-

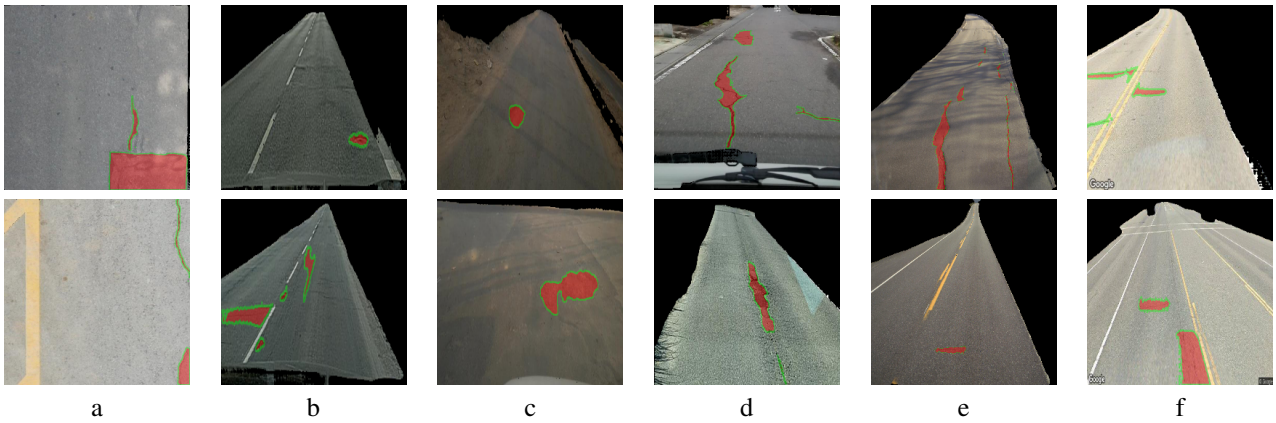

Figure 6. Overview of the Roads-AD Dataset. Each column depicts images of roads from different countries available in the Roads-AD Dataset, more specifically: (a) China, (b) the Czech Republic, (c) India, (d) Japan, (e) Norway, (f) the U.S.A. Anomalous regions are highlighted in red.

ently, which means the assessment of road conditions depends heavily on geographical location and cultural context. For instance, as shown in Figure Fig. 5, roads in Norway and India differ significantly due to various factors such as climate, terrain, infrastructure development, and cultural preferences. Norway's roads are subjected to harsh weather conditions, particularly in winter, necessitating the use of road salt and winter tires. In contrast, roads in India must endure heavy monsoon rains, leading to flooding and erosion, with materials adapted accordingly. Norway's terrain features mountains and fjords, requiring engineering solutions like tunnels and retaining walls, whereas India's diverse terrain includes plains, mountains, and deserts. Traffic and population density also affect road appearance, with densely populated areas in India requiring more traffic signs and narrower roads, while less densely populated Norway allows for wider roads and fewer signs. Cultural differences also play a role, with Norway prioritizing safety, resulting in well-marked pedestrian crossings and dedicated bicycle lanes, giving a distinct appearance compared to Indian roads.

The images in the ROADS-AD dataset were collected from the Road Damage Detection Dataset [40] and free licensed road images from [36]. Following the same process as in the CARS-AD dataset, Grounding DINO was used to localize the roads with bounding boxes, and Segment Anything was used to segment the roads at a pixel level. Anomalous regions were annotated using the bounding boxes provided by the RDD as prompts to SAM. Manual inspection ensured the accuracy of the annotations. An overview of the ROADS-AD dataset is shown in Figure Fig. 6.

4. Benchmark

We used the open-source library *Anomalib* [2] to conduct extensive experiments with relevant state-of-the-art methods for unsupervised image anomaly detection to establish a benchmark on both of our datasets, CARS-AD and ROADS-AD. We analyze the performance of each method for all the different categories in the datasets and discuss the advantages and drawbacks of each architecture. The performance of the evaluated methods is compared by the AUROC at both the image level and pixel level. All methods were trained using a single Nvidia V100 32GB GPU.

4.1. Evaluated Methods

EfficientAD This architecture is based on the student-teacher paradigm, where the student network is trained to imitate the teacher and predict features on healthy samples. At test time, the student network fails to predict the corresponding features for anomalous samples, thus allowing the detection of anomalies. We trained EfficientAD for 400 epochs and used Adam as an optimizer with an initial learning rate of $1e-4$.

Reverse Distillation This is a variation of a conventional student-teacher architecture, where the student receives high-level features from the teacher and attempts to reproduce the teacher's multi-scale representations. We trained Reverse Distillation for 200 epochs and used Adam as an optimizer with an initial learning rate of $5e-3$.

CFA This method proposes a feature adaptation strategy to perform transfer learning on a target dataset to prevent the normality of abnormal features from being overestimated when using a pre-trained CNN. We trained CFA for

	Category				
Method	Broken Part	Scratch	Dent	Wind Screen	All
EfficientAD [4]	0.51 / 0.59	0.57 / 0.67	0.61 / 0.75	0.44 / 0.66	0.54 / 0.64
Reverse Distillation [11]	0.58 / 0.76	0.45 / 0.75	0.44 / 0.62	0.79 / 0.86	0.59 / **0.77**
CFA [28]	0.60 / 0.54	0.66 / 0.45	0.62 / 0.37	0.71 / 0.49	0.64 / 0.49
DRÆM [53]	0.67 / 0.75	0.56 / 0.68	0.55 / 0.62	0.69 / 0.75	0.64 / 0.70
GANomaly [1]	0.68 / -	0.57 / -	0.63 / -	0.62 / -	0.64 / -
Csflow [37]	**0.87** / 0.80	**0.92** / 0.83	**0.91** / **0.88**	**0.93** / 0.90	**0.91** / 0.69
U-Flow [43]	0.78 / **0.80**	0.82 / **0.85**	0.78 / 0.81	0.93 / **0.93**	0.81 / 0.74
Fastflow [52]	0.60 / 0.62	0.67 / 0.59	0.69 / 0.56	0.61 / 0.59	0.64 / 0.56
DSR [54]	0.58 / 0.56	0.69 / 0.59	0.72 / 0.57	0.76 / 0.53	0.69 / 0.56

Table 2. Results of all the methods evaluated on the CARS-AD Dataset. We computed the AUROC metric at the image level and pixel level for every method. The numbers are presented as (AUROC image level/AUROC pixel level) and the best-performing methods are highlighted in bold numbers.

	Category				
Method	Czech Republic	India	Japan	Norway	All
EfficientAD [4]	0.46 / 0.49	0.66 / 0.49	0.23 / 0.55	0.57 / 0.65	0.41 / 0.59
Reverse Distillation [11]	0.58 / **0.81**	0.71 / 0.72	0.55 / **0.76**	0.63 / **0.90**	0.57 / **0.77**
CFA [28]	0.59 / 0.70	0.61 / 0.62	0.50 / 0.70	0.61 / 0.72	0.67 / 0.53
DRÆM [53]	0.57 / 0.70	0.73 / **0.75**	0.49 / 0.60	0.65 / 0.79	0.66 / 0.67
GANomaly [1]	0.43 / -	0.43 / -	0.39 / -	**0.77** / -	0.68 / -
Csflow [37]	0.55 / 0.67	**0.80** / 0.61	**0.96** / 0.62	0.73 / 0.78	0.73 / 0.70
U-Flow [43]	0.57 / 0.65	0.77 / 0.67	0.74 / 0.69	0.66 / 0.86	0.74 / 0.70
Fastflow [52]	**0.58** / 0.63	0.71 / 0.57	0.71 / 0.60	0.68 / 0.78	0.74 / 0.62
DSR [54]	0.53 / 0.38	0.73 / 0.57	0.63 / 0.49	0.67 / 0.49	**0.78** / 0.46

Table 3. Results of all the methods evaluated on the ROADS-AD Dataset. We computed the AUROC metric at the image level and pixel level for every method. The numbers are presented as (AUROC image level/AUROC pixel level) and the best-performing methods are highlighted in bold numbers.

60 epochs and used AdamW as an optimizer with an initial learning rate of $1e-3$.

DRÆM This architecture consists of one reconstructive sub-network and one discriminative sub-network connected in cascade. The reconstructive network is trained to reconstruct an anomaly-free image from synthetically generated anomalous samples, while the discriminative network learns to localize the anomalies using the original image and the reconstructed image as inputs. We trained DRÆM for 600 epochs and used Adam as an optimizer with an initial learning rate of $1e-4$.

GANomaly This approach uses a generative adversarial network (GAN) with a novel encoder-decoder-encoder variation. The encoder-decoder sub-network reconstructs the generated output image while the additional encoder maps the generated image to a lower latent space, allowing the network to learn the distribution of normal samples. We trained GANomaly for 60 epochs and used Adam as an optimizer with an initial learning rate of $2e-4$.

Csflow This method processes multiple feature maps at different scales jointly, using a cross-scale normalizing flow that integrates both global and local contexts. This approach allows for the estimation of the likelihood of a sample being anomalous, as well as anomaly localization at the pixel level. We trained Csflow for 240 epochs and used Adam as an optimizer with an initial learning rate of $2e-4$.

U-Flow This model consists of three main sub-networks connected in a cascade. First, a multi-scale feature extractor; second, a U-shaped normalizing flow sub-network operating on the multi-scale features returned by the feature extractor; and finally, U-flow computes the anomaly score and segmentation, where the threshold for detection is set automatically by controlling the ratio of false detections. We trained U-Flow for 200 epochs and used Adam as an

optimizer with an initial learning rate of $1e-3$.

Fastflow By using normalizing flows, Fastflow maps an input image from the original distribution to a normal distribution. The computed features corresponding to healthy regions on the image fall into the center of the new normal distribution, whereas anomalous regions produce features that fall far from the center of the distribution. We trained Fastflow for 500 epochs and used Adam as an optimizer with an initial learning rate of $1e-3$.

DSR Instead of generating synthetic anomalies directly on the input images for training the model, DSR synthesizes anomalies on a quantized feature space, allowing for the generation of synthetic near-in-distribution anomalies without requiring prior knowledge about the appearance of the anomalies. We trained DSR for 200 epochs and used Adam as an optimizer with an initial learning rate of $2e-4$.

5. Results

The quantitative results for the CARS-AD Dataset and the ROADS-AD Dataset are presented in Tables Sec. 4 and Sec. 4 respectively. In general, the performance of all methods is considerably lower compared to their performance on other standard AD datasets such as MVTec-AD (Table Tab. 1), which is the expected behavior due to the challenging characteristics and heterogeneity of our datasets.

The performance on the CARS-AD Dataset was evaluated separately across different categories as shown in Table Sec. 4, specifically "Broken Part", "Scratch", "Dent", and "Wind Screen". The right-most column "All" comprises all categories including those composed of multiple defects, e.g., "scratch_wind_screen". Csflow and U-Flow outperform the other methods in all categories, and only Reverse Distillation shows better performance in AUROC at the pixel level when all existing categories are evaluated jointly.

Csflow and U-Flow are specifically designed to process feature maps at multiple scales, which gives them an advantage when localizing anomalies in the CARS-AD Dataset. This dataset consists of images with non-uniform surfaces, different types of multi-scale structures, textures, and colors, making it infeasible to localize anomalies accurately unless both local and global features are considered and processed jointly. Furthermore, the anomalies in the CARS-AD dataset vary in scale; for instance, scratches are typically localized in small regions, while other defects such as broken parts are considerably larger. This explains the superior performance of Csflow, followed by the comparable results of U-Flow. Visual results obtained with Csflow are shown in Figure Fig. 7.

The evaluation results of the ROADS-AD Dataset show that no single method consistently dominates across all categories. Similar to the CARS-AD Dataset, all methods exhibit poorer performance compared to their performance on other traditional AD datasets in general. However, Reverse Distillation outperforms most methods in anomaly localization, as measured by the AUROC at the pixel level. The reversed teacher-student architecture of Reverse Distillation, where the student is fed with the low-dimensional features of the teacher, helps the model generalize better to unseen diverse anomalies that are not propagated through the low-dimensional layers at the end of the teacher sub-network. This is enhanced by the one-class bottleneck embedding (OCBE) introduced by Deng and Li [11].

6. Conclusions

We present two novel datasets for unsupervised anomaly detection (AD), namely CARS-AD and ROADS-AD. In contrast with previous AD datasets, our datasets introduce new challenges due to the heterogeneous nature of the images they contain. CARS-AD and ROADS-AD consist of thousands of samples that exhibit significant variation in appearance. We also provide ground truth pixel-wise annotations for all samples, facilitating detailed analysis of AD methods.

We evaluated multiple state-of-the-art methods on both datasets using the AUROC metric for both anomaly detection and localization, thus establishing an initial benchmark. The results indicate that there is considerable room for improvement. Csflow and U-Flow demonstrated the best overall performance on the CARS-AD Dataset, particularly due to their ability to process feature maps at multiple scales. On the ROADS-AD Dataset, Reverse Distillation showed superior performance in anomaly localization.

These findings suggest that the challenging nature of our datasets, with their diverse and complex anomalies, highlights the limitations of current methods. We hope that these datasets will motivate the research community to embrace the challenges introduced here and to develop new techniques and methods for anomaly detection. By doing so, we aim to advance the state of the art in this field and improve the robustness and accuracy of AD systems.

Acknowledgments

References

[1] Samet Akcay, Amir Atapour Abarghouei, and Toby P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Pa-*

pers, Part III, volume 11363 of *Lecture Notes in Computer Science*, pages 622–637. Springer, 2018. 3, 6

[2] Samet Akcay, Dick Ameln, Ashwin Vaidya, Barath Lakshmanan, Nilesh Ahuja, and Utku Genc. Anomalib: A deep learning library for anomaly detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1706–1710, 2022. 5

[3] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver, editors, *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pages 37–49, Bellevue, Washington, USA, 02 Jul 2012. PMLR. 1

[4] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. *CoRR*, abs/2303.14535, 2023. 3, 6

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592, 2019. 1, 2

[6] B. Boghossian and J. Black. The challenges of robust 24/7 video surveillance systems. In *The IEE International Symposium on Imaging for Crime Detection and Prevention, 2005. ICDP 2005.*, pages 33–38, 2005. 1

[7] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 4

[8] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks : the official journal of the International Neural Network Society*, 106:249–259, 2017. 2

[9] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *CoRR*, abs/1901.03407, 2019. 1

[10] Varun Chandola and Vipin Kumar. Outlier detection : A survey. *ACM Computing Surveys*, 41, 01 2009. 1

[11] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9727–9736, 2022. 2, 3, 6, 7

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1, 2

[13] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 1, 2

[14] Andrew Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. Systematic construction of anomaly detection benchmarks from real data. *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, ODD 2013*, pages 16–21, 08 2013. 2

[15] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection - A survey. *CoRR*, abs/2012.02364, 2020. 1

[16] Zahra Ghafoori and Christopher Leckie. Deep multi-sphere support vector data description. In *SDM*, 2020. 2

[17] Joris Guérin and Byron Boots. Improving image clustering with multiple pretrained cnn feature extractors. In *British Machine Vision Conference*, 2018. 2

[18] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 2

[19] Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. *CoRR*, abs/1812.04606, 2018. 2

[20] Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. *CoRR*, abs/2107.13118, 2021. 2

[21] Chaoqin Huang, Jinkun Cao, Fei Ye, Maosen Li, Ya Zhang, and Cewu Lu. Inverse-transform autoencoder for anomaly detection. *ArXiv*, abs/1911.10676, 2019. 2

[22] Humans in the Loop. Car parts and car damages dataset. https://humansintheloop.org/resources/datasets/car-parts-and-car-damages-dataset/, 2024. 3

[23] Ki Hyun Kim, Sangwoo Shim, Yongsub Lim, Jongseob Jeon, Jeongwoo Choi, Byungchan Kim, and Andre S. Yoon. Rapp: Novelty detection with reconstruction along projection pathway. In *International Conference on Learning Representations*, 2020. 2

[24] B. Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *CoRR*, abs/1801.03149, 2018. 1

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE, 2023. 3

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 3

[27] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012. 1, 2

[28] Sungwook Lee, Seunghyun Lee, and Byung Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10:1–1, 01 2022. 3, 6

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *CoRR*, abs/2303.05499, 2023. 3

[30] WenQian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J. Radke, and Octavia I. Camps. Towards visually explaining variational autoencoders. *CoRR*, abs/1911.07389, 2019. 1

[31] Pankaj Mishra, Claudio Piciarelli, and Gian Foresti. A neural network for image anomaly detection with deep pyramidal representations and dynamic routing. *International Journal of Neural Systems*, 30, 07 2020. 1

[32] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. *CoRR*, abs/2104.10036, 2021. 1

[33] Wafa Mousser and Salima Ouadfel. Deep feature extraction for pap-smear image classification: A comparative study. *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, 2019. 2

[34] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *CoRR*, abs/2007.02500, 2020. 1

[35] Claudio Piciarelli, Pankaj Mishra, and Gian Luca Foresti. Image anomaly detection with capsule networks and imbalanced datasets. In *International Conference on Image Analysis and Processing*, 2019. 2

[36] Pixabay. Pixabay. https://pixabay.com/, 2024. Accessed: January 16, 2024. 5

[37] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1829–1838. IEEE, 2022. 3, 6

[38] Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Klaus-Robert Müller, and Marius Kloft. Rethinking assumptions in deep anomaly detection. *CoRR*, abs/2006.00339, 2020. 2

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2

[40] Janpreet Singh and Shashank Shekhar. Road damage detection and classification in smartphone captured images using mask r-cnn. *ArXiv*, abs/1811.04535, 2018. 5

[41] Hongchao Song, Zhuqing Jiang, Aidong Men, and Bo Yang. A hybrid semi-supervised anomaly detection model for high-dimensional data. *Computational Intelligence and Neuroscience*, 2017, 2017. 1, 2

[42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2015. 2

[43] Matías Tailanián, Álvaro Pardo, and Pablo Musé. U-flow: A u-shaped normalizing flow for anomaly detection with unsupervised threshold. *CoRR*, abs/2211.12353, 2022. 3, 6

[44] Tang Tang, Wei-Han Kuo, Jauh-Hsiang Lan, Chien-Fang Ding, Hakiem Hsu, and Hong-Tsu Young. Anomaly detection neural network with dual auto-encoders gan and its industrial inspection applications. *Sensors (Basel, Switzerland)*, 20, 2020. 1

[45] Qian Wan, Liang Gao, Xinyu Li, and Long Wen. Unsupervised image anomaly detection and segmentation based on pretrained feature mapping. *IEEE Transactions on Industrial Informatics*, 19(3):2330–2339, 2023. 2

[46] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. In *British Machine Vision Conference*, 2021. 2

[47] Xinkuang Wang, Wenjing Li, and Zhongcheng Wu. Cardd: A new dataset for vision-based car damage detection. *IEEE Transactions on Intelligent Transportation Systems*, 24(7):7202–7214, 2023. 3

[48] Yawen Wu, Zhepeng Wang, Dewen Zeng, Yiyu Shi, and Jingtong Hu. Synthetic data can also teach: Synthesizing effective data for unsupervised visual representation learning. In *AAAI Conference on Artificial Intelligence*, 2022. 2

[49] Jie Yang, Yong Shi, and Zhiquan Qi. DFR: deep feature reconstruction for unsupervised anomaly segmentation. *CoRR*, abs/2012.07122, 2020. 1

[50] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3973–3981, 2015. 3

[51] Fei Ye, Chaoqin Huang, Jinkun Cao, Maosen Li, Ya Zhang, and Cewu Lu. Attribute restoration framework for anomaly detection. *IEEE Transactions on Multimedia*, 24:116–127, 2019. 1

[52] 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. *CoRR*, abs/2111.07677, 2021. 3, 6

[53] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Dræm - A discriminatively trained reconstruction embedding for surface anomaly detection. *CoRR*, abs/2108.07610, 2021. 1, 2, 3, 6

[54] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. DSR - A dual subspace re-projection network for surface anomaly detection. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pages 539–554. Springer, 2022. 3, 6

[55] David Zimmerer, Peter M. Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, Bjørn Sand Jensen, Alison Q. O'Neil, Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, Bernhard Kainz, Nina Shvetsova, Irina Fedulova, Dmitry V. Dylov, Baolun Yu, Jianyang Zhai, Jingtao Hu, Runxuan Si, Sihang Zhou, Siqi Wang, Xinyang Li, Xuerun Chen, Yang Zhao, Sergio Naval Marimont, Giacomo Tarroni, Victor Saase, Lena Maier-Hein, and Klaus Maier-Hein. Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging*, 41(10):2728–2738, 2022. 1

A. Visual Results

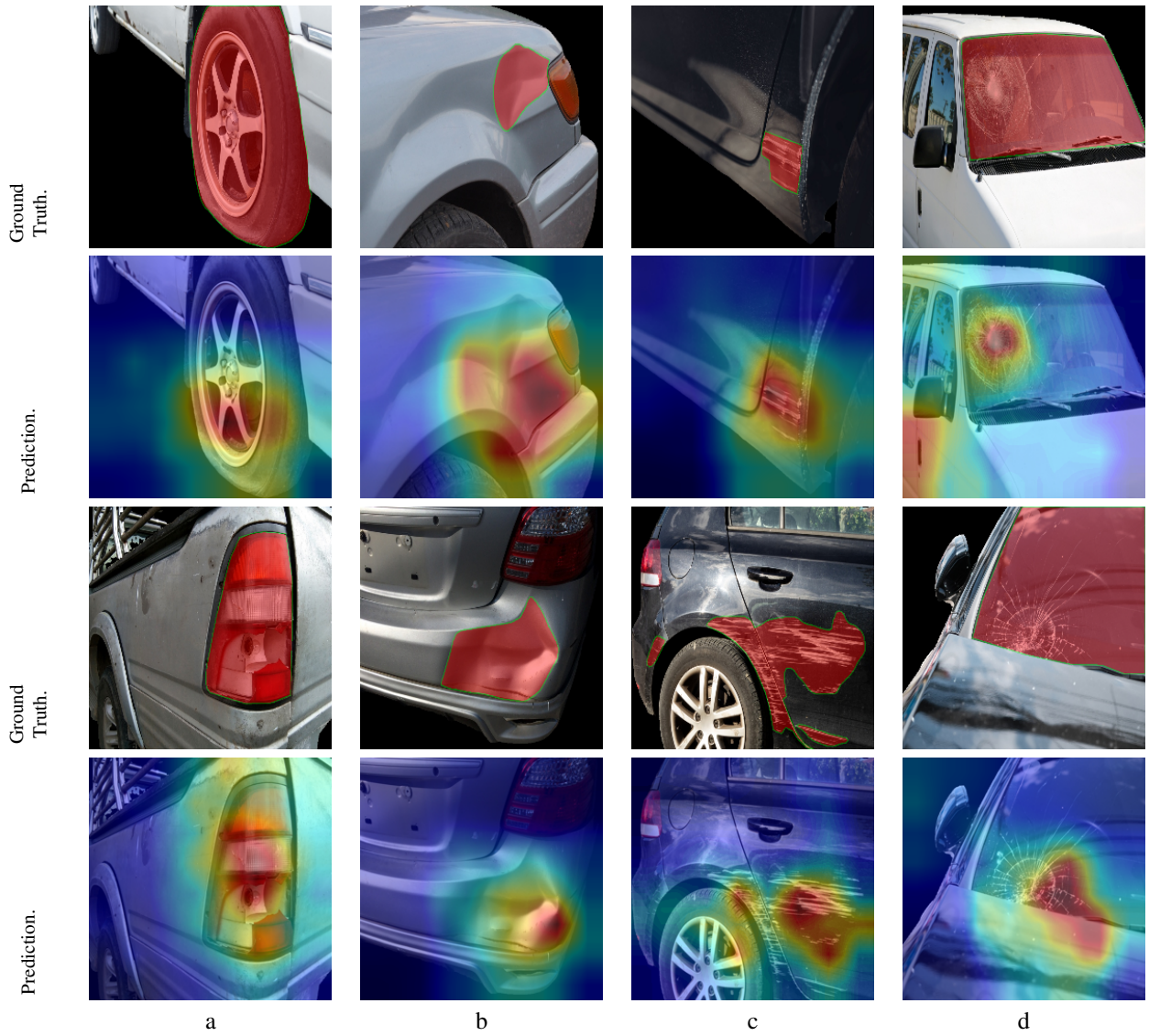

Figure 7. CsFlow results visualization for the predictions done on a subset of the test set. Each column corresponds to a different type of defect: (a) ”broken_part”, (b) ”dent”, (c) ”scratch”, and column (d) ”wind_screen”