

# Same Same But DifferNet: Semi-Supervised Defect Detection with Normalizing Flows

Marco Rudolph

Bastian Wandt

Bodo Rosenhahn

Leibniz University Hanover

{rudolph, wandt, rosenhahn}@tnt.uni-hannover.de

## Abstract

The detection of manufacturing errors is crucial in fabrication processes to ensure product quality and safety standards. Since many defects occur very rarely and their characteristics are mostly unknown *a priori*, their detection is still an open research question. To this end, we propose DifferNet: It leverages the descriptiveness of features extracted by convolutional neural networks to estimate their density using normalizing flows. Normalizing flows are well-suited to deal with low dimensional data distributions. However, they struggle with the high dimensionality of images. Therefore, we employ a multi-scale feature extractor which enables the normalizing flow to assign meaningful likelihoods to the images. Based on these likelihoods we develop a scoring function that indicates defects. Moreover, propagating the score back to the image enables pixel-wise localization. To achieve a high robustness and performance we exploit multiple transformations in training and evaluation. In contrast to most other methods, ours does not require a large number of training samples and performs well with as low as 16 images. We demonstrate the superior performance over existing approaches on the challenging and newly proposed MVTec AD [4] and Magnetic Tile Defects [14] datasets.

## 1. Introduction

In industrial manufacturing processes the quality of the products is constantly monitored and improved. Hence, small defects during fabrication need to be detected reliably. However, manufacturers do not know in advance which types of defects will occur and most of them appear so infrequently that no defective examples are available. Even if some defect types are known, new types can still occur any time due to unforeseeable events during manufacturing. Consequently, reliable defect detection cannot be done with supervised machine learning approaches. We propose a solution for semi-supervised defect detection where only pos-

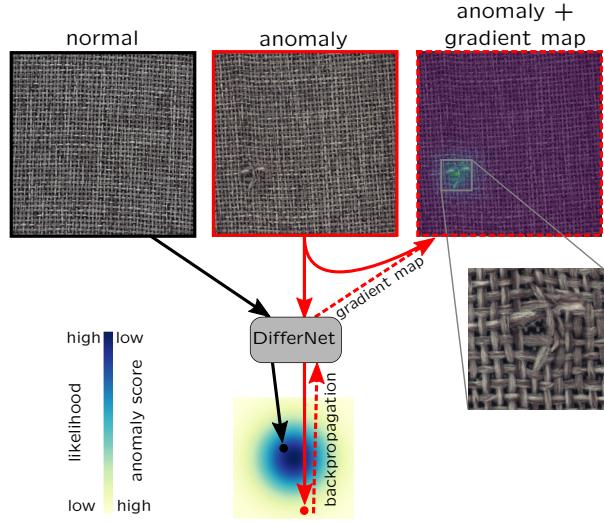


Figure 1. DifferNet assigns likelihoods to inputs which makes it usable to detect defects. In contrast to the top left image without any defect, the top middle image is assigned to a high anomaly score due to the defect (see the enlarged patch on the right side). Additionally, DifferNet identifies the defective region by back-propagating the likelihood loss up to the input which gives a gradient map (top right image). This allows for a detailed analysis of defect position and shape.

itive examples and no defective examples are present during training. This is also known as anomaly detection.

In general, anomaly detection describes the problem of determining whether a data sample differs from a set of given *normal* data. There are various approaches for general anomaly detection on images, summarized in Section 2. Defect detection is a specific sub-problem where visually similar normal samples and only slightly different anomalous samples are present. While traditional anomaly detection methods are well-suited to data with high intra-class variance, they are not able to capture subtle differences. We tackle this problem by employing an accurate density estimator on image features extracted by a convolutional neural network. The feature distribution of normal samples is

captured by utilizing the latent space of a normalizing flow [22]. Unlike other generative models such as variational autoencoders [17] or GANs [12], there exists a bijective mapping between feature space and latent space in which each vector is assigned to a likelihood. This enables DifferNet to calculate a likelihood for each image. From this likelihood we derive a scoring function to decide if an image contains an anomaly. Figure 1 visualizes the core idea behind our method. The most common samples are assigned to a high likelihood whereas uncommon images are assigned to a lower likelihood. Since defects are not present during training, they are mapped to a low likelihood. We further improve the descriptiveness of the feature extractor by using multi-scale inputs. To derive a meaningful scoring function, we include likelihoods of several transformations of the image. Thereby DifferNet gains flexibility and robustness to detect various types of defects. We show that our method considers even small changes in data while other approaches struggle to detect them. Moreover, due to the efficiency of the feature extractor and the proposed image transformations, our approach even outperforms the state of the art when trained with a low number of training samples. Besides defect detection, DifferNet’s architecture allows for localization of defects by having expressive gradients of input pixels regarding the scoring functions. A high magnitude of the gradient in an image region signals an area with anomalous features which helps to identify the defect.

Our work comprises the following contributions:

- Detection of anomalies via the usage of likelihoods provided by a normalizing flow on multi-scale image features with multi-transform evaluation.
- Anomaly localization without training labels, the necessity of any pixel-wise optimization and sub-image detection.
- Applicability on small training sets. Even with a low number of training examples our approach achieves competitive results.
- State-of-the-art detection performance on MVTec AD and Magnetic Tile Defects.
- Code is available on GitHub<sup>1</sup>.

## 2. Related Work

### 2.1. Anomaly Detection

Existing methods for anomaly detection can be roughly divided into approaches based on generative models and pretrained networks. The most relevant methods to our work are briefly presented in the following subsections. Note that we focus on works that deal with image anomaly

detection rather than anomaly localization to keep the focus on our main problem.

#### 2.1.1 Detection with Pretrained Networks

There are several works which use the feature space of a pretrained network to detect anomalies. In most cases simple traditional machine learning approaches are used to obtain an anomaly score. Andrews *et al.* [2] apply a One-Class-SVM on VGG [28] features of the training images. Nazare *et al.* [21] evaluated different normalization techniques to have a 1-Nearest-Neighbor-classifier on a PCA-reduced representation of the feature space. Localization of the anomalies is achieved by evaluating the method on many overlapping patches. However, this is very costly. Sabokrou *et al.* [25] models the anomaly-free feature distribution as an unimodal Gaussian distribution. Therefore, they cannot capture multimodal distributions as opposed to our method.

These techniques only work for particular classes in defect detection. In contrast to our proposed DifferNet, existing approaches do not appear to be robust and powerful enough for defect detection. None of the above techniques take advantage of the flexibility of another neural network on top of the pretrained model. Another benefit of our method compared to these approaches is being able to compute gradients w.r.t. the inputs which can be utilized to compute anomaly maps.

#### 2.1.2 Generative Models

Generative models, such as autoencoders [19, 17, 23] and GANs [12], are able to generate samples from the manifold of the training data. Anomaly detection approaches using these models are based on the idea that the anomalies cannot be generated since they do not exist in the training set.

Autoencoder-based approaches try to detect anomalies by comparing the output of an autoencoder to its input. Thus, a high reconstruction error should indicate an anomalous region. Bergmann *et al.* [5] proposes SSIM-loss to make the reconstruction error dependent on visual similarity. In many cases autoencoder-based methods fail because they generalize too strongly, *i.e.* anomalies can be reconstructed as good as normal samples. Gong *et al.* [11] tackle the generalization problem by employing memory modules which can be seen as a discretized latent space. Zhai *et al.* [30] connect regularized autoencoders with energy-based models to model the data distribution and classify samples with high energy as an anomaly.

GAN-based approaches assume that only positive samples can be generated. Schlegl *et al.* [26] propose a two-stage training method: The GAN is learned first and an encoder is optimized as an inverse generator. Using the generator as decoder enables the calculation of a reconstruc-

<sup>1</sup><https://github.com/marco-rudolph/differnet>

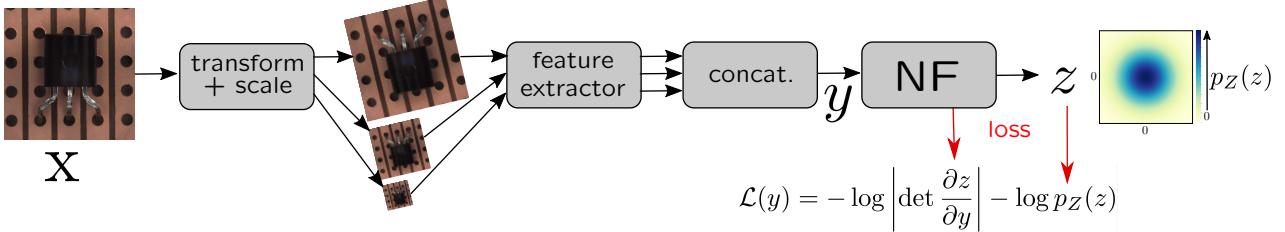


Figure 2. Overview of our pipeline: Multiple scales of a transformed input image are fed into a feature extractor. The distribution of its concatenated outputs is captured by transforming it via a normalizing flow (NF) into a normal distribution by maximum likelihood training.

tion loss alongside the difference in discriminator features of original and reconstructed image to obtain an anomaly score. Akcay *et al.* [1] make use of adversarial training by letting an autoencoder directly act as generating part of the GAN. This enforces the property of the decoder to only generate normal-like samples which can be measured by the difference between the embedding of the original and the reconstructed data.

We argue that generative models are appropriate for a wide range of defect detection scenarios since they strongly depend on the anomaly type. For example, the size and structure of the defective area heavily influence the anomaly score. If the region of interest shows high frequency structures, they cannot be represented accurately. Often, other instance-specific structures influence the reconstruction error more than the anomaly. In contrast, we show that DifferNet handles significantly more and various defect types. Additionally, our method does not rely on a large number of training samples compared to generative models.

## 2.2. Normalizing Flows

Normalizing Flows (NF) [22] are neural networks that are able to learn transformations between data distributions and well-defined densities. Their special property is that their mapping is bijective and they can be evaluated in both directions. First, they assign likelihoods to a given sample. Second, data is generated by sampling from the modeled distribution. The bijectivity is ensured by stacking layers of affine transforms which are fixed or autoregressive. A common class of such autoregressive flows is *MADE* (Germain *et al.* [9]) which makes use of the Bayesian chain rule to decompose the density. These models can learn distributions of large datasets containing mostly small images. In contrast, we capture the distribution of a comparably small number of images at a high resolution. Autoregressive flows compute likelihoods fast, but are slow at sampling. Inverse autoregressive flows, proposed by Kingma *et al.* [16], show the exact opposite behavior. Real-NVP [8] can be seen as a special inverse autoregressive flow which is simplified such that both forward and backward pass can be processed quickly. Similar to Ardizzone *et al.* proposed for *Invertible Neural Networks* [3], we integrate a parametrized clamping

mechanism to the affine transformation of the Real-NVP-layers to obtain a more stable training process. Details to the so-called coupling layers and the clamping mechanism of Real-NVP are explained in Section 3.1.

The property of normalizing flows as an adequate estimator of probability densities to detect anomalies has not raised much attention yet, although some works present promising results using Real-NVP and MADE [24, 27, 7]. However, none of the works deal with visual data.

## 3. Method

Figure 2 shows an overview of our pipeline. Our method is based on density estimation of image features  $y \in Y$  from the anomaly-free training images  $x \in X$ . Let  $f_{\text{ex}} : X \rightarrow Y$  be the mapping of a pretrained feature extractor which is not further optimized. The estimation of  $p_Y(y)$ , provided by  $f_{\text{ex}}(x)$ , is achieved by mapping from  $Y$  to a latent space  $Z$  – with a well-defined distribution  $p_Z(z)$  – by applying a normalizing flow  $f_{\text{NF}} : Y \rightarrow Z$ . Likelihoods for image samples are directly calculated from  $p_Z(z)$ . Features of anomalous samples should be out of distribution and hence have lower likelihoods than normal images. Likelihoods of multiple transforms on the image are maximized in training and used in inference for a robust prediction of the anomaly score. To capture structures at different scales and thus having a more descriptive representation in  $y$ , we further define  $f_{\text{ex}}$  as the concatenation of features at 3 scales.

### 3.1. Normalizing Flow

The normalizing flow acts as a bijector between feature space  $Y$  and latent space  $Z$  by using affine transformations. Likelihoods are computed from  $Z$  according to the modelled distribution. We model  $z \sim \mathcal{N}(0, I)$  in our latent space which gives us a well-defined density  $p_Z(z)$  for  $z$ . Following from bijectivity, for every feature input there is a unique  $z$  of the same dimension and vice versa.

#### 3.1.1 Architecture

We use an architecture of coupling layers as proposed in Real-NVP [8]. The detailed structure of one block is shown in Figure 3. The design of  $f_{\text{NF}}$  is a chain of such blocks,

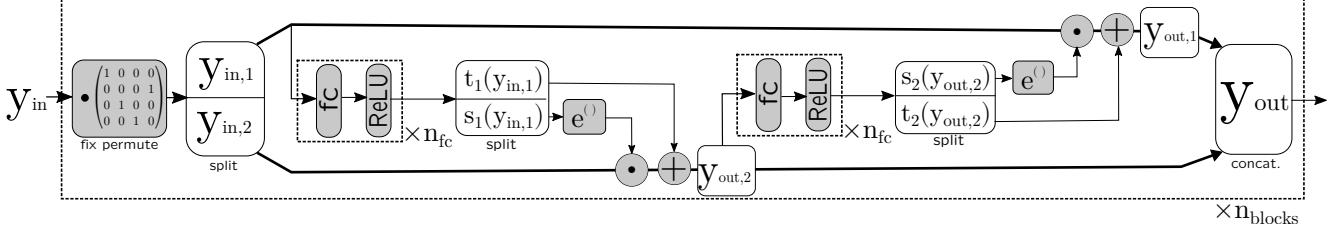


Figure 3. Architecture of one block inside the normalizing flow: After a fixed random permutation, the input  $y_{in}$  is split into two parts that regress scale and shift parameters to transform their respective counterpart. Symbols  $\odot$  and  $\oplus$  denote element-wise multiplication and addition, respectively. Numerical operations are symbolized by grey blocks. White blocks contain variable names.

consisting of permutations followed by scale and shift operations.

To apply the scale and shift operations, the input  $y_{in}$  is split into  $y_{in,1}$  and  $y_{in,2}$  that manipulate each other by regressing multiplicative and additive components in subnetworks  $s$  and  $t$ ; these manipulations are applied to their respective counterpart successively. The scale and shift operations are described by

$$\begin{aligned} y_{out,2} &= y_{in,2} \odot e^{s_1(y_{in,1})} + t_1(y_{in,1}) \\ y_{out,1} &= y_{in,1} \odot e^{s_2(y_{out,2})} + t_2(y_{out,2}), \end{aligned} \quad (1)$$

with  $\odot$  as the element-wise product. Using an exponential function before scaling preserves the affinity property by ensuring non-zero coefficients. The internal functions  $s$  and  $t$  can be any differentiable function, which in our case is implemented as a fully-connected network that regresses both components by splitting the output (see Figure 3). Similar to Ardizzone *et al.* [3], we apply soft-clamping to the values of  $s$  to preserve model stability which is crucial in our case for better convergence. This is achieved by using the activation

$$\sigma_\alpha(h) = \frac{2\alpha}{\pi} \arctan \frac{h}{\alpha} \quad (2)$$

as the last layer of  $s$ . This prevents large scaling components by restricting  $s(y)$  to the interval  $(-\alpha, \alpha)$ .

Each block first performs a predefined random permutation on the features to allow each dimension to affect all other dimensions at some point. The output of one coupling block is given by the concatenation of  $y_{out,1}$  and  $y_{out,2}$ .

### 3.1.2 Training

The goal during training is to find parameters for  $f_{NF}$  that maximize likelihoods for extracted features  $y \in Y$  which are quantifiable in the latent space  $Z$ . With the mapping  $z = f_{NF}(y)$  and according to the change-of-variables formula Eq. 3, we describe this problem as maximizing

$$p_Y(y) = p_Z(z) \left| \det \frac{\partial z}{\partial y} \right|. \quad (3)$$

This is equivalent to maximizing the log-likelihood, which is more convenient here, since the terms simplify when inserting the density function of a standard normal distribution as  $p_Z(z)$ . We use the negative log-likelihood loss  $\mathcal{L}(y)$  to obtain a minimization problem:

$$\begin{aligned} \log p_Y(y) &= \log p_Z(z) + \log \left| \det \frac{\partial z}{\partial y} \right| \\ \mathcal{L}(y) &= \frac{\|z\|_2^2}{2} - \log \left| \det \frac{\partial z}{\partial y} \right|. \end{aligned} \quad (4)$$

Intuitively, we want  $f_{NF}$  to map all  $y$  as close as possible to  $z = 0$  while penalizing trivial solutions with scaling coefficients close to zero<sup>2</sup>. The latter is ensured by the negative log determinant of the Jacobian  $\frac{\partial z}{\partial y}$  in  $\mathcal{L}(y)$ . In our case the log determinant of the Jacobian is the sum of scaling coefficients before exponentiation.

During Training,  $\mathcal{L}(y)$  is optimized for features  $y$  of different transformations of an input image for a fixed epoch length. Section 4.2 describes the training in more detail.

### 3.2 Scoring Function

We use the calculated likelihoods as a criterion to classify a sample as anomalous or normal. To get a robust anomaly score  $\tau(x)$ , we average the negative log-likelihoods using multiple transformations  $T_i(x) \in \mathcal{T}$  of an image  $x$ :

$$\tau(x) = \mathbb{E}_{T_i \in \mathcal{T}} [-\log p_Z(f_{NF}(f_{ex}(T_i(x)))]). \quad (5)$$

As  $\mathcal{T}$  we choose rotations and manipulations of brightness and contrast. An image is classified as anomalous if the anomaly score  $\tau(x)$  is above the threshold value  $\theta$ . Thus, the decision can be expressed as

$$\mathcal{A}(x) = \begin{cases} 1 & \text{for } \tau(x) \geq \theta \\ 0 & \text{for } \tau(x) < \theta \end{cases}, \quad (6)$$

where  $\mathcal{A}(x) = 1$  indicates an anomaly. In Section 4  $\theta$  is varied to calculate the *Receiver Operating Characteristic* (ROC).

<sup>2</sup>The exponentiation inhibits the coefficients from being zero.

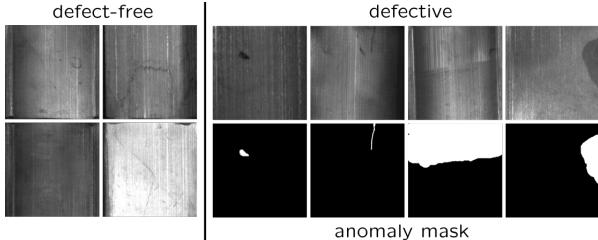


Figure 4. Samples of defect-free and defective images from Magnetic Tile Defects [14].

### 3.3. Localization

In contrast to several other approaches, ours is not optimized for localizing the defects on the image. Nevertheless, our method localizes areas where anomalous features occur. Our method allows for propagating the negative log-likelihood  $\mathcal{L}$  back to the input image  $x$ . The gradient  $\nabla x_c$  of each input channel  $x_c$  is a value indicating how much the pixels influence the error which relates to an anomaly. For better visibility we blur these gradients with a Gaussian kernel  $G$  and sum the absolute values over the channels  $C$  according to

$$g_x = \sum_{c \in C} |G * \nabla x_c|, \quad (7)$$

with  $*$  as 2D convolution and  $|\cdot|$  as the element-wise absolute value, which results in the gradient map  $g_x$ . Averaging the maps of multiple rotations of one single image – after rotating back the obtained maps – gives a robust localization.

## 4. Experiments

### 4.1. Datasets

In this paper, we evaluate our approach on real-world defect detection problems. We use the challenging and newly proposed datasets MVTec AD [4] and Magnetic Tile Defects (MTD) [14]. The difficulty in these datasets lies in the similarity of anomalies and normal examples.

To the best of our knowledge, MVTec AD is the only publicly available multi-object and multi-defect anomaly dataset. It contains 5354 high-resolution color images of 10 object and 5 texture categories. The number of training samples per category ranges from 60 to 320, which is challenging for the estimation of the distribution of normal samples. Several defect types per category, such as little cracks, deformations, discolorizations and scratches are occurring in the test set. Some of which are shown in Figure 8. In total, the dataset includes 70 defect types. The anomalies differ in their size, shape and structure and thus cover several scenarios in industrial defect detection.

Magnetic Tile Defects [14] comprises grayscale images of magnetic tiles under different illuminations with and

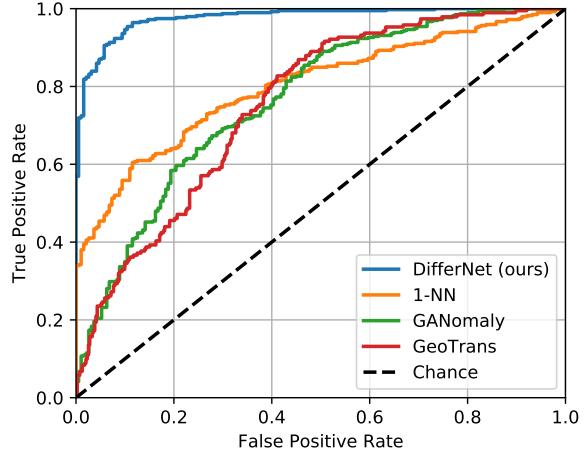


Figure 5. ROC-Curve for different methods for detecting defects in MTD. DifferNet is significantly more accurate in detecting the defects compared to other approaches. Best viewed in color.

without defects. Such tiles should provide a constant magnetic potential in engines. We split the 952 defect-free images randomly into a train and a test set, where the test set contains 20% of the data. All 392 defect images are used for testing. These show frayed or uneven areas, cracks, breaks and blowholes as anomalies, as shown in Figure 4. A lot of defect-free images contain variations that are similar to anomalies.

### 4.2. Implementation Details

For all experiments, we use the convolutional part of AlexNet [18] as the feature extractor and apply global average pooling on each feature map at each scale. We tested more complex topologies, for instance ResNet [13] and VGG [28], but did not observe better performance. We prefer the smaller AlexNet since it is sufficient for our purpose. The feature extractor is pretrained on ImageNET and remains fixed during training. We use features at 3 scales with input image sizes of  $448 \times 448$ ,  $224 \times 224$  and  $112 \times 112$  pixels - resulting in  $3 \cdot 256 = 768$  features. The normalizing flow consists of 8 coupling blocks with fully connected networks as internal functions  $s$  and  $t$ . These include 3 hidden dense layers with a size of 2048 neurons and ReLU activations. We set the mentioned clamping parameter  $\alpha = 3$ . For training, we use the Adam Optimizer [15] with the author-suggested  $\beta$ -parameters and a learning rate of  $2 \cdot 10^{-4}$ . As transformations  $\mathcal{T}$ , random rotations, which are uniformly distributed in the interval  $[0, 2\pi]$ , are applied. In addition, we manipulated the contrast and brightness of the magnetic tiles with an uniformly distributed random factor in the interval  $[0.85, 1.15]$ . In each case of training and inference the same transformations are applied. We train our models for 192 epochs with a batch size of 96.

	Category	GeoTrans [10]	GANomaly [1]	DSEBM [30]	OCSVM [2]	1-NN [21]	<b>DifferNet (ours)</b>	DifferNet (16 shots)
Textures	Grid	61.9	70.8	<u>71.7</u>	41.0	55.7	<b>84.0</b>	65.8
	Leather	84.1	84.2	<u>41.6</u>	88.0	90.3	<b>97.1</b>	<u>92.9</u>
	Tile	41.7	79.4	69.0	87.6	96.9	<b>99.4</b>	<u>98.9</u>
	Carpet	43.7	69.9	41.3	62.7	<u>81.1</u>	<b>92.9</b>	77.0
	Wood	61.1	83.4	95.2	95.3	93.4	<b>99.8</b>	<u>99.2</u>
Objects	Bottle	74.4	89.2	81.8	<b>99.0</b>	98.7	<b>99.0</b>	98.5
	Capsule	67.0	<u>73.2</u>	59.4	54.4	71.1	<b>86.9</b>	61.4
	Pill	63.0	74.3	80.6	72.9	<u>83.7</u>	<b>88.8</b>	65.1
	Transistor	<u>86.9</u>	79.2	74.1	56.7	75.6	<b>91.1</b>	76.6
	Zipper	82.0	74.5	58.4	51.7	<u>88.6</u>	<b>95.1</b>	88.3
	Cable	78.3	75.7	68.5	80.3	<u>88.5</u>	<b>95.9</b>	86.4
	Hazelnut	35.9	78.5	76.2	91.1	<u>97.9</u>	<b>99.3</b>	97.3
	Metal Nut	<u>81.3</u>	70.0	67.9	61.1	76.7	<b>96.1</b>	77.7
	Screw	50.0	74.6	<b>99.9</b>	74.7	67.0	<u>96.3</u>	75.9
	Toothbrush	<u>97.2</u>	65.3	78.1	61.9	91.9	<b>98.6</b>	92.3
<b>Average</b>		67.2	76.2	70.9	71.9	83.9	<b>94.9</b>	<u>87.3</u>

Table 1. Area under ROC in % for detected anomalies of all categories of MVTec AD [4] grouped into textures and objects. Best results are in bold, second best underlined. OCSVM and 1-NN are calculated on the same feature extractor outputs our NF is trained on. *16 shots* denotes a model trained on only 16 images.

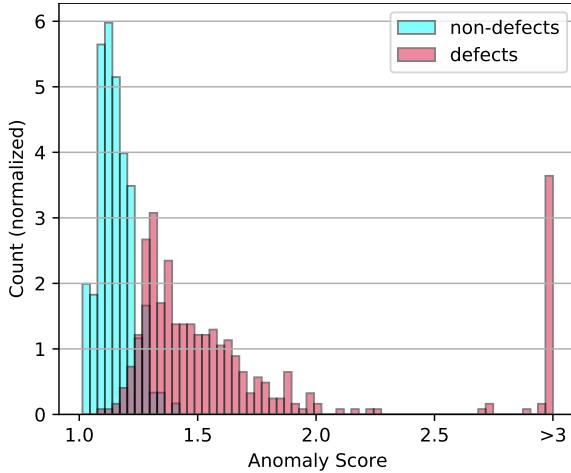


Figure 6. Normalized histogram of DifferNet’s anomaly scores for the test images of MTD. As can be seen, the score is a reliable indicator for defects except for a narrow range of some borderline cases. Note that the rightmost bar summarizes all scores above 3.

### 4.3. Detection

For reporting the performance of our method regarding the detection of anomalies, we follow [1] and compute the Area Under Receiver Operator Characteristics (AUROC) depending on the scoring function which we obtained as described in Section 3. It measures the area under the true positive rate as a function of the false positive rate. The AUROC metric is not sensitive to any threshold or the percentage of anomalies in the test set. Besides other anomaly

detection methods, we compare our method to the baselines one-class SVM (OCSVM) [2] and the distance to the nearest neighbor (1-NN) after PCA reduction to 64 dimensions and z-score normalization [21]. Note that both methods OCSVM and 1-NN are adapted to our setting: We used every technique of our pipeline (see Figure 2) but replaced the normalizing flow with them. We evaluate several transformations and used the mean respective score. Apart from these approaches and some state-of-the-art models, we compare our method with GeoTrans [10] which cannot be assigned to generative and pretrained methods described in Section 2. GeoTrans computes an anomaly score based on the classification of conducted transformations. Table 1 shows the results for MVTec AD. Compared to other approaches, our method outperforms existing methods in almost every category, up to a large margin of 15%. In all of the 15 categories our method achieves an AUROC of at minimum 84%, which shows that our approach is not limited to a specific set of defects or features. The fact that 1-NN outperforms other competitors except us, demonstrates that our feature extraction and evaluation is well-suited for the problem.

We can observe similar characteristics on MTD, seen in Table 2. The ROC-Curve in Figure 5 shows that our method provides a much higher true positive rate for any false positive rate. DifferNet achieves a recall of about 50% without any false positive among 191 defect-free test images. The histogram of anomaly scores is visualized in Figure 6. There is a large subset of defective samples whose scores differ significantly from all scores of non-defective samples.

Method	AUROC [%]
GeoTrans[10]	75.5
GANomaly[1]	76.6
DSEBM [30]	57.2
ADGAN [6]	46.4
OCSVM [2]	58.7
1-NN [21]	80.0
<b>DifferNet (ours)</b>	<b>97.7</b>

Table 2. Area under ROC in % for detecting anomalies on MTD

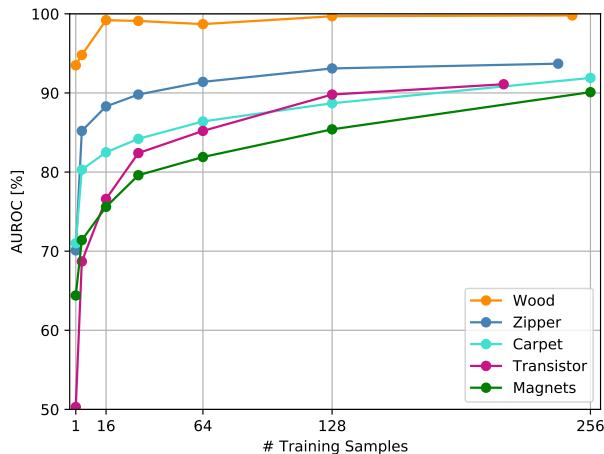


Figure 7. Detection performance of DifferNet, measured by AUROC, depending on the training set size of MTD and of some categories of MVTec AD. Best viewed in color.

The assignment of extremely high scores without any false positive is a characteristic of our method and can be similarly observed for other evaluated product categories.

#### 4.4. Localization

Results of the localization procedure described in Section 3.3 are shown in Figure 8. The localizations are accurate for many types, sizes and shapes of anomalies; despite the average pooling of feature maps before being processed by the normalizing flow. Our architecture produces meaningful gradients which can be explained by the models architecture: First, AlexNet is relatively shallow such that noisy or vanishing gradients are prevented. Second, the bijectivity of the normalizing flow causes a direct relation between all image features  $y$  and all values of  $z$  with non-zero gradients. The gradients tend to appear speckled for larger anomalous regions. We conject the reason is that pixels, leading to features influencing the anomaly score, are usually not located evenly distributed in the corresponding region. However, our method enables the human to perceive the defective region and interpret which areas influenced the networks decision to what extent.

Config.	A	B	C	D	E	F	G
multi-scale	<u>X</u>	<u>X</u>	✓	✓	✓	✓	✓
train transf.	<u>X</u>	✓	<u>X</u>	✓	✓	✓	✓
# test transf.	<u>1</u>	64	<u>1</u>	<u>1</u>	4	<u>16</u>	64
AUROC [%]	84.4	90.2	86.6	86.5	91.6	94.1	<b>94.9</b>

Table 3. Average detection performance for all categories of MVTec AD when modifying our proposed training and evaluation strategy. The columns show parameter configurations named from A to F. Parameters that differ from our proposed configuration are underlined.

#### 4.5. Ablation Studies

To quantify the effects of individual strategies used in our work, we performed an ablation study by comparing the performance on MVTec AD [4] when modifying the strategies. In addition, the model’s behavior for different characteristics of the training set is analyzed.

**Preprocessing Pipeline and Evaluation.** Table 3 compares the detection performance on MVTec AD for different configurations regarding multi-scaling, the usage of transformations in training and the number of used transformations  $T_i$  for evaluation. Having one test transformation means that only the original image was used for evaluation. Note that we outperform existing methods even without the proposed transformations and multi-scale strategy. Since relevant features could appear at any scale, it is beneficial to include features at multiple scales which is shown by an AUROC improvement of 4.7%. Having transformed samples in training is crucial as it enables multi-transform evaluation and helps for generalization and data augmentation. The similar performances of configuration C and D reflect that applying transformations in training is only useful if they are performed in inference as well. The more of these transformations are then used, the more meaningful the score is, as the rising performance of configurations D to G shows.

**Number of Training Samples.** We investigate the effect of the training set size on the detection performance as shown on Figure 7 and on the right of Table 1. The reported results are the average over three runs with different random subsets per training set size. It can be seen that our model and training procedure allows for a stable training even on small training sets. This can be explained by the usage of multiple transformations and the averaging of feature maps. Our model profits from this strategy which is a mixture of augmentation and compression. DifferNet requires only 16 training samples to outperform existing approaches that use the whole training set with 369 samples per class on average. For some classes, the subsets cannot represent the feature variation of normal samples.

**Multimodality.** The feature distributions of the evaluated categories are unimodal. We also investigated the performance on multimodal distributions. Therefore, we also ob-

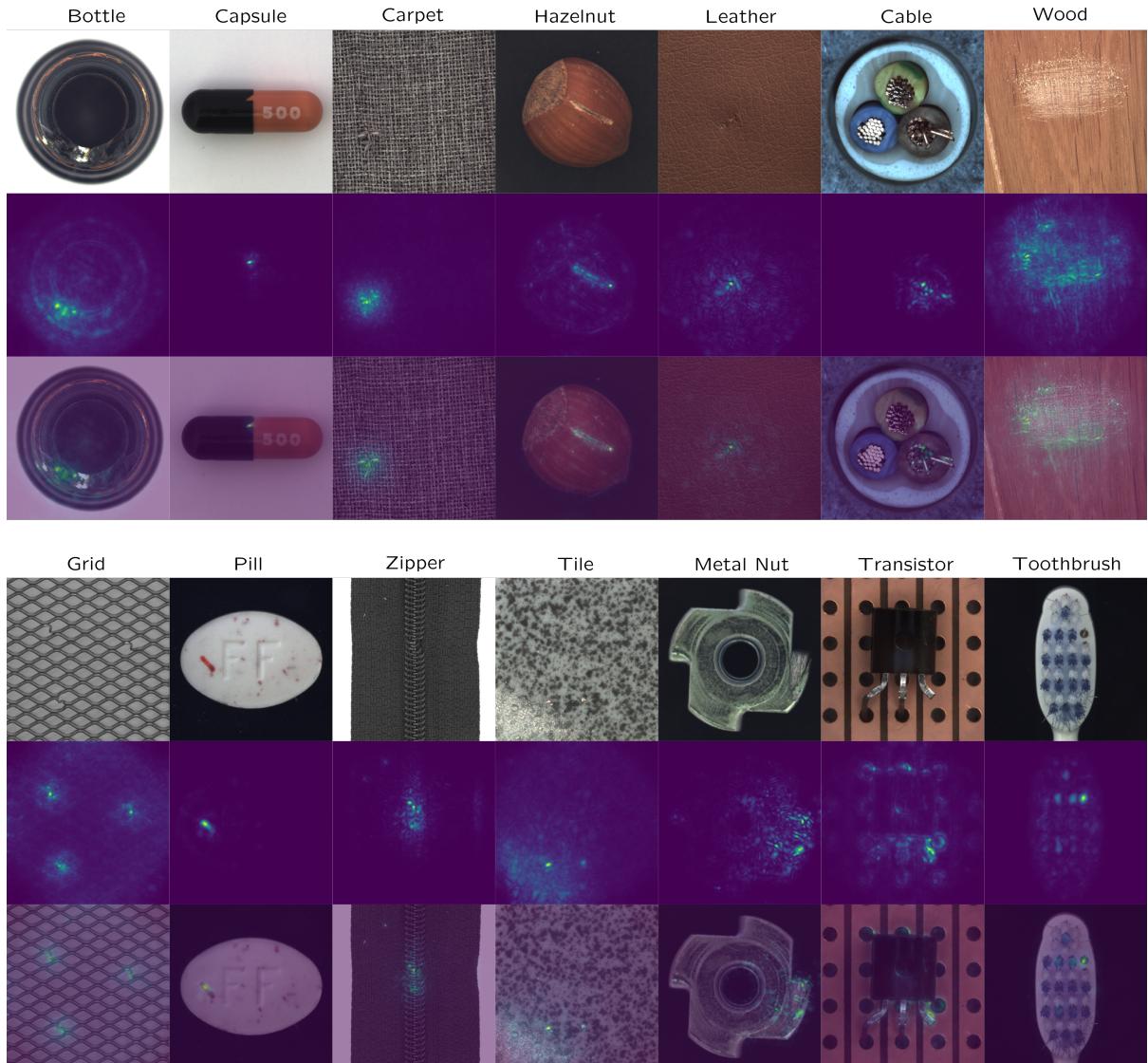


Figure 8. Localization of anomalous regions of different categories in MVTec AD. The upper rows shows the original anomaly images, the mid rows the localizations provided by DifferNet and the lower rows the superimposition of both. They were generated by backpropagating the negative log-likelihood loss to the input image.

served the detection performance when using all 15 categories of MVTec as training data. To capture this more complex distribution, we used 12 coupling blocks. The result is a mean AUROC of 90.2% which shows that our method is able to handle multimodal distributions well. The regressing sub-blocks inside the NF appear to capture the modes and switch between them depending on their input.

## 5. Conclusion

We presented *DifferNet* to detect defects in images by utilizing a normalizing-flow-based density estimation of image features at multiple scales. Likelihoods of several transformations of a single image are used to compute a ro-

bust anomaly score. Therefore, there is no need for a large amount of training samples. The design and scoring function is chosen such that image gradients can be exploited to localize defects. As shown, the method also scales to multimodal distributions which resembles real-world settings. In the future we plan to refine the concept in order to find anomalies in video data comparable to [29, 20].

## Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122).

## References

- [1] Samet Akcay, Amir Atapour Abarghouei, and Toby P. Breckon. Gandomaly: Semi-supervised anomaly detection via adversarial training. *CoRR*, abs/1805.06725, 2018.
- [2] Jerome Andrews, Thomas Tanay, Edward Morton, and Lewis Griffin. Transfer representation-learning for anomaly detection. *JMLR*, 2016.
- [3] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Guided image generation with conditional invertible neural networks. *arXiv preprint arXiv:1907.02392*, 2019.
- [4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtex ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019.
- [5] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.
- [6] Haoqing Cheng, Heng Liu, Fei Gao, and Zhuo Chen. Adgan: A scalable gan-based architecture for image anomaly detection. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (IT-NEC)*, volume 1, pages 987–993. IEEE, 2020.
- [7] Madson LD Dias, César Lincoln C Mattos, Ticiana LC da Silva, José Antônio F de Macedo, and Wellington CP Silva. Anomaly detection in trajectory data with normalizing flows. *arXiv preprint arXiv:2004.05958*, 2020.
- [8] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *ICLR 2017*, 2016.
- [9] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889, 2015.
- [10] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pages 9758–9769, 2018.
- [11] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1705–1714, 2019.
- [12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [14] Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 36(1):85–96, 2020.
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- [16] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.
- [17] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.
- [18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [19] Y. LeCun. Generalization and network design strategies. In R. Pfeifer, Z. Schreter, F. Fogelman, and L. Steels, editors, *Connectionism in Perspective*, Zurich, Switzerland, 1989. Elsevier. an extended version was published as a technical report of the University of Toronto.
- [20] Wentong Liao, Bodo Rosenhahn, and Yang Michael. Gaussian process for activity modeling and anomaly detection. In *International Society for Photogrammetry and Remote Sensing ISA workshop, La Grande Motte, France*, Sept. 2015.
- [21] Tiago Nazare, Rodrigo de Mello, and Moacir Ponti. Are pre-trained cnns good feature extractors for anomaly detection in surveillance videos? *arXiv preprint arXiv:1811.08495*, 2018.
- [22] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.
- [23] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Structuring autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.
- [24] Artem Ryzhikov, Maxim Borisyak, Andrey Ustyuzhanin, and Denis Derkach. Normalizing flows for deep anomaly detection. *arXiv preprint arXiv:1912.09323*, 2019.
- [25] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, Zahra Moayed, and Reinhard Klette. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, 2018.
- [26] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019.
- [27] Maximilian Schmidt and Marko Simic. Normalizing flows for novelty detection in industrial time series data. *arXiv preprint arXiv:1906.06904*, 2019.
- [28] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.
- [29] Michael Ying Yang, Wentong Liao, Yanpeng Cao, and Bodo Rosenhahn. Video event recognition and anomaly detection by combining gaussian process and hierarchical dirichlet process models. In *Photogrammetric Engineering & Remote Sensing*, 2018.
- [30] Shuangfei Zhai, Yu Cheng, Weinig Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly

detection. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning- Volume 48*, pages 1100–1109, 2016.