

DSR – A dual subspace re-projection network for surface anomaly detection

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj

University of Ljubljana, Faculty of Computer and Information Science
`{vitjan.zavrtanik, matej.kristan, danijel.skocaj}@fri.uni-lj.si`

Abstract. The state-of-the-art in discriminative unsupervised surface anomaly detection relies on external datasets for synthesizing anomaly-augmented training images. Such approaches are prone to failure on near-in-distribution anomalies since these are difficult to be synthesized realistically due to their similarity to anomaly-free regions. We propose an architecture based on quantized feature space representation with dual decoders, DSR, that avoids the image-level anomaly synthesis requirement. Without making any assumptions about the visual properties of anomalies, DSR generates the anomalies at the feature level by sampling the learned quantized feature space, which allows a controlled generation of near-in-distribution anomalies. DSR achieves state-of-the-art results on the KSDD2 and MVTec anomaly detection datasets. The experiments on the challenging real-world KSDD2 dataset show that DSR significantly outperforms other unsupervised surface anomaly detection methods, improving the previous top-performing methods by 10% AP in anomaly detection and 35% AP in anomaly localization. Code is available at: https://github.com/VitjanZ/DSR_anomaly_detection.

Keywords: Surface anomaly detection, discrete feature space, simulated anomaly generation

1 Introduction

Surface anomaly detection addresses localization of image regions that deviate from normal object appearance. This is a fundamental problem in industrial inspection, in which the anomalies are defects on production line objects. In the most challenging situations, the distribution of the normal appearance of the inspected objects is very close to the distribution of anomaly appearances, while anomalies often occupy only a small portion of the object. Furthermore, the anomalies are rare in practical production lines, making the acquisition of a suitable data set for training supervised methods infeasible. The methods thus focus on leveraging only anomaly-free images, since these can be abundantly obtained.

Most anomaly detection approaches are based on computing the difference between the inspected image and its image-level or feature-level reconstruction [1,2,18,5,22,3], with their reconstruction method trained only on anomaly-

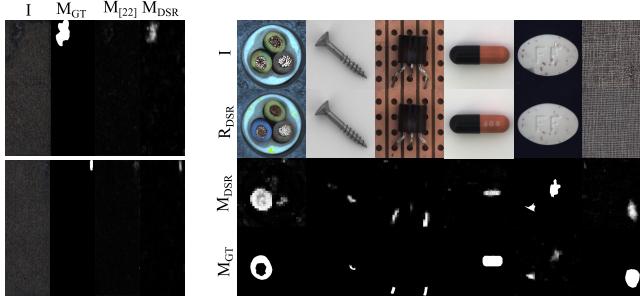


Fig. 1. The dual decoder architecture with discrete feature space allows DSR robust object-specific reconstruction (R_{DSR}) and accurate detection of near-in-distribution anomalies, which present a considerable challenge for the recent state-of-the-art ($M_{[22]}$).

free images. These approaches assume that anomalies will be poorly reconstructed since they have never been observed during training and that the reconstruction failure on anomalies can be well detected by L_2 or SSIM [20] difference with the input image.

However, L_2 and SSIM measures can only detect anomalies that differ substantially from normal appearance. Subsequent works have addressed this problem by either learning the distance measure with a discriminative network [21] or by classifying the anomalies directly on the input image [11]. These methods require annotated anomalies at training time, and resort to simulation of anomalies from auxiliary datasets and copy-pasting and blending them with the anomaly-free training images. While these methods by far outperform the reconstruction-only methods, they rely substantially on the quality of the auxiliary dataset and the simulation process quality; their performance still degrades on near-in-distribution anomalies (Figure 1) since it is difficult to simulate these realistically.

In this paper we address two drawbacks of the surface anomaly detection state-of-the-art: the reliance on the auxiliary anomaly simulation datasets and poor near-in-distribution anomaly detection. We propose a dual subspace re-projection surface anomaly detection network (DSR). The network leverages the framework of discretized latent feature space image representation [16], and jointly learns a general and a normal-appearance-specific subspace re-projection to emphasize the anomaly detection capability.

The proposed architecture avoids reliance on auxiliary anomaly datasets in training anomaly discrimination. We propose a new anomaly simulation technique that generates the anomalies directly from the network’s discretized latent space of natural images, leading to significant performance improvements on near-in-distribution anomaly detection (Figure 1).

Our contribution is thus twofold: (i) the dual image reconstruction branch architecture with discretized latent representation and (ii) the latent space anomaly generation method that leverages the learned representation of natural images.

DSR substantially outperforms the state-of-the-art in near-in-distribution anomaly detection on the recent challenging KSDD2 [6] and delivers state-of-the-art performance on the standard MVTec anomaly detection dataset [3].

2 Related Work

Many recent surface anomaly detection methods are based on image reconstruction [1,2,5], where an encoder-decoder network is trained for image reconstruction on anomaly free images. The anomaly detection capability of these methods is based on the assumption that the trained networks will be unable to accurately reconstruct anomalous regions due to never seeing them during training, making anomalies detectable by comparing the input image to its reconstruction. The core assumption often does not hold, especially in more diverse datasets, as the reconstruction networks learn to generalize well which enables accurate anomaly reconstruction, hampering the downstream anomaly detection performance. Similarly, in [22] iterative inpainting is used for image reconstruction, however, the method is sensitive to random pattern regions that are difficult to inpaint and which cause false positive detections.

Several recent approaches utilize the ability of pretrained networks to extract informative features from an image. In [4] a network is trained on anomaly free images to reconstruct features extracted by the pretrained network. In [19] a feature map is generated as a concatenation of several layers of a pretrained network. An auto-encoder is then trained to reconstruct the resulting feature map. As with image reconstruction based anomaly detection methods, the networks in [4,19] are assumed to be unable to reconstruct the features extracted by the pretrained networks. In [17], a pretrained network is also used to extract informative features from anomaly free data, however instead of relying on reconstruction for anomaly detection, a multivariate gaussian is fit to the anomaly-free data. A Mahalanobis distance is then used as an anomaly score. A similar approach is used in [7], however, a Gaussian is fit at each location in the feature map.

Discriminative unsupervised anomaly detection methods [21,12,11] utilize synthetically generated anomalies to train a discriminative anomaly detection network. In order to alleviate overfitting on the synthetic anomaly appearance, in [21] a reconstruction network is used to restore the normal appearance of the synthetic anomalies. The discriminative network then learns a distance function between the original image and its reconstruction to perform anomaly detection. Due to a limited distribution of the generated synthetic anomalies, the reconstruction network may overfit to the synthetic appearance and fail to restore normality in near-distribution anomalies, leading to a reduction in performance in downstream anomaly detection.

3 DSR

The DSR architecture is outlined in Figure 2. The input image is projected into a quantized latent space (Section 3.1) and then decoded in parallel by two decoders

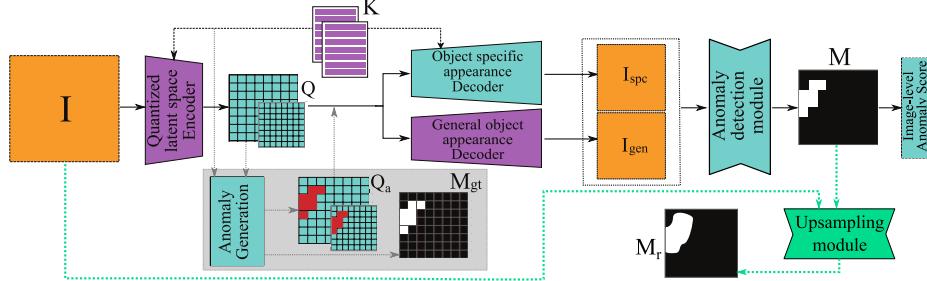


Fig. 2. The DSR architecture. During training, the non-anomalous image quantized feature maps (Q_{hi} , Q_{lo}) are replaced by the anomaly augmented feature maps ($Q_{a,hi}$, $Q_{a,lo}$) generated by the latent space sampling procedure (shaded block). The pathway marked with green arrows are used when training the Upsampling module with simulated smudges and at inference.

specialized on different appearance subspaces. One decoder, the general object appearance decoder (Section 3.2), is trained for high-fidelity reconstruction of arbitrary natural images, while the second, object-specific decoder (Section 3.3), is restricted to reconstructing only normal local appearances of the selected object. The two reconstructed images are then analyzed by the anomaly detection module (Section 3.4). The output localization map is at feature resolution and is upsampled to the input image resolution by the Upsampling module (Section 3.5).

DSR is trained by a novel technique of sampling the quantized latent feature space for generating near-in-distribution anomalies (Section 3.6). The DSR training regime is explained in Section 3.7.

3.1 Quantized latent space encoder

DSR leverages quantized feature space representation, which has recently demonstrated strong modelling capabilities of complex natural image distributions ,e.g., [9,15]. The approach is based on quantizing the extracted features with features from a codebook \mathbf{K} which has been trained for optimal decoding of spatial configurations of quantized features into high-fidelity images.

In particular, the quantized latent space encoder module in Figure 2 accepts the input image \mathbf{I} and projects it into a feature space \mathbf{F} using a Resnet-based encoder. A quantized feature representation \mathbf{Q} of the input image is obtained by replacing each feature vector \mathbf{F}_{ij} with its nearest neighbor \mathbf{e}_l in \mathbf{K} , i.e,

$$Q_{i,j} = q(F_{i,j}) = \operatorname{argmin}_{\mathbf{e}_l \in \mathbf{K}} (\|\mathbf{F}_{i,j} - \mathbf{e}_l\|). \quad (1)$$

In the following, we refer to this operation as vector quantization (VQ). Note that the input image is encoded at two levels of detail using low- and high-resolution codebooks (\mathbf{K}_{lo} , \mathbf{K}_{hi}), producing \mathbf{Q}_{lo} and \mathbf{Q}_{hi} . The two-level VQ has recently

been reported to produce superior reconstructions [16]. The architecture of the quantized latent space encoder is shown in Figure 3. The quantized feature maps Q_{hi} and Q_{lo} produced by the quantized latent space encoder are $4\times$ and $8\times$ smaller than the original input image, respectively.

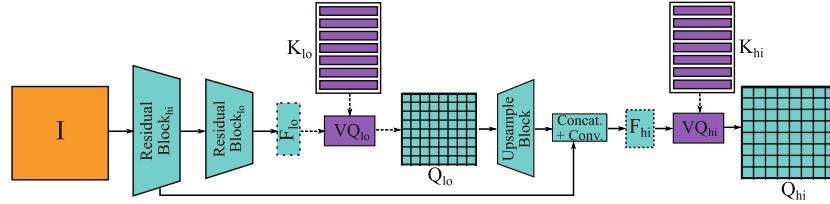


Fig. 3. The quantized latent space encoder architecture. Two residual blocks extract image features at different spatial resolutions. The low-resolution features F_{lo} are quantized by the codebook K_{lo} into a feature representation Q_{lo} . The high-resolution features are concatenated by upsampled Q_{lo} , followed by a convolutional block and quantized by the codebook K_{hi} into a high-resolution representation Q_{hi} .

3.2 General object appearance decoder

The subspace of VQ encoded natural images is captured by specific *spatial configurations* of quantized feature vectors. We apply a *general object appearance decoder* to learn decoding of these configurations into image reconstructions. The decoder first upsamples the low resolution Q_{lo} and concatenates it with the Q_{hi} , which is followed by two ResNet blocks and two transposed convolution upsampling blocks that map into the reconstructed image I_{gen} .

3.3 Object-specific appearance decoder

The tasks of the second decoder, the *object-specific appearance decoder* (see Figure 4) is to restore local visual anomalies into feasible normal appearances of the object instances observed during training. In particular, we would like to restrict the appearance subspace, i.e., the allowed spatial VQ feature configurations, into configurations that agree with normal appearances. This is achieved by a *subspace restriction module* (Figure 4), which transforms both high- and low-resolution general input VQ representations ($\mathbf{Q}=\{Q_{hi}, Q_{lo}\}$) into non-quantized object-specific subspace configurations ($\tilde{\mathbf{F}}=\{\tilde{\mathbf{F}}_{lo}, \tilde{\mathbf{F}}_{hi}\}$). This is followed by a VQ projection (with codebooks $\mathbf{K}=\{K_{hi}, K_{lo}\}$) into object subspace-restricted quantized feature configurations ($\tilde{\mathbf{Q}}=\{\tilde{Q}_{hi}, \tilde{Q}_{lo}\}$). The quantized representations are concatenated (the low-resolution representation is upsampled first) and decoded into reconstructed anomaly-free image I_{spc} using a convolutional decoder. The subspace restriction modules (Figure 4) are encoder-decoder convolutional networks with three downsampling and corresponding upsampling blocks. The image reconstruction network consists of two downsampling convolutional blocks,

followed by four transposed convolution upsampling blocks. Examples of images with anomalies present and their anomaly-free reconstructions are shown in Figure 5.

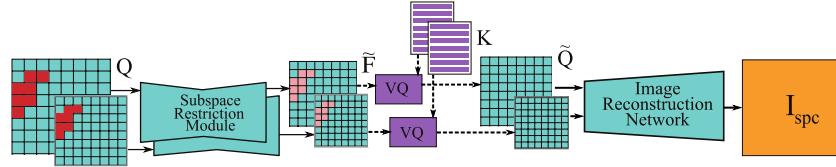


Fig. 4. The object-specific appearance decoder architecture. The features in the quantized features maps \mathbf{Q} are reduced by a subspace restriction module into non-quantized features $\tilde{\mathbf{F}}$ that are then vector-quantized (VQ) by the codebooks \mathbf{K} into $\tilde{\mathbf{Q}}$, which are then decoded into an anomaly-free image \mathbf{I}_{spc} .

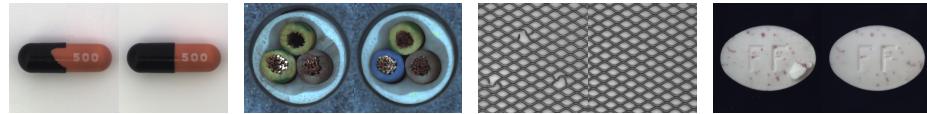


Fig. 5. The object-specific decoder reconstructs the anomalous images (left) into anomaly-free images (right) with remarkable fidelity. In the second example, it even correctly reconstructs the anomalous green wire into blue.

3.4 Anomaly detection module

The purpose of the anomaly detection module is to localize the anomaly by inspecting the input image reconstruction generated by the general object appearance decoder (\mathbf{I}_{gen}) and the object-specific appearance decoder (\mathbf{I}_{spc}). The reconstructed images are concatenated depth-wise and decoded into a segmentation mask \mathbf{M} by a Unet-based architecture. \mathbf{M} is the output anomaly map indicating the pixel-level location of the anomalies in the image. To compute also the image-level anomaly score, we apply a simple segmentation mask interpretation procedure as in [22] – the segmentation mask is smoothed by a 21×21 averaging filter and globally max-pooled into a single score.

3.5 Upsampling module

The generated segmentation mask \mathbf{M} is of the same resolution as the feature maps. A simple Unet-like upsampling module is thus used to resample the mask to full resolution. The input to the network is a depth-wise concatenation of the input image \mathbf{I} and bilinearly upsampled mask \mathbf{M} . The output is the final full-resolution mask \mathbf{M}_r (Figure 2).

3.6 Feature-space surface anomaly generation

The purpose of anomaly generator is to simulate near-in-distribution anomalies of various shapes and sizes with diverse visual appearances to (i) learn normal appearance subspace restriction in the object-specific appearance decoder (Section 3.3) and (ii) to specialize the anomaly detection module (Section 3.4) for detection of potentially gentle appearance deviations of anomalies from a diverse within-class normal appearance.

We propose a method that leverages the learned quantized subspace in DSR to generate such training anomalies as follows. An anomaly-free input image \mathbf{I} is encoded into a VQ subspace representation \mathbf{Q} and an anomaly mask \mathbf{M}_{gt} is generated by sampling a Perlin noise [14], with values 1 indicating anomalous pixels. The features in \mathbf{Q} corresponding to the anomaly indicators in \mathbf{M}_{gt} are replaced by sampling from the set of codebook features \mathbf{K} .

Sampling without constraints will likely lead to significant appearance changes on anomalous pixels, resulting in trivial out-of-distribution reconstructed images. On the other hand, if the closest vectors are sampled that are too similar to the normal appearance, they will likely lead to false-positive detections.

We thus first define a similarity bound on all features for a given image $n \sim \mathcal{U}[\lambda_s N_K, N_K]$, where N_K is the number of codebook vectors and λ_s is the similarity bound parameter. Then we replace each feature in \mathbf{Q} , indicated by \mathbf{M}_{gt} , with one of its near neighbors from the codebook feature vectors, sampled uniformly, i.e., $k \sim \mathcal{U}[\lambda_s N_K, n]$. In all experiments λ_s is 0.05, therefore, excluding the 5% of most similar vectors to prevent false-positive generation, while prioritizing the selection of the features close to the vector to be replaced to encourage the generation of near-in-distribution anomalies.

3.7 DSR training procedure

The DSR is trained in three stages. In the **first stage**, the quantized latent space encoder (Section 3.1), along with the VQ codebook and the general object appearance decoder are trained on ImageNet [8] to learn the subspace of natural images, which allows a high-fidelity reconstruction of general images. We use the procedure from [13] that minimizes the image reconstruction loss as well as the difference between the feature space projection \mathbf{F} computed in the quantized latent space encoder (Figures 2 and 3) and its quantized version \mathbf{Q} , i.e,

$$\mathcal{L}_{st1} = L_2(\mathbf{I}, \mathbf{I}_{gen}) + L_2(sg[\mathbf{F}], \mathbf{Q}) + \lambda_1 L_2(\mathbf{F}, sg[\mathbf{Q}]), \quad (2)$$

where $L_2(\cdot)$ is the Euclidean distance and $sg[\cdot]$ is the stop gradient operator constraining the operand to be a non-updated constant [13]. After training, the discrete latent space encoder, the codebook and the general object appearance decoder (coloured in magenta in Figure 2) are fixed.

In the second, **anomaly detection training stage**, the detection parts of DSR are trained on images of the selected object type. Anomaly-free training images are projected through the quantized latent space encoder into their

quantized feature representation \mathbf{Q} . The surface anomaly generation method presented in Section 3.6 then generates the anomalies at the feature level, \mathbf{Q}_a , along with their ground truth masks \mathbf{M}_{gt} . The representation \mathbf{Q}_a (that replaces \mathbf{Q} in Figure 2) is then forward passed and decoded into the anomaly mask \mathbf{M} .

The anomaly detection module and the object-specific appearance decoder are trained by minimizing the focal loss between the ground truth \mathbf{M}_{gt} and predicted anomaly mask \mathbf{M} , the L_2 distance between the subspace-restricted configurations $\tilde{\mathbf{F}}$ computed in the object-specific appearance decoder (Figure 4) and the non-anomalous input image quantized representation \mathbf{Q} , and the L_2 distance between the non-anomalous input image \mathbf{I} and its object-specific reconstruction \mathbf{I}_{spc} , i.e.,

$$\mathcal{L}_{st2} = \mathcal{L}_{foc}(\mathbf{M}_{gt}, \mathbf{M}) + \lambda_2 L_2(\tilde{\mathbf{F}}, \mathbf{Q}) + \lambda_3 L_2(\mathbf{I}, \mathbf{I}_{spc}). \quad (3)$$

Finally, the Upsampling module is trained. After the anomaly detection network has been trained, images with copy-pasted smudges are generated and the low-resolution anomaly masks are computed as detection network outputs. Since the full-resolution masks are known from the smudge pasting on the original image, the upsampling network can be trained with a focal loss.

4 Experiments

4.1 Implementation details

In the first training stage (Section 3.7) quantized latent space encoder, general object appearance decoder and the latent vector codebook are trained for image reconstruction on ImageNet [8] for 200,000 iterations with a batch size of 32 and a learning rate of 0.0002. The codebooks K_{hi} and K_{lo} each contain 4096 latent vectors of dimension 128. In the second training stage (Section 3.7) the object-specific appearance decoder and the anomaly detection model are trained for 100,000 iterations with a batch size of 8 with a learning rate of 0.0002. The learning rate is decreased by a factor of 10 after 80,000 iterations. The λ_2 and λ_3 values in Eq.3 are set to 1 and 10, respectively. In the third stage the Upsampling module is then trained for 20,000 iterations with a learning rate of 0.0002 and a batch size of 8. The training hyperparameters and network architectures remain constant throughout our experiments.

4.2 Experimental results on KSDD2

The recently proposed KSDD2 [6] surface anomaly detection dataset is currently the most challenging dataset with near-in-distribution surface anomalies (see Figure 6). KSDD2 was acquired on real industrial production lines and contains a wide variety of anomalies, many of which are particularly challenging due to their similarity to the normal appearances in the training set.

The dataset contains 2085 anomaly-free and 246 anomalous training images to test anomaly detection under unsupervised and supervised setups. For example, in the unsupervised setup, only anomaly-free training images may be

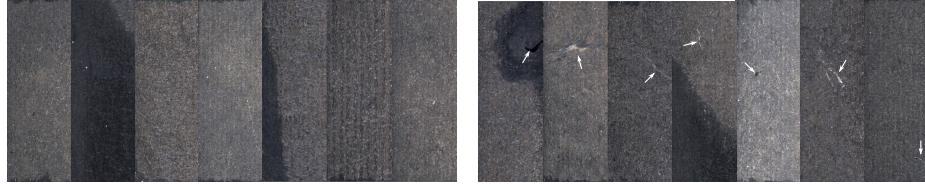


Fig. 6. Normal (left) and anomalous (right) images from KSDD2. The anomalies (marked with white arrows) are very similar to the normal appearance in this dataset.

used. The test set contains 894 anomaly-free and 110 anomalous images. We follow the evaluation procedure defined in [6], with the AP metric for image-level anomaly detection (AP_{det}), and present also anomaly localization results in terms of pixel-level AP (AP_{loc}).

Additional anomalous training samples available in KSDD2 [6] enable comparison of DSR with the supervised defect detection methods. Note that, although the acquisition of anomalous examples is difficult in practice, in many cases these are available, albeit in small quantities. Most of unsupervised anomaly detection methods, however, do not, or even can not, make use of this additional information, if available. The proposed DSR method can be easily adapted to take such annotations of anomalous images into account. In addition to the generated synthetic anomalies, the real-world anomalies with known ground truth can be utilised to train the anomaly detection module.

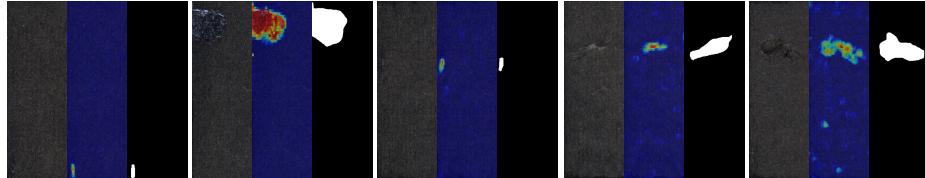


Fig. 7. Qualitative results of the unsupervised DSR on the KSDD2 dataset: the input image, the overlaid predicted mask and the ground truth.

Comparison with the state-of-the-art The results of unsupervised surface anomaly detection methods are shown in Table 1. DSR significantly outperforms other state-of-the-art unsupervised methods such as MAD [17], PaDim [7] and DRAEM [21], achieving higher AP scores for anomaly detection and localization. It outperforms the previous best image-level AP score by 7.9 p.p. Qualitative examples of the unsupervised DSR are presented in Figure 7 and show that despite a quite heterogeneous normal appearance, that can also be observed in Figure 6, the visual defects are successfully detected and localized.

Extension to supervised learning In contrast to most of the unsupervised visual anomaly detection methods, DSR can also utilise pixel-level annotations, if they are available, making it applicable in low-shot anomaly detection scenarios. We evaluated the proposed approach in the supervised setting. Table 2 shows the comparison with the method from [6] that is specifically designed for supervised defect detection and can operate with various levels of supervision. In KSDD2, there are 246 anomalous samples available, and if image-level labels are available but none of them is segmented, the method [6] operates in the weakly supervised mode ($N = 0$).

	Method US [4]	MAD [17]	DRÆM [21]	PaDim [7]	DSR
AP_{det}	65.3	79.3	77.8	55.6	87.2
AP_{loc}	-	-	42.4	45.3	61.4

Table 1. Anomaly detection (AP_{det}) and localization (AP_{loc}) on the KSDD2 dataset.

The completely unsupervised DSR, without taking these additional positive training samples into account, outperforms the weakly supervised method [6] for 13.9 p.p. in AP_{det} and achieves a good localization result, whereas [6] is unable to produce meaningful segmentation maps. When a number of anomalous training images are also segmented ($N > 0$), the results improve even further, and significantly outperform the results of [6]. When using the full annotated training set of 246 segmented examples, DSR achieves anomaly detection performance near that of the fully supervised method [6] and outperforms the recent fully supervised method presented in [10]. Furthermore, DSR achieves the highest localization performance AP_{loc} under all training settings, significantly outperforming [6]. These results demonstrate that DSR can efficiently work in supervised settings as well, utilising all information available.

	Method	$N:$	0	16	53	246
AP_{det}	[6]		73.3	83.2	89.1	95.4
	DSR		87.2	91.4	94.6	95.2
	[10]		-	-	-	93.3
AP_{loc}	[6]		1.0	45.1	52.2	67.6
	DSR		61.4	71.2	81.6	85.5

Table 2. Anomaly detection and localization on the KSDD2 dataset in a supervised settings w.r.t. number of used anomalous training images N with ground truth masks.

4.3 Experimental results on MVTec

We also perform experiments on the MVTec anomaly detection dataset [3], which contains 15 different texture and object classes, and has been established as the standard surface anomaly evaluation benchmark [11,7,21,17]. The training set

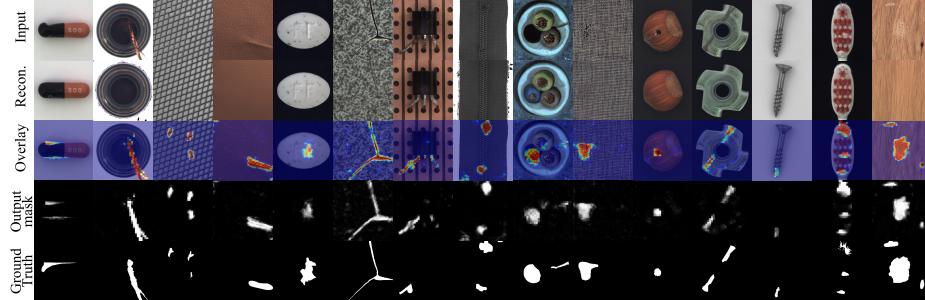


Fig. 8. Qualitative results on the MVTec dataset: the input images, outputs of the reconstruction network, input images with overlaid output masks, the output masks and the ground truth masks are shown in individual rows.

consists only of anomaly-free images, while the test set is comprised of anomalous as well as anomaly-free images. The widely used AUROC metric is applied for image-level anomaly detection. Because only a fraction of the pixels in the test set are anomalous, the pixel-wise average-precision metric AP [21] is used to evaluate the anomaly localization performance. It is more robust to class imbalance and better suited for anomaly localization evaluation than the commonly used pixel-wise AUROC.

Method	bottle	capsule	grid	leather	pill	tile	trans.	zipper	cable	carpet	hazelnut	m.	nut	screw	toothbrush	wood	average
[4]	99.0	86.1	81.0	88.2	87.9	99.1	81.8	91.9	86.2	91.6	93.1	82.0	54.9	95.3	97.7	87.7	
[22]	99.9	88.4	99.6	100	83.8	98.7	90.9	98.1	81.9	84.2	83.3	88.5	84.5	100	93.0	91.7	
[17]	100	92.3	92.9	100	83.3	97.4	95.9	97.9	94.0	95.5	98.7	93.1	81.2	95.8	97.6	94.4	
[7]	99.8	91.5	95.7	100	94.4	97.4	97.8	90.9	92.2	99.9	93.3	99.2	84.4	97.2	98.8	95.5	
[11]	98.2	98.2	100	100	94.9	94.6	96.1	99.9	81.2	93.9	98.3	99.9	88.7	99.4	99.1	96.1	
[21]	99.2	98.5	99.9	100	98.9	99.6	93.1	100	91.8	97.0	100	98.7	93.9	100	99.1	98.0	
DSR	100	98.1	100	100	97.5	100	97.8	100	93.8	100	95.6	98.5	96.2	99.7	96.3	98.2	

Table 3. Results of anomaly detection on MVTec dataset (AUROC) with the average score over all classes (*avg*) in the last column.

Comparison with the state-of-the-art We evaluate DSR against recent state-of-the-art surface anomaly detection methods. The experimental results are presented in Tables 3 and 4 for image-level anomaly detection and for pixel-level anomaly localization, respectively, and show that DSR outperforms the recent state-of-the-art approaches. It achieves an average AUROC of 98.2%, while maintaining a strong anomaly localization performance with an AP score of 70.2%. DSR outperforms the previous top performer in anomaly detection DRAEM [21] on the mean AUROC score by 0.2 percentage points (p.p.) and achieves superior performance on classes such as transistor, cable, carpet and screw, where near-distribution anomalies such as deformations are more prevalent. Qualitative results on the MVTec dataset are shown in Figure 8 and demonstrate that

the detected anomalous regions resemble the ground truth anomaly maps to a high degree.

Method	bottle	capsule	grid	leather	pill	tile	trans.	zipper	cable	carpet	hazelnut	m.	nut	screw	toothbrush	wood	average
[4]	74.2	25.9	10.1	40.9	62.0	65.3	27.1	36.1	48.2	52.2	57.8	83.5	7.8	37.7	53.3	45.5	
[22]	76.4	38.2	36.4	49.1	51.6	52.6	39.2	63.4	24.4	61.4	33.8	64.3	43.9	50.6	38.2	48.2	
[7]	77.3	46.7	35.7	53.5	61.2	52.4	72.0	58.2	45.4	60.7	61.1	77.4	21.7	54.7	46.3	55.0	
[21]	86.5	49.4	65.7	75.3	48.5	92.3	50.7	81.5	52.4	53.5	92.9	96.3	58.2	44.7	77.7	68.4	
DSR	91.5	53.3	68.0	62.5	65.7	93.9	41.1	78.5	70.4	78.2	87.3	67.5	52.5	74.2	68.4	70.2	

Table 4. Results (AP) of anomaly localization on MVTec dataset.

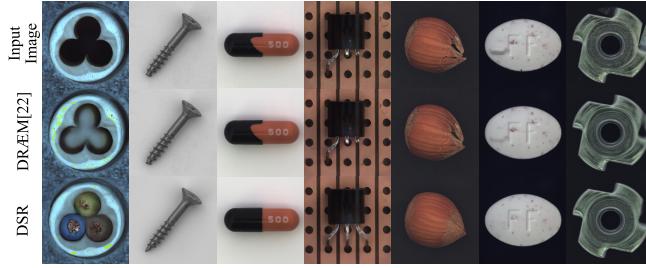


Fig. 9. Qualitative reconstruction results. DSR far better reconstructs the anomalies by their corresponding in-distribution appearance than the state-of-the-art DRÆM [21].

Further qualitative comparison is shown in Figure 9. Both DSR and DRÆM [21] are trained to restore the normality of images corrupted by simulated anomalies. However, DRÆM [21] generates the anomalies in the image space from an out-of-distribution dataset, while DSR generates the anomalies within the quantized feature space, making it more difficult for the reconstruction network to overfit to the simulated anomaly appearance. This results in DSR having a more robust image normality restoration capability that is insensitive to near-in-distribution anomalies such as deformations. Figure 9 compares the reconstruction results of both methods. Note that DSR produces a more realistic reconstruction. Due to the reconstruction network not overfitting to the synthetic anomaly appearance, DSR can recognize deformations as deviations from normality and reconstruct them accordingly.

4.4 Ablation Study

Additional experiments on the MVTec dataset provide further insights into DSR.

Anomaly source evaluation. We evaluate the impact of training using our feature-based anomalies and compare it to the training with image-based

anomalies generated from out-of-distribution datasets as proposed in [21]. The results are shown in Table 6(a), where DSR_{img} denotes the results when using image-based anomalies during training. We can see that feature-based anomaly generation (DSR) outperforms the image-based approach in anomaly localization and detection.

Anomaly feature sampling. To generate simulated anomalies, DSR samples vectors from the codebook K according to a designed process presented in Section 3.6. We compare this approach to one that samples the codebook with a uniform probability for each vector. The results (DSR_{random}) presented in Table 6(b) show a significant drop in performance. The sampling method, therefore, plays an important role. The proposed approach generates anomalous regions from vectors that are close to the extracted vectors generating simulated near-distribution anomalies that leads to superior results in terms of anomaly detection as well as localization. DSR is robust to the choice of the similarity bound parameter, retaining a good anomaly detection performance for a wide range of λ_s values (Table 5).

λ_s	0.01	0.02	0.05	0.1	0.2	0.5
AUROC _{det}	97.5	98.0	98.2	97.7	97.1	95.5

Table 5. Results on MVTec using various similarity bound λ_s values.

Reconstruction loss components. Table 6(c) shows the impact of individual reconstruction loss components of Eq. (3), where the reconstruction is conditioned both on the feature L_{feat} and image L_{img} reconstruction losses, which are the second and the third components in Eq. (3), respectively. During training $\text{DSR}_{L_{img}}$ uses only the image reconstruction loss and $\text{DSR}_{L_{feat}}$ uses only the feature reconstruction loss. There is a significant loss in performance when using only individual loss components. Relying on both image and feature reconstruction losses during training results in a more robust normality reconstruction model leading to a 2 p.p. higher average image-level AUROC as well as significantly higher localization AP scores.

Upsampling module Table 6(d) shows the effect of removing the Upsampling module of the network, leading to a drop in localization performance. The anomaly detection performance remains the same as the image-level score is extracted from the lower resolution output anomaly map M .

5 Conclusion

We proposed DSR, a discriminative surface anomaly detection method based on dual image reconstruction branch architecture with discretized latent representation. Such representation allows controlled generation of synthetic anomalies in feature space, which, in contrast to the state-of-the-art methods that generate anomalies in image space, makes no assumption about the anomaly ap-

Experiment	(a)		(b)		(c)		(d)
	DSR	DSR _{img}	DSR _{random}	DSR _{L_{img}}	DSR _{L_{feat}}	DSR _{U-}	
AUROC _{det}	98.2	97.8	95.6	96.3	95.2	98.2	
AP _{loc}	70.2	67.5	62.5	67.0	61.8	65.2	

Table 6. Ablation study results on MVTec: (a) using out-of-distribution texture-based anomalies (DSR_{img}) in training; (b) unconstrained uniform anomaly sampling (DSR_{random}); (c) training with only image reconstruction loss (DSR_{L_{img}}) and with only the feature reconstruction loss (DSR_{L_{feat}}); (d) Performance without the Upsampling module (DSR_{U-}).

pearance and does not rely on image-level heuristics. Our anomaly generation approach produces near-in-distribution synthetic anomalies resulting in more robustly trained reconstruction capabilities and detection of anomalies whose appearance is close to the normal appearance.

On the recent challenging real-world KSDD2 dataset [6], it outperforms all other unsupervised surface anomaly detection methods by 10% AP in anomaly detection and 35% AP in anomaly localization tasks. DSR also significantly outperforms the weakly-supervised model presented in [6]. Moreover, we showed that the proposed approach can be extended to utilise also pixel-wise annotated anomalous training samples. When used in such supervised settings, it considerably improves the results of related supervised methods [6] when only a few annotated examples are available. This demonstrates the potential of DSR to be used in real-world settings where the number of available anomalous images is typically too low to train fully supervised methods. DSR also achieves state-of-the-art results on the standard MVTec anomaly detection dataset [3].

The ablation study shows that the sampling method used for anomaly generation impacts the final anomaly detection results significantly, which suggests that a more complex feature sampling method may improve results further, which we leave for our future research endeavours.

In addition to its good performance across several tasks and its ability to utilize anomalous training samples that are available in some practical scenarios, DSR is fast, running at 58 FPS, making it a good choice for real-world industrial applications with real-time requirements.

Acknowledgement *This work was supported by Slovenian research agency programs J2-2506, L2-3169, P2-0214. Vitjan Zavrtanik was supported by the Young researcher program of the ARRS.*

References

1. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: GANomaly: Semi-supervised anomaly detection via adversarial training. In: Asian Conference on Computer Vision. pp. 622–637. Springer (2018)
2. Akçay, S., Atapour-Abarghouei, A., Breckon, T.P.: Skip-GANomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In: 2019 In-

- ternational Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (jul 2019)
3. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: MVTec AD – A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9592–9600 (2019)
 4. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4183–4192 (2020)
 5. Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., Steger, C.: Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In: 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics theory and Applications. vol. 5, pp. 372–380 (2018)
 6. Božič, J., Tabernik, D., Skočaj, D.: Mixed supervision for surface-defect detection: from weakly to fully supervised learning. Computers in Industry **129**, 103459 (2021)
 7. Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: A patch distribution modeling framework for anomaly detection and localization. In: International Conference on Pattern Recognition. pp. 475–489. Springer (2021)
 8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
 9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12873–12883 (June 2021)
 10. Lei, L., Sun, S., Zhang, Y., Liu, H., Xu, W.: PSIC-Net: Pixel-wise segmentation and image-wise classification network for surface defects. Machines 2021 **9**, 221 (2021). <https://doi.org/10.3390/MACHINES9100221>, <https://www.mdpi.com/2075-1702/9/10/221>
 11. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9664–9674 (2021)
 12. Lv, C., Shen, F., Zhang, Z., Xu, D., He, Y.: A novel pixel-wise defect inspection method based on stable background reconstruction. IEEE Transactions on Instrumentation and Measurement **70**, 1–13 (2021). <https://doi.org/10.1109/TIM.2020.3038413>
 13. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural discrete representation learning. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30, pp. 6306–6315. Curran Associates, Inc. (2017), <https://proceedings.neurips.cc/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf>
 14. Perlin, K.: An image synthesizer. ACM Siggraph Computer Graphics **19**(3), 287–296 (1985)
 15. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092 (2021)
 16. Razavi, A., van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F.,

- Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32, pp. 14866–14876. Curran Associates, Inc. (2019), <https://proceedings.neurips.cc/paper/2019/file/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Paper.pdf>
- 17. Rippel, O., Mertens, P., Merhof, D.: Modeling the distribution of normal data in pre-trained deep features for anomaly detection. ICPR (2020)
 - 18. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis **54**, 30–44 (2019)
 - 19. Shi, Y., Yang, J., Qi, Z.: Unsupervised anomaly segmentation via deep feature reconstruction. Neurocomputing **424**, 9–22 (2021)
 - 20. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
 - 21. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8330–8339 (October 2021)
 - 22. Zavrtanik, V., Kristan, M., Skočaj, D.: Reconstruction by inpainting for visual anomaly detection. Pattern Recognition **112**, 107706 (2021). <https://doi.org/https://doi.org/10.1016/j.patcog.2020.107706>, <https://www.sciencedirect.com/science/article/pii/S0031320320305094>