Introducing DIAD: A Novel Metric for Assessing the Difficulty of Anomaly Detection Problems

Jure Pahor[1], **Danijel Skočaj**[1]

[1] *Faculty of Computer and Information Science, University of Ljubljana*
E-mail: jp7316@student.uni-lj.si

Abstract

*Assessing the complexity of anomaly detection tasks is essential for benchmarking datasets and models as well as for understanding the problem domains. While numerous anomaly detection methods have been developed, there remains a need for a simple, learning-free metric to estimate the difficulty of a given anomaly detection task. In this paper, we introduce DIAD (**D**ifficulty **I**ndex for **A**nomaly **D**etection), a lightweight metric designed to quantify task difficulty without requiring model training or inference. DIAD builds upon and extends the recently proposed AD3 metric by incorporating both the saliency of anomalies and the heterogeneity of normal appearance across the dataset. We evaluate DIAD on five widely used visual anomaly detection datasets and compare its scores with the observed performance of three state-of-the-art detection models. Results show that DIAD correlates more consistently with model performance than AD3, offering a practical and interpretable tool for assessing the complexity of anomaly detection problems.*

1 Introduction

Visual anomaly detection is a fundamental task in computer vision that involves identifying patterns that deviate from expected visual characteristics. The primary challenge lies in distinguishing subtle irregularities from normal data, particularly under diverse and complex visual conditions.

While a wide range of anomaly detection methods are proposed each year, new and diverse problem domains continue to emerge. This raises the need for a simple, learning-free metric that can estimate the inherent difficulty of an anomaly detection problem—without relying on the execution of computationally expensive models.

To address this, the AD3 metric [3] was recently proposed as a quick method for assessing surface anomaly detection difficulty based on anomaly saliency. It considers two key visual factors: anomaly size, as larger anomalies are typically easier to detect, and appearance similarity, since anomalies that closely resemble the object's surface are naturally harder to identify.

However, AD3 operates at the level of individual images and does not account for broader dataset-level characteristics, such as the variability of normal appearance. We hypothesise that anomaly detection difficulty is also determined by how homogeneous the normal class is - across the dataset: the more heterogeneous the normal appearance, the harder it is to model and, in turn, to detect deviations from it.

In this paper, we introduce DIAD (Difficulty Index for Anomaly Detection), a metric that builds on both insights: anomalies are easier to detect when they are salient and occur on visually uniform surfaces, and conversely, non-salient anomalies in heterogeneous contexts pose a greater challenge.

To demonstrate the practicality and interpretability of DIAD, we apply it to several widely used anomaly detection datasets [3, 4, 2, 8, 7]. Our experiments show that DIAD correlates well with the observed performance of state-of-the-art anomaly detection models [1, 5, 6], serving as a lightweight and effective proxy for estimating task difficulty.

2 Difficulty Index for Anomaly Detection

We build our work on the AD3 score [3]. The AD3 score aims to estimate the salience of anomalies—that is, how easily they can be detected. It considers two key factors: the relative size of the anomalies and the appearance difference between anomalies and the normal object. In our work, we slightly modify this metric and extend it to also account for the heterogeneity of normal appearance, which can significantly affect detection difficulty.

2.1 Saliency of anomalies

Our adjusted version of the AD3 metric uses the same two components as the original [3] but slightly altered, to better model the problem. The relative area of defect A_r is described by

$$A_r = \frac{A_d}{A_i}, \tag{1}$$

where A_d represents the number of defective pixels and A_i the total number of pixels in the image.

The appearance difference between anomalies and normal regions is quantified using the Similarity of Histograms (SiH) of the corresponding areas. Unlike the original formulation in [3], we derive the colour reference solely from the immediate neighbourhood of the defect, as this region has a significantly stronger influence on the anomaly's saliency than the rest of the object. We define

the neighbourhood by applying two iterations of morphological dilation using an elliptical kernel, where the kernel's width and height are set to 5% of the image's width and height, respectively.

The adjusted Similarity of Histograms SiH^* is calculated channel-wise, and then averaged over all three channels:

$$\mathrm{SiH}^* = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{A_{d\cup n}}\sum_{b=1}^{B}\left|h_{c,d}(b) - h_{c,n}(b)\right|, \quad (2)$$

where C is the number of channels, B is the number of bins, $h_{c,d}$ is the colour histogram of the defective area and $h_{c,n}$ is the histogram of its neighbourhood. Note that $d \cap n = \emptyset$. The value is normalised with the total number of pixels in the anomaly and in its neighbourhood $A_{d\cup n}$.

Using A_r and SiH^*, we compute our proposed metric, similarly as in [3], as the harmonic mean of the two, highlighting the equal importance of both factors. To obtain the average anomaly saliency for a given object class, we average the resulting values across all N_{defect} defective samples:

$$\mathrm{AD3}^* = \frac{1}{N_{\mathrm{defect}}}\sum_{k=1}^{N_{\mathrm{defect}}} 2\cdot\frac{\mathrm{SiH}^*\cdot\sqrt{A_r}}{\mathrm{SiH}^* + \sqrt{A_r}}. \quad (3)$$

$\mathrm{AD3}^*$ now represents the average anomaly saliency for the whole object category, i.e. the problem domain.

2.2 Heterogeneity of normality

To quantify both the heterogeneity of the object's normal appearance as well as the variability of its rotations and positions within the dataset, we utilise mean squared error (MSE) between the current image and the mean image of the object $\overline{I} = \frac{1}{N_{\mathrm{train}}}\sum_{i=1}^{N_{\mathrm{train}}} I_i$:

$$\mathrm{MSE}_i = \frac{1}{CWH}\sum_{c,x,y}\left(p_{c,x,y}(I) - p_{c,x,y}(\overline{I})\right)^2, \quad (4)$$

where W, H are image dimensions and $p_{c,x,y}$ is the pixel value at position (x, y) for channel c.

We can now compute the average mean squared error for the chosen object:

$$\mathrm{MSE} = \frac{1}{N_{\mathrm{train}}}\sum_{i=1}^{N_{\mathrm{train}}}\mathrm{MSE}_i, \quad (5)$$

where N_{train} is the number of training samples.

When comparing different datasets, it is useful to constrain the scores to the interval $[0, 1]$. To achieve this, we first standardise the input score x and then apply the *sigmoid function*:

$$\sigma : \mathbb{R} \to (0,1), \quad \sigma(x,\mu,\sigma) = \frac{1}{1 + e^{-\frac{x-\mu}{\sigma}}}, \quad (6)$$

where μ and σ are the mean value and the standard deviation of the target distribution. This transformation ensures that the resulting values are comparable across datasets, regardless of their original scale or distribution.

Using (6), we map the values of MSE to the interval between 0 and 1 and get a representative score of Heterogeneity of Normality - *HtgN*, as defined by the following equation:

$$\mathrm{HtgN} = \sigma(\mathrm{MSE}, \mu_{\mathrm{MSE}}, \sigma_{\mathrm{MSE}}). \quad (7)$$

Since a larger number of training images typically leads to a better representation and understanding of the problem, it inherently reduces the impact of the *heterogeneity of normality* and the dataset's variability on anomaly detection difficulty. This is because such variability becomes well represented within the training set. To account for this, we introduce the *unpredictability coefficient* u, defined as:

$$u = \sigma(-N_{\mathrm{train}}, -\mu_{N_{\mathrm{train}}}, \sigma_{N_{\mathrm{train}}}), \quad (8)$$

and compute the raw difficulty score as:

$$\mathrm{DIAD}^* = \frac{\mathrm{HtgN}^u}{\mathrm{AD3}^*}, \quad (9)$$

where u reduces the influence of HtgN when the number of training samples is sufficiently large, effectively pulling it closer to 1.

Finally, to ensure that the final difficulty score is in the range $[0, 1]$, we apply the sigmoid function (6):

$$\mathrm{DIAD} = \sigma(\mathrm{DIAD}^*, \mu_{\mathrm{DIAD}^*}, \sigma_{\mathrm{DIAD}^*}). \quad (10)$$

The equations (7), (8), and (10) require corresponding mean and standard deviation parameters. We estimate these parameters by computing the approximate mean and standard deviation values across five different datasets [3, 4, 2, 7, 8], which together provide a sufficiently large and diverse set of data points for robust modelling of the problem. The values used are defined in Table 1.

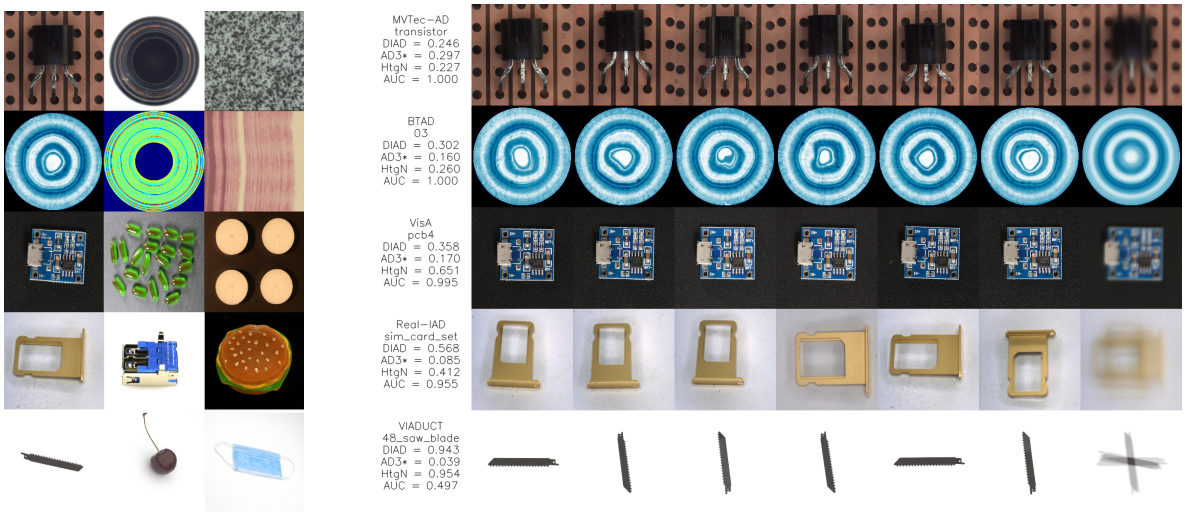
Table 1: Mean and standard deviation values used for normalisation.

$\mu_{\mathrm{MSE}} = 2000$	$\sigma_{\mathrm{MSE}} = 1300$
$\mu_{N_{\mathrm{train}}} = 600$	$\sigma_{N_{\mathrm{train}}} = 900$
$\mu_{\mathrm{DIAD}^*} = 8$	$\sigma_{\mathrm{DIAD}^*} = 6$

3 Experiments

3.1 Datasets

In our work we used five widely used anomaly detection datasets: MVTec-AD [2], a benchmark dataset commonly employed in anomaly detection, consisting of 15 object categories; VisA [8], another frequently used anomaly detection dataset containing images from 12 object categories, including some with multiple objects per image; BTAD, featuring three object categories; and two recently introduced multi-view datasets - meaning a single object is observed from multiple cameras, VIADUCT [3] with 49 object categories and Real-IAD [7] with 15. Examples of three objects from each dataset are shown in Figure 1, as well as the intra-class variability present



(a) (b)

Figure 1: (a) Images from 3 selected object categories for each dataset. (b) Description consisting of dataset name, object name, DIAD, AD3*, HtgN and detection AUC for PatchCore, and 6 random training images along with the mean image $\overline{I}$.

within a single object category for selected examples. By selecting this diverse set of datasets, we cover all major publicly available anomaly detection benchmarks and ensure broad representation of the problem distribution across different domains.

We used these five datasets both for estimating the general anomaly detection parameters, as described in the previous section, and for assessing the difficulty of these datasets as well as for evaluating the effectiveness of the proposed DIAD score.

3.2 Methods

We expect the difficulty score to negatively correlate with the performance of anomaly detection methods on the same dataset, as more difficult problems are inherently harder to solve. To validate this assumption, we selected three state-of-the-art anomaly detection models [1, 5, 6], each representing one of the main paradigms commonly found in the literature.

Reconstruction methods learn to reproduce only normal images—since anomalies (absent from the training set) are reconstructed poorly, comparing the input to its reconstruction reveals anomalies. For this approach, we have chosen *EfficientAD* [1].

Building on this, *discriminative methods* rely on the idea that anomalies can be simulated. In practice, anomalies are artificially injected into otherwise normal training images, and the model learns to identify these synthetic anomalies. For this approach, we used *SuperSimpleNet* [5].

Embedding-based methods rely on similarity embeddings: a deep convolutional neural network is used to extract feature vectors. The key idea is that features from normal samples form tight clusters in the embedding space, whereas those representing anomalies lie far apart. For this approach, we employed *PatchCore* [6].

3.3 Results

To obtain the results, we used the anomaly detection methods mentioned above [1, 5, 6] with an input size of 256×256. For each object in each dataset [3, 4, 2, 8, 7], we first computed the DIAD score, and then the image-level AUC (AUC Image) as the detection performance metric for all three anomaly detection models. For the two multi-view datasets [3, 7], the experiments were conducted by treating all images—regardless of viewpoint—as part of a single dataset.

Five individual results—one object per dataset—are shown in Figure 1. The figure depicts the calculated values of DIAD and AD3*, alongside the performance of the PatchCore method. We observe that performance, measured as Image AUC, tends to decrease as DIAD increases, i.e., when the corresponding object dataset is deemed more difficult for anomaly detection.

More comprehensive results, covering all object datasets, are presented in Figure 2. The figure compares the computed AD3* and DIAD scores with the performance of the PatchCore method [6]. In both cases, we observe a moderate correlation between the dataset difficulty score and the performance of the detection model. The correlation is positive for AD3*, as it reflects the saliency of anomalies — more salient (i.e., easier-to-detect) anomalies lead to higher model performance. In contrast, the correlation is negative for DIAD, since it explicitly measures detection difficulty — higher DIAD scores indicate harder problems and thus lower expected performance. Notably, the correlation is slightly stronger in the case of DIAD. The results indicate that it is indeed possible to estimate the problem's difficulty without running compute-heavy models.

It is also worth noting that all values of AD3* are clustered within a narrow range and consistent with the findings reported in the original paper [3]. In contrast, our proposed metric DIAD spans most of the interval

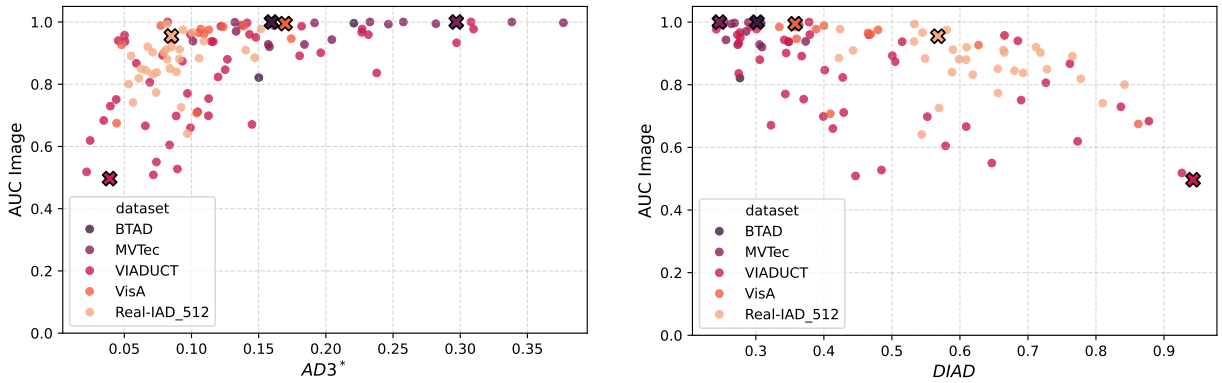


Figure 2: Scatter plot comparison of AD3* and DIAD scores with detection AUC for PatchCore [6]. Each point represents the average score for a (dataset, object) pair. Points marked by **x** correspond to the objects shown in Fig. 1.

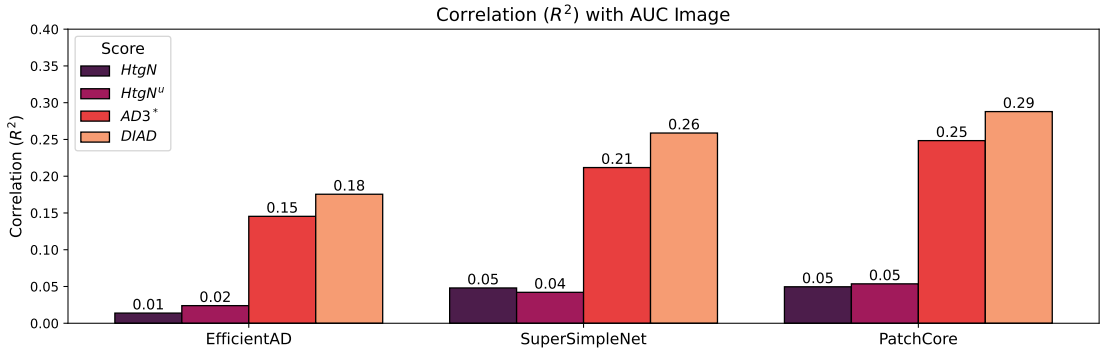


Figure 3: The correlation between HtgN, HtgN^u, AD3*, and DIAD and the detection performance (AUC Image).

from 0 to 1, providing better insight into the problem's difficulty, as its values are more interpretable and spread across a wider range. This interpretability is important: if the DIAD value is close to 1, we can confidently say that the anomaly detection task is very challenging.

Figure 3 summarizes the results and shows the correlation for all three evaluated methods. The correlation is computed as R^2, the squared Pearson coefficient, across all data points, each represented as a (dataset, object) pair. It demonstrates that the proposed DIAD score exhibits a stronger (negative) correlation with detection performance than using AD3* alone [3], regardless of which anomaly detection method is used. Furthermore, Table 2 shows that, when metrics are evaluated per dataset, the proposed DIAD score generally outperforms AD3*[3]. The only notable exception is the BTAD dataset [4], which contains only three object categories. In such a small dataset, even a single outlier can significantly skew the correlation coefficient, which likely explains DIAD's weaker performance in this case.

4 Conclusion

We have presented a new approach for estimating the difficulty of surface anomaly detection problems that considers not only the saliency of anomalies, but also the variability of training images and the heterogeneity of normal object appearance. Our proposed metric, DIAD, builds on and extends the AD3 framework by incorporating dataset-level characteristics that affect anomaly detection performance.

Dataset / Model	*EfficientAD*		*SuperSimpleNet*		*PatchCore*	
	AD3*↑	DIAD↓	AD3*↑	DIAD↓	AD3*↑	DIAD↓
BTAD	**0.60**	0.98	**0.60**	0.98	**0.59**	0.98
MVTec-AD	0.05	**–0.30**	0.43	**–0.60**	0.61	**–0.64**
Real-IAD	0.25	**–0.29**	0.36	**–0.45**	0.32	**–0.35**
VIADUCT	0.38	**–0.43**	0.46	**–0.55**	0.53	**–0.58**
VisA	**0.08**	–0.02	0.48	**–0.61**	0.48	**–0.65**

Table 2: Pearson's correlation coefficient between AD3* (higher is better) and DIAD (lower is better) to the detection AUC. The coefficient for the better score is bold - if the DIAD score is negative and its absolute value is higher than AD3 score, it is considered better.

Through extensive evaluation on five publicly available datasets, we have shown that DIAD correlates more reliably with model performance than AD3 across multiple detection methods. This makes DIAD a practical and interpretable proxy for task difficulty, suitable for benchmarking new datasets or guiding method selection without requiring training or inference.

In future work, DIAD could be further refined to support other types of anomaly detection, and incorporated into automated dataset analysis pipelines, helping practitioners better understand and design anomaly detection problems.

Acknowledgements

This research was partially funded by the ARIS MUXAD research project (grant number J2-60055), the research program P2-0214, and supported by the SLING supercomputing infrastructure (ARNES, EuroHPC Vega - IZUM).

References

[1] Kilian Batzner, Lars Heckler, and Rebecca Konig. " EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies ". In: *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, Jan. 2024, pp. 127–137. DOI: 10.1109/WACV57701.2024.00020. URL: https://doi.ieeecomputersociety.org/10.1109/WACV57701.2024.00020.

[2] Paul Bergmann et al. "MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 9584–9592. DOI: 10.1109/CVPR.2019.00982.

[3] Jan Lehr et al. "AD3: Introducing a Score for Anomaly Detection Dataset Difficulty Assessment Using VIADUCT Dataset". In: *Computer Vision – ECCV 2024*. Ed. by Aleš Leonardis et al. Cham: Springer Nature Switzerland, 2025, pp. 449–464. ISBN: 978-3-031-73113-6.

[4] Pankaj Mishra et al. "VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization". In: *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*. 2021, pp. 01–06. DOI: 10.1109/ISIE45552.2021.9576231.

[5] Blaž Rolih, Matic Fučka, and Danijel Skočaj. "SuperSimpleNet: Unifying Unsupervised and Supervised Learning for Fast and Reliable Surface Defect Detection". In: *Pattern Recognition*. Ed. by Apostolos Antonacopoulos et al. Cham: Springer Nature Switzerland, 2025, pp. 47–65. ISBN: 978-3-031-78192-6.

[6] Karsten Roth et al. " Towards Total Recall in Industrial Anomaly Detection ". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2022, pp. 14298–14308. DOI: 10.1109/CVPR52688.2022.01392. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01392.

[7] Chengjie Wang et al. "Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 22883–22892.

[8] Yang Zou et al. "SPot-the-Difference Self-supervised Pre-training for Anomaly Detection and Segmentation". In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Cham: Springer Nature Switzerland, 2022, pp. 392–408. ISBN: 978-3-031-20056-4.