

From Lab to Factory: Pitfalls and Guidelines for Self-/Unsupervised Defect Detection on Low-Quality Industrial Images

Sebastian Hönel¹ (✉)^[0000–0001–7937–1645] and Jonas Nordqvist¹^[0000–0002–0510–6782]

Linnæus University, Växjö, Sweden {sebastian.honel,jonas.nordqvist}@lnu.se

Abstract. The detection and localization of quality-related problems in industrially mass-produced products has historically relied on manual inspection, which is costly and error-prone. Machine learning has the potential to replace manual handling. As such, the desire is to facilitate an unsupervised (or self-supervised) approach, as it is often impossible to specify all conceivable defects ahead of time. A plethora of prior works have demonstrated the aptitude of common reconstruction-, embedding-, and synthesis-based methods in laboratory settings. However, in practice, we observe that most methods do not handle low data quality well or exude low robustness in unfavorable, but typical real-world settings. For practitioners it may be very difficult to identify the actual underlying problem when such methods underperform. Worse, often-reported metrics (e.g., AUROC) are rarely suitable in practice and may give misleading results. In our setting, we attempt to identify subtle anomalies on the surface of blasted forged metal parts, using rather low-quality RGB imagery only, which is a common industrial setting. We specifically evaluate two types of state-of-the-art models that allow us to identify and improve quality issues in production data, without having to obtain new data. Our contribution is to provide guardrails for practitioners that allow them to identify problems related to, e.g., (lack of) robustness or invariance, in either the chosen model or the data reliably in similar scenarios. Furthermore, we exemplify common pitfalls in and shortcomings of likelihood-based approaches and outline a framework for proper empirical risk estimation that is more suitable for real-world scenarios.

Keywords: Unsupervised Anomaly Detection · Anomaly Localization · Normalizing Flows · Industrial Visual Inspection · Low-Quality Imagery · Robust Evaluation · Out-of-Distribution Detection · Computer Vision in Manufacturing

1 Introduction

Industrial manufacturing requires quality assurance. A common approach to quality evaluation is through manual visual inspection, which is expensive, repetitive, and error-prone [10]. Recent advances in deep learning and computer vision have boosted interest in automated self- or unsupervised anomaly detection

methods [5]. In such approaches, acquiring nominal (defect-free) data is affordable, but exhaustively defining abnormal defect variants often proves difficult or impossible. Thus, anomaly detection (AD) is framed as an out-of-distribution challenge, distinguishing nominal samples from those outside the known distribution [24]. A related task, anomaly localization (AL), further identifies where anomalies occur, benefiting interpretability by providing visual cues and enabling human-in-the-loop systems [30].

AD/AL methods in visual inspection generally fall into three categories: embedding-, reconstruction-, and synthesis-based approaches [6]. Embedding-based approaches use pretrained feature extractors. Reconstruction approaches exploit differences between input and reconstructed images, or encodings and decodings thereof, to detect anomalies [32]. Synthesis methods train classifiers by generating artificial defects to differentiate them from nominal samples, effectively combining unsupervised and semi-supervised approaches [33]. Semi-supervised approaches are now generally preferred. Unsupervised methods often lack reliability in precisely detecting out-of-distribution cases, whereas fully-supervised methods require costly data labeling, suffer limited data availability, and cannot easily handle unknown defect types or label noise [30].

Visual-inspection AD methods are commonly evaluated on standard datasets such as MVTec AD [1] or Magnetic Tile Defects (MTD) [13]. However, these benchmark datasets largely come from controlled laboratory environments. Hence, most images feature simple setups, single centered objects with uniform backgrounds, static orientations and camera distances, and lack realistic disturbances such as reflections, shadows, or blur. Recently, Jezek et al. [16] introduced the Metal Parts Defect Detection (MPDD) dataset, a more challenging real-world scenario encompassing some typical practical issues. Their evaluation demonstrates a significant deterioration of state-of-the-art performance, especially for image-level AD, highlighting the gap between lab-based benchmarks and real production environments.

In view of this gap, this paper explores practical problems related to data quality, modeling choices, and proper evaluation. Important data-related challenges include excessively subtle anomalies relative to natural variance, unwanted background disturbances, reflections, and other unfavorable image-capturing conditions. Additionally, certain conditions can impact models differently. While introducing rotations or varying viewpoints can broaden data diversity and generalization, the effectiveness of these augmentations strongly depends on the model chosen. Some models handle image variations like object scale or position robustly, while others decline significantly in performance if these transformations occur among the objects themselves.

Regarding performance evaluation, anomaly detection results frequently use area under the receiver operating characteristic curve (AUROC). Although AUROC is valuable when comparing methods within identical datasets, interpreting reported AUROC values across different publications or scenarios should be done with caution. Its interpretation strongly depends on the underlying data and evaluation strategies. Often only maximum or average AUROC scores

are reported, lacking uncompromising replication. Importantly, classifiers built from selected thresholds for anomaly scores are rarely evaluated rigorously using proper repeated cross-validation or bootstrapping methods that provide reliable estimates of empirical risk and realistic performance in deployment scenarios.

Our contributions are summarized as follows:

1. We overview critical data-acquisition and data-quality issues encountered in practice, and suggest directions for improvement.
2. We investigate how core architectural components in current AD/AL methods respond differently to specific real-world issues, offering practical recommendations for properly pairing datasets with models.
3. We propose an evaluation framework based on adequate outer resampling to accurately estimate empirical model risk and derive confidence intervals, better predicting factory-floor model performance.

The remainder of this paper is structured as follows. In Section 2, we present the relevant background, related work, and describe the concrete dataset and problem we attempt solving. In Section 3, we elucidate limitations of flow-based architectures. Section 4 is dedicated to outlining and validating a robust estimation framework. Lastly, Section 5 offers a discussion and an outlook on future work.

2 Background and Related Work

Anomaly detection in imagery is a critical task in various domains, including industrial inspection, medical imaging, and security, where identifying irregularities is essential. Due to the rarity and diversity of anomalies, unsupervised and semi-supervised learning approaches are commonly employed, as they enable detection without the need for extensively labeled (complementary) datasets.

Unsupervised methods typically identify anomalies by modeling the distribution of normal samples. One widely used approach involves reconstruction-based techniques, such as autoencoders [12], where the model is trained to reconstruct normal data, and anomalies are inferred from high reconstruction errors [28]. Another method leverages generative adversarial networks (GANs) [8], assessing deviations from the learned distribution to detect anomalies [19]. A third strategy involves normalizing flows, which estimate the likelihood of feature maps and classify low-likelihood regions as anomalous [25,26], where the latter case is further discussed in Section 3.1. Additionally, clustering techniques such as k -means and DBSCAN can be utilized, where data points identified as outliers are considered anomalies.

In the context of self-supervised learning, techniques such as synthetic anomaly generation or auxiliary tasks have been explored to enhance feature representation learning, thereby improving anomaly detection performance (e.g., [11]). The current state-of-the-art in the realm of AD is a method which utilizes several techniques in AD. This method is further described in Section 3.2.

2.1 Normalizing Flows

Let $d \geq 1$ be an integer, and consider a vector $\mathbf{x} \in \mathbb{R}^d$. A normalizing flow models the distribution of \mathbf{x} using a diffeomorphic transformation T , which is bijective and smooth along with its inverse. The transformation is applied to a latent variable \mathbf{z} drawn from a base distribution $p_{\mathbf{z}}(\mathbf{z})$. Given $\mathbf{x} = T(\mathbf{z})$, the density of \mathbf{x} follows from the change-of-variables formula

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{z}}(\mathbf{z}) |\det J_T(\mathbf{z})|^{-1}, \quad \text{where } \mathbf{z} = T^{-1}(\mathbf{x})$$

and $J_T(\mathbf{z})$ is the Jacobian of T . In practice, T is parameterized, and in the context of normalizing flows, the set of parameters θ for a transformation T is learned by a so-called *conditioner*, typically implemented by some fully-connected neural network. Training a normalizing flow involves minimizing the forward Kullback–Leibler divergence between the data distribution $p_{\mathbf{x}}(\mathbf{x})$ and the transformed base distribution $\hat{p}_{\mathbf{x}}(\mathbf{x}; \theta)$. The loss function is:

$$\begin{aligned} \mathcal{L}(\theta) &= D_{\text{KL}}(p_{\mathbf{x}}(\mathbf{x}) \parallel \hat{p}_{\mathbf{x}}(\mathbf{x}; \theta)) = -\mathbb{E}_{p_{\mathbf{x}}(\mathbf{x})} [\log(\hat{p}_{\mathbf{x}}(\mathbf{x}; \theta))] + C \\ &= -\mathbb{E}_{p_{\mathbf{x}}(\mathbf{x})} [\log(p_{\mathbf{z}}(T^{-1}(\mathbf{x}; \theta))) + \log|\det J_{T^{-1}}(\mathbf{x})|] + C, \end{aligned}$$

where C is a constant. This may be estimated via Monte Carlo sampling. For further details, see, e.g., [18].

2.2 Some Problems of Purely Likelihood-Based Approaches of AD

Likelihood-based deep generative models have been widely used in anomaly detection. However, there are a number of critical limitations which makes these prone to failure. One of these problems is that the model may assign higher likelihood to data which may be semantically different although still normal in some sense. For instance, models trained on Cifar-10 have been shown to assign higher likelihoods to images from SVHN. One explanation to this problem has been proposed in [4], where the argument is that there is a dependence on the entropy of the dataset, which gives rise to unwanted behavior. Another problem is that likelihood-based models struggle in high-dimensional spaces where background statistics dominate the likelihood estimation. However, to circumvent these issues it is suggested in [21] to use likelihood ratios instead of crude likelihoods to handle the background statistics.

2.3 Our Problem and Dataset

The dataset consists of medium-resolution images of blasted forged steel coupling links (“G-links”) captured using an area-scan camera as the coupling link was translated and rotated by an industrial robot [14]. This means that each individual piece is captured by roughly 120 images over different angles. Damaged pieces were collected using the regular manual inspection process in the production environment over an extended time frame. In the training set, there

are 3,424 images of non-defect coupling links for training. The validation set is comprised of 568 images, 165 of which contain a defect. Figure 1 shows four defective samples with surface dents, deeper deformations, and scratches. These kind of anomalies arise naturally in production. Defects can be comparatively subtle and typically exhibit less variance than the embossing on the items. Furthermore, the links and the background both reflect the camera’s flash.

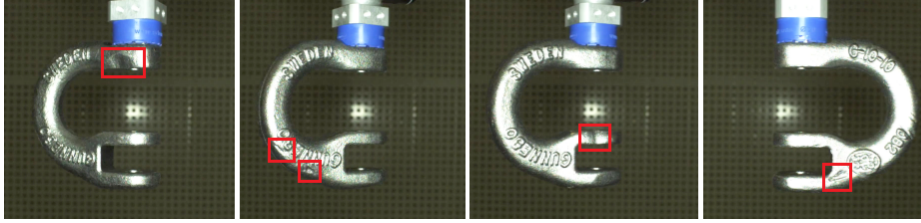


Fig. 1: Four defective examples from the G-link dataset. Anomalies are coarsely marked using a red bounding box (zoom for more details).

3 Discovery of Limitations through Normalizing Flows

When our project started in early 2023, one of the state of the art methods in unsupervised anomaly detection and -localization was a method called “*Fully Convolutional Cross-Scale-Flows for Image-based Defect Detection*”, here and after called “CS-Flow” [26]. Over the course of approx. 1.5 years, we would attempt to facilitate it to solve our own problem (see Section 2.3). In mid-2024, CS-Flow was superseded by a new state of the art method called “*A Unified Anomaly Synthesis Strategy with Gradient Ascent for Industrial Anomaly Detection and Localization*”, or, as its authors abbreviate it, “GLASS” [6]. As of writing this, GLASS is still undefeated.

In this section, we outline our journey of attempting to facilitate CS-Flow and adapting it to our problem. This journey is characterized by an iterative and hypothesis-/evidence-based approach. Our endeavors allowed us to unveil the intricate and inconspicuous issues related to data quality, choice of model, and evaluation approach. For full disclosure, comprehensibility, and reproducibility, each step is associated with a notebook in the replication package [15]. The remainder of this section is dedicated to the larger steps in this process. All experiments were run on an NVIDIA 8xH100/80GB machine.

3.1 Previous SotA: Convolutional Cross-Scale Normalizing Flow

CS-Flow [26] is a technique that processes image embeddings through a specialized normalizing flow to compute likelihoods of feature maps, which are then used to distinguish anomalies from normal samples.

The method begins by resizing each image into three different scales and passing them through a frozen feature extractor, EfficientNet-B5 [29], to generate three corresponding feature maps representing the image at multiple resolutions. A key component of CS-Flow is a unique coupling block that captures interactions between feature maps at different scales. These coupling blocks are stacked to form a normalizing flow, whose output is evaluated against a multivariate standard normal distribution with diagonal covariance matrix. The resulting likelihood determines whether an image is anomalous, based on a threshold τ .

In the paper, CS-Flow is evaluated on the MVTec AD dataset, achieving an average AUROC of 98.7. It reports state-of-the-art performance in eleven out of 15 categories. Furthermore, their approach processes feature maps in a fully convolutional (equivariant) manner, thus retaining positional information, enabling the identification of specific regions of anomalies. To achieve this, an anomaly score is assigned to each local position (i, j) of the feature map by aggregating values along the channel dimension using the L_2 -norm. Regions with high norm values in the output feature tensors indicate potential anomalies.

3.2 Current SotA: Global and Local Anomaly co-Synthesis Strategy

GLASS is a three-way discriminator (*i.e.*, not a probabilistic model) that draws from all three branches of AD: embedding, reconstruction, and especially synthesis [6]. It uses a frozen feature extractor, as well as a feature adaptor, which is a fully-connected network that learns a useful representation of those features. It requires foreground masks of the objects of interest in order to synthesize so-called local anomalies. For a local anomaly, GLASS blends the foreground mask with a randomly generated mask first. Then, using a texture from the Describable Textures Dataset [7], the resulting patches are overlayed, thereby simulating an anomaly. This results in pixel-accurate augmentations, making the entire method self-supervised. Another type of synthetic anomaly that is generated are so-called global anomalies. For each batch, GLASS takes the derivative of itself with respect to the nominal sample. Then, Gaussian noise is added, and the sample is changed using truncated projection of the gradient (*i.e.*, the sample is changed such that its loss under the current model increases). Finally, the loss is computed along three branches: The discriminator shall predict all zeros for nominal samples, all ones for global anomalies (*i.e.*, anomalous everywhere), and all zeros except in the augmented regions for local anomalies.

Note that GLASS operates optionally under one of two distribution hypotheses: the hypersphere and manifold hypotheses. Under the first hypothesis, it is assumed that all nominal samples can be encompassed by a compact hypersphere and that all out-of-distribution samples have a distance from the center that is greater than that of any in-distribution sample [9]. Under the second hypothesis, it is assumed that all nominal samples lie within a lower-dimensional, locally linear manifold distribution. The local linearity and the fact that the manifold is homeomorphic to Euclidean space allows GLASS to define a distance under which a sample is considered to lie outside the in-distribution manifold [27]. By synthesizing global anomalies, a nominal sample is gradually worsened until it

is considered to be located outside the in-distribution manifold. Interestingly, our dataset of coupling links conforms to a *cyclic manifold*, because the objects themselves repeatedly assume previously-assumed angles [20]. This might also explain why we generally achieve better results by operating GLASS under the manifold hypothesis.

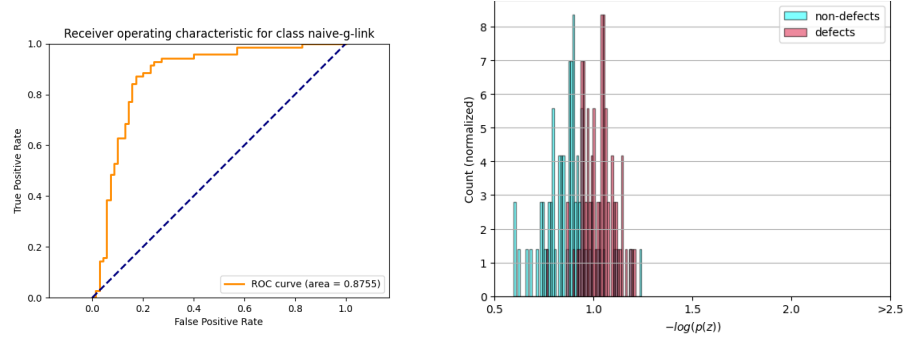


Fig. 2: ROC-curve (left) and anomaly scores (right) for the naive G-Link application under CS-Flow. Scores of nominal and anomalous samples overlap significantly.

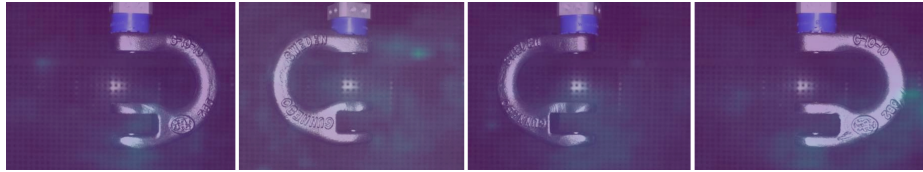


Fig. 3: Two defective samples (left) and two nominal samples (right). Anomaly scores of nominal samples sometimes greatly exceed those of actual defective samples (increased gamma and contrast). Often, high scores are assigned to the background.

3.3 Naïve Application to Raw Dataset

The first step was to attempt to use CS-Flow as an off-the-shelf solution and apply it to raw imagery of our dataset of forged coupling links. We have not made changes to the default configuration, which was used to achieve state-of-the-art results on the MVTec AD dataset. By default, images are resized to 768x768 pixels. Two smaller versions of the same image are used, one in 384x384 and one in 192x192, as CS-Flow facilitates a multi-scale approach that extracts and combines features from all three images per sample. The default architecture comprises four coupling blocks, applies gradient-clipping, and uses 1,024 units

in the fully-connected networks that are the conditioners to each coupling block. The training runs for 240 epochs by default, with an intermediate evaluation step every 60 epochs.

Table 1: Summary of AUROC scores across models and datasets. Note that the AUROC of the MVTec AD datasets are based on a rotated dataset as described in Section 3.5.

Model	Trained On	AUROC
CS-Flow	MVTec AD	0.631
	MVTec AD (rot.)	0.939
	Naïve G-Link	0.871
	Masked G-Link	0.802
	CE G-Link	0.867
GLASS	CE G-Link	0.941

The evaluation of the obtained results from this step promised a skillful model. Typically, the reported AUROC on the test set would be ≥ 0.87 and the plot of the anomaly scores (which are derived from the negative log-likelihood an image produces) would indicate a somewhat good separability of scores. An example histogram and ROC-curve are shown in Figure 2 and results in Table 1. The next evaluation step was to inspect the anomalous regions. Four typical examples are shown in Figure 3. It becomes apparent that most of the low-likelihood pixels are assigned to the background. The highlighted anomalous regions are not concordant with the actual location of the defect. Even worse, it would appear that the residual anomaly maps roughly cover the area where the coupling link is *usually* spatially present. What we mean by that is that most images in our dataset show the coupling link rotated (to a varying degree) along the fixture’s axis to one or the other side. One could argue that if we were to interpolate all images of our dataset, the result would perhaps show the silhouette of a somewhat “double-link”, which represents the average of it. Since a normalizing flow learns a complex probability distribution, it would approximate this average image, which is in disparity with any single image. This in turn allows us to conclude and hypothesize three things:

1. We hypothesize that strong object-variance can have detrimental effects when the underlying model is a deep density estimator.
2. The background should be blacked out in order to prohibit the model from assigning high (or any positive) anomaly score to it.
3. We have a Husky–Snow problem, likely caused by too-large spatial object variance.

The “Husky–Snow” problem, sometimes also referred to as “Clever–Hans” problem, is used to demonstrate how a model may learn unintended background features—in this case, a model distinguishing between wolves and huskies using

background snow presence—rather than the animal itself (or discriminative features thereof) as the distinguishing factor. More formally, these so-called local interpretable model-agnostic explanations were introduced by Ribeiro et al. [23]. What this means for us is that the model has not learned any useful representation of the underlying problem. It might have memorized specialized patterns or shortcuts from the training set without extracting broader features to achieve the relatively high AUROC. One may also consider the average image (silhouette) as intrinsic noise in the training set: Without sufficient capacity for generalization (or compensating invariances) in the model, it appears we are posing a contradictory problem in the first place. Recall that in a standard normalizing flow architecture, the available learning capacity is restricted to the conditioners, which learn the parameters θ , required by the transformation $T(\mathbf{x}; \theta)$. In other words, between the (frozen) feature extractor and the flow itself, CS-Flow does not currently offer some form of learnable embedding.

3.4 Segmentation and Background Removal

The primary consequence from our naïve application to the coupling links dataset is to remove (or to black out) the background from each image, including removal of the sometimes visible fixture atop. However, since we find ourselves in an inherent unsupervised setting, manual removal of the backgrounds for each image would be too costly. While not technically completely unsupervised, we used GroundedSAM [22] to segment the bulk ($\approx 98\%$) of our dataset. GroundedSAM is an open-vocabulary detection and segmentation model. It facilitates textual prompts to instruct SAM (Segment Anything, [17]) to detect and segment corresponding objects. While it took a handful of different prompts, this final prompt “*the curved metal part in the center of the image without the blue ring*” was used to segment the entire dataset. As a micro ablation, we segment images from the MPDD dataset [16] using similarly trivial prompts and on-par segmentation results¹.



Fig. 4: Examples of spurious anomaly maps of defect samples using the masked dataset of coupling links (increased gamma and contrast).

Although the performance of training a vanilla CS-Flow on the segmented data has declined, the typical AUROC scores in the range of $[0.79, 0.81]$ would still

¹ Among our contributions are two command-line applications that enable efficient mass-wise processing of images and their masks using a single prompt.

suggest a somewhat skillful model. Upon inspecting the identified anomalous regions, we realize that higher scores are still assigned to the now-black background in many cases, meaning a worse version of the Husky–Snow problem is still present. Since the model cannot learn any shortcuts through spurious features of the background any longer, this result explains the lower AUROC and presents evidence for our silhouette theory. In short, it appears that the varying angles of the object seem to be disadvantageous for the chosen model.

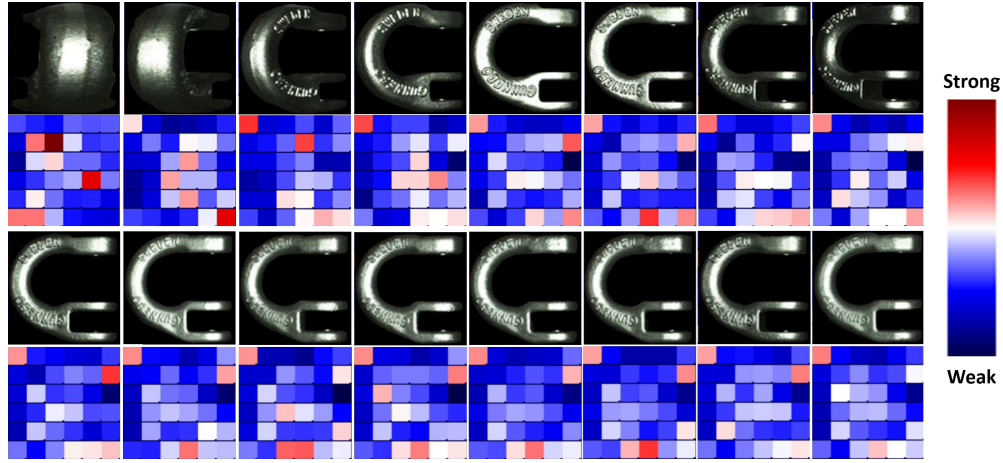


Fig. 5: A highly varying dataset causes an equivariantly varying feature map, too. Aggregated were the $6 \times 6 \times 304$ feature maps, each using $\max()$. The lowest low is the supremum of the maximum activations across all patches and across all images (*i.e.*, $\max()$ is the global maximum).

3.5 Controlled Rotation Experiment on MVTec AD

Following our naïve approach, we hypothesized that a pronounced lack of object invariance leads to detrimental performance. However, it remains unclear whether the same holds true for image invariance—particularly relevant for equivariant models. Convolutions inherently provide some approximate local translation invariance. Pooling further amplifies this, capturing only the maximum, average, or minimum in local regions. Thus, for limited translation and scale variance, equivariant models offer partial image invariance. To quantify translation invariance, we designed a two-fold controlled experiment using a vanilla CS-Flow model with the “cable” category from the MVTec AD dataset [1]. We first trained on unmodified nominal data containing three strands: a central yellow strand at the top and blue and brown strands in the left and right corners. For testing, we added rotated versions (90° , 180° , 270°) of these nominal images. In a second test, training images were rotated using the same angles,

while inference used original, as well as non-rotated images. One goal of this test was to find out to what degree CS-Flow is resilient to rotation variances. Ideally, a rotated nominal sample should not yield anomaly scores higher than typical nominal test images, indicating robust invariance. For the first experiment, the optimal threshold was ≈ 0.957 with AUROC ≈ 0.631 across all anomaly categories (e.g., bent wire, poked insulation). Rotated images had anomaly scores between ≈ 1.28 and ≈ 1.64 , resulting in all of the rotated images being wrongly classified as anomalous. In contrast, the second experiment achieved a significantly higher AUROC (≈ 0.936) and an optimal threshold of ≈ 1.251 . These two tests indicate that there is only very low robustness with regard to rotational invariance. Only one out of 58 non-rotated nominal samples was incorrectly classified (score ≈ 1.278). Thus, the second model generalized better, successfully handling rotated images. In summary, we demonstrated that rotation-based image invariances can be effectively handled when the training set explicitly includes such variance. We hypothesize that careful and limited artificially introduced image variance helps equivariant models generalize better. However, this approach may simultaneously increase model fragility through learning contradictory information—especially prominent when object invariance is introduced. Figure 5 visualizes activation magnitudes of latent features under an equivariant feature extractor (EfficientNet-B5, [29]). The first row shows rotated coupling links, revealing activations tightly localized around object regions. The second row depicts coupling links with only minimal viewpoint changes. While activation intensities vary slightly, the activation locations remain consistent as the object’s position changes minimally.

3.6 Restriction to Similar Angles

Since CS-Flow is a deliberate equivariant pipeline, we alter our dataset one more time, by limiting the imagery to those shot from a very similar, front-facing angle. The idea is, again, to reduce the natural variance in the data. However, that same restriction also applies to the test data, meaning that we alter and simplify our initial problem. Therefore, the results cannot (and should not) be compared to the others. However, this method achieves an AUROC of ≈ 0.909 . The qualitative evaluation shows that the anomaly maps often are closely located to the true region of the anomaly. However, having occasional spurious high anomaly scores assigned to the background is still a problem in this model.

4 Robust Dataset Preparation and Performance Estimation

We produce a segmented, centered, and padded version of our original dataset (“center-embedded” or “CE” short). In it, all objects are segmented and cropped to the minimum bounding box first. Then, we determine the minimum width and height that can encompass all coupling links, regardless of their spatial arrangement. We find that a resolution of 600x800 suffices. Finally, padding is

added to each image such that the object remains vertically and horizontally centered. This way, the coupling links do not rotate longer around the fixture’s axis, but rather around the image’s center. This allows us to effectively reduce natural variation while retaining all information.

4.1 Results using CS-Flow

We perform a last test using CS-Flow and this dataset, as previously, images were resized by it to a fixed resolution, which likely introduced more unwanted object variance, as narrow-appearing objects are perhaps stretched very wide. With an AUROC of ≈ 0.867 , the model does surprisingly not perform better than our naïve tests. Spurious anomalous regions pose still a problem. It would finally appear that the too-large natural object variance makes the flow model inherently fragile as it attempts to learn contradictory information depending on object position. Recall that vanilla Normalizing Flows, even with equivariantly extracted features, lack a distinct embedding space robust enough to cope with large object-level variations like rotation and position shifts.

4.2 Results using GLASS

Contrary to CS-Flow, GLASS is not a multi-scale model. Instead, it accepts images of a single resolution only. We identified a champion model that uses a lowered resolution of 160x160 that achieves an image-level AUROC of ≈ 0.941 . The qualitative inspection of the results shows strong agreement between true and predicted anomaly location. Similarly, anomaly scores are practically zero for nominal parts (see Figure 6). While sometimes the predicted anomaly scores are too high for some regions, GLASS never assigns any scores to the background, so it seems the silhouette-problem vanished. Also, judging by the qualitative results, it appears that the strong object-variance in our dataset is not problematic here, either. Interestingly though, GLASS also uses an equivariant (frozen) feature extractor, here WideResNet-50 by default. However, compared to Normalizing Flows, GLASS is a purpose-built model for AD, with two nonlinear high-capacity components, a feature adaptor (learning a suitable embedding and salient features), as well as a dense discriminator. We conclude that it is this architecture that allows the model to learn a more generalized representation of our problem.

4.3 Obtaining Unbiased Robust Performance Estimates

Once we are confident in our model and want to take it from the laboratory to the factory, we require robust estimates as to the usefulness for unsupervised industrial inspection. So far, we have only computed the AUROC. However, its usefulness lies in comparing models and/or datasets and is limited to the dataset it was determined on. In order to estimate other metrics, such as (balanced) accuracy, F1, precision and recall, etc., we have to consider the previously

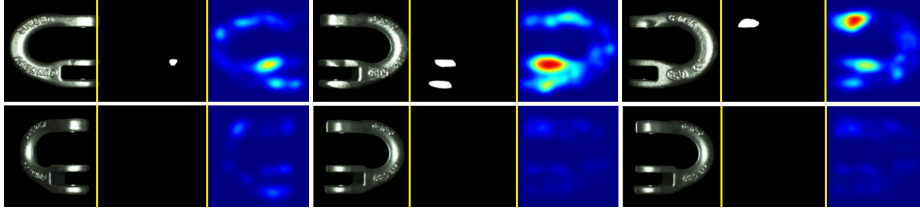


Fig. 6: Three anomalous (top row) and three nominal samples (bottom row) as predicted by GLASS. The ground truth mask for each is in the middle.

computed optimal threshold and use it in a decision rule on a *new* dataset. The most common choice for selecting a threshold is to use Youden’s index to determine the maximum value of a ROC curve. However, common alternatives are to select the threshold by, e.g., maximizing the sensitivity/specificity trade-off, maximizing a particular cost-based metric, or another criterion, such as taking into account the real-world cost of a false positive/negative (e.g., “*what is the cost of an overlooked defective part?*”).

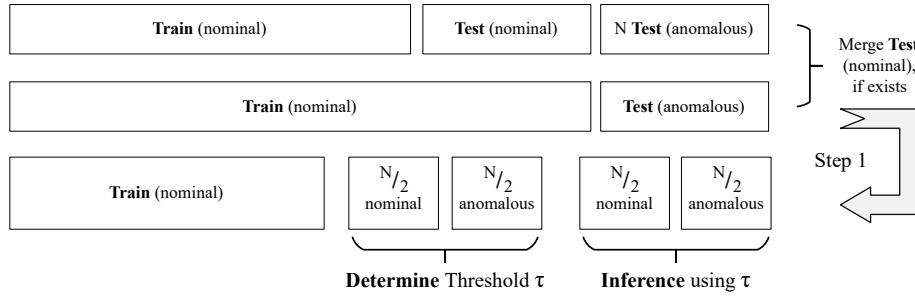


Fig. 7: Repartitioning of the training and testing data is the first step in the pipeline for obtaining robust factory-level estimates.

In the following, we describe the necessary steps that allow obtaining an unbiased estimator for the empirical risk. These steps are in accordance with existing guidelines that are recommended also in cases of limited samples sizes [31]. We assume that a three-part dataset exists to begin with: nominal training samples, as well nominal and anomalous test samples. We further assume that either the nominal test samples were drawn i.i.d. from the same distribution as the train samples, or that no nominal test set exists (second row, Figure 7).

- Step 0)** (Optional) If a nominal test set exists and the i.i.d. assumption holds, merge it into the nominal train set.
- Step 1)** Perform a random, but deterministic three-way split. Separate the anomalous test data randomly into one half each for determining the threshold and for inferencing. Retain two partitions of the same size

of nominal samples. Separate the remaining nominal samples into the training data.²

- Step 2)** Train model on remaining nominal data until convergence.
- Step 3)** Predict anomaly scores on *threshold* dataset and determine the optimal threshold τ for separating nominal/anomalous samples.
- Step 4)** Apply threshold τ to *inference* dataset and compute desired metrics. Record all results.
- Step 5)** Repeat K times for K -fold cross-validation (go to Step 1).

As for the number of repeats, a typical recommendation is at least 5–10. However, there exist insightful empirical guidelines for cross-validation experimentation for obtaining stable model assessments [2]. Choosing partition sizes different from $N/2$ in Step 1 will not result in an unbiased estimation. This is because ROC curve and AUROC can be overly optimistic for severely imbalanced classification problems and unsuitable when false negatives and false positives have significantly different costs [3].

5 Discussion and Future Work

Approaching the G-link problem naïvely and without prior assumptions inspired us to look beyond too-good-looking AUROCs and to iteratively identify and correct a problem that was inherently rooted in data quality and model choice. We identify an overall fragility of likelihood-based methods without representation learning (embedding) under real industrial conditions. We learn that the amount of natural variance present in a dataset strongly affects model performance and its ability to generalize and discriminate. Lastly, we highlight the importance of inspecting reported metrics and examining their applicability to factory-level operations.

We hypothesize that the synthesis of local and global anomalies is what allows GLASS to better concretize its in-distribution manifold. While the synthesis of local anomalies was done many times in previous works, the global anomaly synthesis strategy is novel. However, GLASS is specialized to imagery and, as such, the procedure using clamped gradient ascent is hard to control (hence the introduction of stochasticity). We have since begun to work on a flow-based model that exploits the fact that under a known base distribution (e.g., isotropic Gaussian), we can take controlled steps that allow us to gain insights into the geometry of the latent data manifold. This technique, in combination with controlled noise, allows us to effectively traverse the loss space linearly along all dimensions and to controllably synthesize anomalies.

Acknowledgments. Our project “*In-line visual inspection using unsupervised learning*” is a 2022 Vinnova project for Advanced Digitization, application number 2022-03018. We would like to sincerely thank our project partners and co-financers, namely

² If the image set contains multiple recordings of the same object, as is the case in our “G-link” dataset, one must ensure images of the same object are not in both train and test to prevent information leakage.

Linnaeus University and its High-Performance Computing Center, Gimic AB, SKF, and Gunnebo Industries.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C.: The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *Int. J. Comput. Vis.* **129**(4), 1038–1059 (2021). <https://doi.org/10.1007/S11263-020-01400-4>
2. Bouckaert, R.R.: Estimating replicability of classifier learning experiments. In: Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4–8, 2004. vol. 69. ACM (2004). <https://doi.org/10.1145/1015330.1015338>
3. Branco, P., Torgo, L., Ribeiro, R.P.: A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.* **49**(2) (Aug 2016). <https://doi.org/10.1145/2907070>
4. Caterini, A.L., Loaiza-Ganem, G.: Entropic issues in likelihood-based OOD detection. In: Pradier, M.F., Schein, A., Hyland, S.L., Ruiz, F.J.R., Forde, J.Z. (eds.) I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021, Virtual Workshop, December 13, 2021. Proceedings of Machine Learning Research, vol. 163, pp. 21–26. PMLR (2021), <https://proceedings.mlr.press/v163/caterini22a.html>
5. Chai, J., Zeng, H., Li, A., Ngai, E.W.: Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications* **6**, 100134 (Dec 2021). <https://doi.org/10.1016/j.mlwa.2021.100134>
6. Chen, Q., Luo, H., Lv, C., Zhang, Z.: A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXVII. Lecture Notes in Computer Science, vol. 15125, pp. 37–54. Springer (2024). https://doi.org/10.1007/978-3-031-72855-6_3
7. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23–28, 2014, pp. 3606–3613. IEEE Computer Society (2014). <https://doi.org/10.1109/CVPR.2014.461>
8. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada. pp. 2672–2680 (2014), <https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html>
9. Goyal, S., Raghunathan, A., Jain, M., Simhadri, H.V., Jain, P.: DROCC: deep robust one-class classification. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 3711–3721. PMLR (2020), <http://proceedings.mlr.press/v119/goyal20c.html>

10. Gu, C., Ren, S., Duan, Q.: Unsupervised anomaly detection of industrial images based on dual generator reconstruction networks. *Inf. Technol. Control.* **53**(2), 331–341 (2024). <https://doi.org/10.5755/J01.ITC.53.2.36018>
11. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*. pp. 15637–15648 (2019), <https://proceedings.neurips.cc/paper/2019/hash/a2b15837edac15df90721968986f7f8e-Abstract.html>
12. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. *Science* **313**(5786), 504–507 (2006). <https://doi.org/10.1126/science.1127647>, <https://www.science.org/doi/abs/10.1126/science.1127647>
13. Huang, Y., Qiu, C., Yuan, K.: Surface defect saliency of magnetic tile. *Vis. Comput.* **36**(1), 85–96 (2020). <https://doi.org/10.1007/S00371-018-1588-5>
14. Hönel, S.: 3,992 Images of Blasted Forged Steel Coupling Links (“G-Links”) Including Ground Truth Masks (Jun 2025). <https://doi.org/10.5281/zenodo.15693611>
15. Hönel, S.: Github repository: Replication package for paper: “From Lab to Factory: Pitfalls and Guidelines for Self-/Unsupervised Defect Detection on Low-Quality Industrial Images” (Jun 2025). <https://doi.org/10.5281/zenodo.15693842>
16. Jezek, S., Jonak, M., Burget, R., Dvorak, P., Skotak, M.: Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In: *13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, ICUMT 2021, Brno, Czech Republic, October 25–27, 2021*. pp. 66–71. IEEE (2021). <https://doi.org/10.1109/ICUMT54235.2021.9631567>
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1–6, 2023*. pp. 3992–4003. IEEE (2023). <https://doi.org/10.1109/ICCV51070.2023.00371>
18. Papamakarios, G., Nalisnick, E.T., Rezende, D.J., Mohamed, S., Lakshminarayanan, B.: Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.* **22**, 57:1–57:64 (2021), <https://jmlr.org/papers/v22/19-1028.html>
19. Park, S., Lee, K.H., Ko, B., Kim, N.: Unsupervised anomaly detection with generative adversarial networks in mammography. *Scientific Reports* **13**(1), 2925 (2023). <https://doi.org/10.1038/s41598-023-29521-z>
20. Pless, R., Souvenir, R.: A survey of manifold learning for images. *IPSJ Trans. Comput. Vis. Appl.* **1**, 83–94 (2009). <https://doi.org/10.2197/IPSJTCA.1.83>
21. Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., DePristo, M.A., Dillon, J.V., Lakshminarayanan, B.: Likelihood ratios for out-of-distribution detection. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*. pp. 14680–14691 (2019), <https://proceedings.neurips.cc/paper/2019/hash/1e79596878b2320cac26dd792a6c51c9-Abstract.html>

22. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded SAM: assembling open-world models for diverse visual tasks. *CoRR abs/2401.14159* (2024). <https://doi.org/10.48550/ARXIV.2401.14159>
23. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144. ACM (2016). <https://doi.org/10.1145/2939672.2939778>
24. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.V.: Towards total recall in industrial anomaly detection. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. pp. 14298–14308. IEEE (2022). <https://doi.org/10.1109/CVPR52688.2022.01392>
25. Rudolph, M., Wandt, B., Rosenhahn, B.: Same same but different: Semi-supervised defect detection with normalizing flows. In: *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*. pp. 1906–1915. IEEE (2021). <https://doi.org/10.1109/WACV48630.2021.00195>
26. Rudolph, M., Wehrbein, T., Rosenhahn, B., Wandt, B.: 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*. pp. 1829–1838. IEEE (2022). <https://doi.org/10.1109/WACV51458.2022.00189>
27. Ruff, L., Görnitz, N., Deecke, L., Siddiqui, S.A., Vandermeulen, R.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: Dy, J.G., Krause, A. (eds.) *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018. Proceedings of Machine Learning Research*, vol. 80, pp. 4390–4399. PMLR (2018), <http://proceedings.mlr.press/v80/ruff18a.html>
28. Sakurada, M., Yairi, T.: Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: Rahman, A., Deng, J.D., Li, J. (eds.) *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, Gold Coast, Australia, QLD, Australia, December 2, 2014. p. 4. ACM (2014). <https://doi.org/10.1145/2689746.2689747>
29. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research*, vol. 97, pp. 6105–6114. PMLR (2019), <http://proceedings.mlr.press/v97/tan19a.html>
30. Tao, X., Gong, X., Zhang, X., Yan, S., Adak, C.: Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Trans. Instrum. Meas.* **71**, 1–21 (2022). <https://doi.org/10.1109/TIM.2022.3196436>
31. Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J.: Machine learning algorithm validation with a limited sample size. *PLOS ONE* **14**(11), 1–20 (11 2019). <https://doi.org/10.1371/journal.pone.0224365>
32. Wang, S., Li, Q., Luo, H., Lv, C., Zhang, Z.: Produce once, utilize twice for anomaly detection. *IEEE Trans. Circuits Syst. Video Technol.* **34**(11), 11751–11767 (2024). <https://doi.org/10.1109/TCSVT.2024.3420775>

33. Zavrtanik, V., Kristan, M., Skocaj, D.: Dræm - A discriminatively trained reconstruction embedding for surface anomaly detection. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 8310–8319. IEEE (2021). <https://doi.org/10.1109/ICCV48922.2021.00822>