Cheating Depth: Enhancing 3D Surface Anomaly Detection via Depth Simulation

Vitjan Zavrtanik Matej Kristan Danijel Skočaj
Faculty of Computer and Information Science, University of Ljubljana
{vitjan.zavrtanik, matej.kristan, danijel.skocaj}@fri.uni-lj.si

Abstract

RGB-based surface anomaly detection methods have advanced significantly. However, certain surface anomalies remain practically invisible in RGB alone, necessitating the incorporation of 3D information. Existing approaches that employ point-cloud backbones suffer from suboptimal representations and reduced applicability due to slow processing. Re-training RGB backbones, designed for faster dense input processing, on industrial depth datasets is hindered by the limited availability of sufficiently large datasets. We make several contributions to address these challenges. (i) We propose a novel Depth-Aware Discrete Autoencoder (DADA) architecture, that enables learning a general discrete latent space that jointly models RGB and 3D data for 3D surface anomaly detection. (ii) We tackle the lack of diverse industrial depth datasets by introducing a simulation process for learning informative depth features in the depth encoder. (iii) We propose a new surface anomaly detection method 3DSR, which outperforms all existing state-of-the-art on the challenging MVTec3D anomaly detection benchmark, both in terms of accuracy and processing speed. The experimental results validate the effectiveness and efficiency of our approach, highlighting the potential of utilizing depth information for improved surface anomaly detection. Code is available at: https://github.com/VitjanZ/3DSR

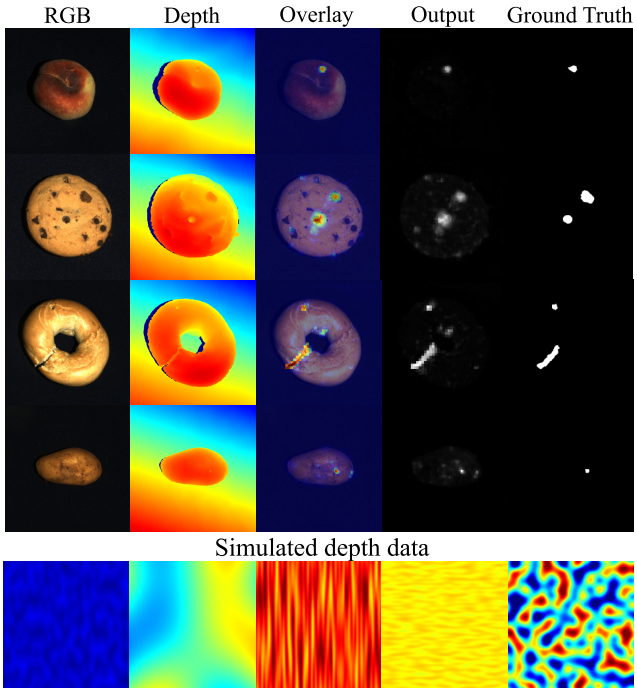

Figure 1. Certain anomalies are practically imperceptible in RGB, requiring depth for precise detection. Parameterized generative model yields images sufficiently describing depth statistics for training general depth-reconstruction backbones.

1. Introduction

Surface anomaly detection addresses localization of image regions that deviate from normal object appearance. Most of the works consider the problem of RGB-based detection, which has witnessed remarkable progress in recent years, with several methods approaching perfection on the widely adopted MVTec anomaly detection dataset [1]. However, since certain surface anomalies in practical applications are not detectable in RGB (Figure 1), recent works [2, 3, 9, 17] consider a new research problem of RGB+3D anomaly detection.

The state-of-the-art methods typically rely on features extracted by general backbone networks [14, 15, 17, 20], pretrained on large datasets. The current state-of-the-art 3D anomaly detection method M3DM [19] applies two backbones, one pretrained on RGB, the other on point-cloud datasets. However, the general point-cloud datasets do not represent well the depth appearance distribution of industrial setups, leading to suboptimal representations. Furthermore, the point-cloud backbone substantially slows down processing, reducing the method's practicality. The inference could be sped up by considering depth image as a grayscale image and replacing the point-cloud back-

bone with an RGB-pretrained one, but evidence shows [9] that RGB-pretrained backbones insufficiently represent the depth properties relevant for anomaly detection.

Alternatively, RGB backbones could be re-trained on industrial depth datasets, but the current industrial depth datasets are too small to efficiently train the large backbones. Recent work DSR [22] proposed utilizing Vector-quantized autoencoders (VQVAE) [13], which learn only a fixed number of discrete latent representation vectors, thus potentially enable learning from smaller datasets. Nevertheless, our experience shows that training DSR on the available industrial depth dataset MVTec3D [2] leads to suboptimal results, indicating that the existing data is too small even for the representation-efficient VQVAEs. Advances in RGB+3D surface anomaly detection are thus hindered by the lack of sufficiently large datasets that would enable pretraining general depth backbones and allow development of methods fast enough for practical applications.

We address the aforementioned issues by making an insight that it is possible to summarize statistical properties of spatial and intensity content typical for industrial surface anomaly inspection depth data, which allows the creation of a parameterized generative models for such data. We hypothesize that such a model can then be used to generate a large training set for pre-training a depth-specific backbone (Figure 1, bottom). Furthermore, we note that the processing time constraint requires efficient architecture design for encoding of the RGB and depth data to jointly exploit both modalities for detecting complex anomalies.

The main contributions of this work are three-fold: (i) A novel Depth-Aware Discrete Autoencoder (*DADA*) architecture is proposed, that enables learning a general discrete latent space that jointly models RGB and depth data for 3D surface anomaly detection. (ii) We directly address the lack of a diverse industrial depth dataset by proposing an industrial depth data simulation process that facilitates the learning of informative depth features by the DADA module. (iii) We demonstrate the effectiveness of the learned feature space by proposing *3DSR*, a novel discriminative 3D surface anomaly detection method that significantly outperforms all competing state-of-the-art methods on the most challenging MVTec3D anomaly detection benchmark [2]. Owing to its efficient design, 3DSR also outperforms the current state-of-the-art [19] in speed by by an order of magnitude.

2. Related work

The MVTec anomaly detection dataset [1] has been widely adopted for unsupervised surface anomaly detection research. The training dataset comprises only normal cases, while the testing dataset includes both normal and anomalous instances. Most top performing anomaly detection methods [5, 14, 15, 17, 20] rely on strong pretrained backbones for informative feature extraction. After obtaining a good representation of each anomaly-free image in the training set a simple statistical model is typically built that tightly binds the anomaly-free feature space [5, 14]. This enables the detection of anomalies based on a chosen distance function to the anomaly-free representation model. Flow-based methods that utilize pretrained feature extractors to train a flow model also achieve state-of-the-art results [15–17, 21]. Certain discriminative methods [10, 22, 23] do not need a strong pretrained backbone but instead rely on simulated anomalies using an out-of-distribution dataset to build a strong classifier that generalizes well to real anomalies at test-time. It has been shown that top-performing RGB anomaly detection methods do not generalize well to 3D surface anomaly detection [2, 9, 17].

In the problem setup of 3D surface anomaly detection, the MVTec-3D anomaly detection dataset [2] is the most comprehensive dataset containing RGB and 3D information. In [2] reconstruction-based anomaly detection methods have been applied to the 3D anomaly detection problem as an initial baseline. In [3], a 3D student-teacher network was proposed that focuses on point cloud geometry descriptors. In [9] a memory-based method akin to [14] was proposed, that utilizes pretrained backbone features for RGB and FPFH [18] features for 3D representation. In [17] a teacher-student flow-based model is proposed that uses pretrained backbone features for RGB data but raw depth pixel values are used for 3D representation. In [19] 3D features are extracted by a point transformer [24] pretrained on a general point cloud dataset not adapted to industrial 3D data. A lack of 3D data from the industrial domain and the lack of strong domain-specific feature extractors presents a significant challenge since better 3D representations are required for improving the accuracy of 3D surface anomaly detection methods.

3. Our approach: 3DSR

We propose a novel 3D surface anomaly detection method based on dual subspace reprojection (3DSR). The input image is encoded into a discrete feature space and is then reprojected into image space by two decoders. The object-specfic decoder and the general object decoder reconstruct the anomaly-free and the anomalous appearance, respectively. The anomaly detection module then segments potential anomalies based on the difference between the two reprojections. 3DSR is trained in two stages. First, a novel Depth-Aware Discrete Autoencoder (DADA) is trained on RGB and depth image pairs to learn a general joint discrete representation of RGB+3D depth data.

In the second stage, DADA is integrated into the DSR [22] surface anomaly detection framework, producing 3DSR, which is then trained on 3D anomaly detection

datasets [2, 4]. In Section 3.1 the architecture of the proposed Depth-Aware Discrete Autoencoder (DADA) module is described. The industrial depth data simulation process is then described in Section 3.2. The final 3DSR surface anomaly detection pipeline is described in Section 3.3.

3.1. Depth-aware discrete Autoencoder

To learn a representation of both RGB and depth data, a naive approach may be to train a vector quantized autoencoder [13] with 4 input channels to represent both RGB and depth. This approach has some drawbacks due to the properties of both RGB and depth images. In Figure 2 an example of a cable gland is shown. A region of interest is marked with a green rectangle in I_{RGB} and I_D and is shown in more detail in RGB (T_{RGB}) and depth (T_D). T_{RGB} exhibits significant local variation due to shadows but barely any variation is visible in T_D. In depth images, even slight depth variations can be informative for defect detection so representing variations in T_D is vital. Discrete autoencoders are typically trained using the $L2$ loss, which is less sensitive to variation types in T_D, where values change minimally, but is sensitive to changes in T_{RGB}, where local gradients are higher. Subtle changes in T_D therefore contribute very little to the final loss leading to a higher emphasis on the reconstruction of RGB data during training.

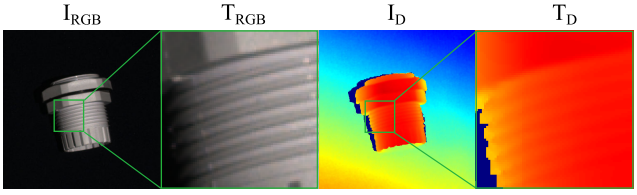

Figure 2. Example of a cable gland from the MVTec3D dataset [2]. Local variations are clearly visible in T_{RGB} and barely perceptible in T_D.

To ensure a good representation of both RGB and depth, a new Depth-Aware Discrete Autoencoder (DADA) architecture is required and is shown in Figure 3. The input I_{in} to the discrete autoencoder is a 4 channel tensor, a concatenation of an RGB image I and a depth image D. The encoder is a convolutional network, where RGB and depth information is separated by grouped convolution layers [8] which ensure that RGB and Depth features do not interact. This channel-wise separation is necessary to prevent the overwhelming influence of a single modality in the loss function.

To minimize the information loss due to discretization at a low spatial resolution, a two-stage discretization architecture is used [13]. First, the DADA Encoder 1 encodes the input and downsamples the spatial resolution by $4\times$ producing features f_1, where f_{I1} and f_{D1} stand for RGB and depth features, respectively. The second encoder stage, DADA Encoder 2, further downsamples the features to a total $8\times$ downsampling, producing f_2. f_2 features are quantized to the nearest neighbors of the codebook VQ_1 in terms of the L2 distance. The quantized features Q_1 are then input into a decoder module which upsamples the spatial resolution $2\times$ to f_U. The features f_1 and f_U are then concatenated and channel-wise reordered to group image features f_{I1} and f_{IU} and depth features f_{D1} and f_{Du}. This reordering of f_R is necessary to maintain the separation of RGB and depth features in the grouped convolutions of the decoder. The resulting features f_R are then quantized to nearest neighbor codebook vectors in VQ_2, producing Q_2. Q_1 is then upsampled to fit the spatial resolution of Q_2, after which Q_1 and Q_2 are input into the second decoder module which outputs the reconstructions of RGB I_o and depth D_o concatenated as I_{out}. We use the VQ-VAE [13] loss function, modified to address the added depth information, to train DADA:

$$\begin{aligned}\mathcal{L}_{\text{ae}} = \lambda_D L_2(\mathbf{D}, \mathbf{D}_\text{o}) + \lambda_I L_2(\mathbf{I}, \mathbf{I}_\text{o}) \\ + L_2(sg[\mathbf{f_2}], \mathbf{Q_1}) + \lambda_K L_2(\mathbf{f_2}, sg[\mathbf{Q_1}]) \\ + L_2(sg[\mathbf{f_1}], \mathbf{Q_2}) + \lambda_K L_2(\mathbf{f_1}, sg[\mathbf{Q_2}]),\end{aligned} \quad (1)$$

where $L_2(\cdot)$ is the Euclidean distance and $sg[\cdot]$ is the stop gradient operator. λ_I and λ_D are loss weighing factors and are both set to 1 in all experiments unless stated otherwise. λ_K controls the reluctance to change the codebook vectors corresponding to f_1 and is fixed to 0.25 in all experiments following [13].

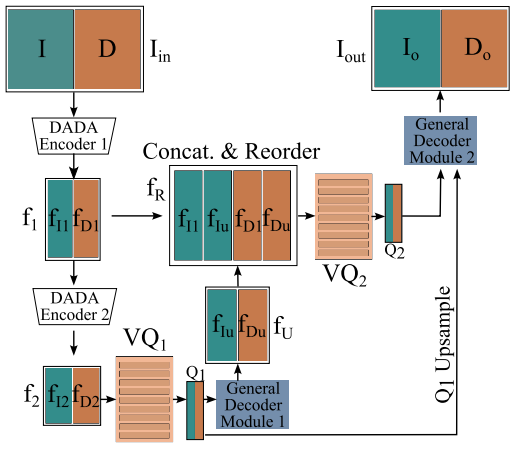

Figure 3. The Depth-Aware Discrete Autoencoder (DADA) module.

3.2. Depth data generative model

Simulated data is necessary for training DADA due to the absence of industrial depth datasets. Consideration of key properties of industrial depth data is required for an effective simulation process. Firstly, object depth can vary

continuously from closest to farthest from the sensor. Secondly, small dents and bumps can either cause significant intensity changes in RGB or are completely invisible, depending on the lighting. In depth, such minor changes are always detectable through a minimal local depth value alteration. Finally, the average of a depth image can vary significantly. Simulated data must capture local changes and variable average object depth in industrial images. The simulated depth image generation process is thus designed to explicitly address these properties. The core generator of the simulated images is the Perlin noise generator [12] that produces a variety of locally smooth textures that simulate the gradual changes in depth well, addressing the first property. Subtle local changes and varied average object distance are then simulated by adapting the Perlin noise image with a randomized affine transform.

To generate a single simulated depth image a Perlin noise image P is first generated and normalized between 0 and 1. P is then multiplied by a uniformly sampled $\alpha \in (0.0, 1.0)$ to produce P_α. At lower α values the maximum and minimum values of P_α will differ only slightly. To model the variation of the average object distance, P_α is translated with a uniformly sampled $\beta \in (0, 1-\alpha)$. The final simulated depth image D is therefore $D = \alpha P + \beta$.

Examples of simulated depth images with various α and β parameters are shown in Figure 4. α controls the difference between the minimum and maximum values of D, simulating local changes in depth and β controls the minimum value of D.

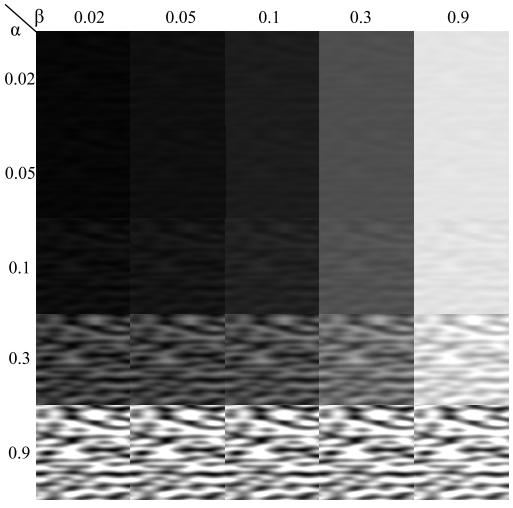

Figure 4. Impact of parameters α and β on the simulated depth maps. α is the scaling parameter and β is the translation parameter.

DADA is then trained on RGB and depth image pairs. RGB images are sampled from ImageNet [6] and depth images are simulated (Figure 4) and unrelated to the input RGB data.

3.3. 3D anomaly detection pipeline

In the second stage, DSR [22] is used as a discriminative anomaly detection framework. The VQ-VAE-2 [13] network that is used by DSR for RGB surface anomaly detection is replaced with DADA, pretrained to extract informative representations from both 3D and RGB data. Additionally, DADA's vector-quantized feature space enables efficient simulated anomaly sampling. The architecture of 3DSR is shown in Figure 5.

The DADA encoder modules extract and quantize the features Q_1 and Q_2. Then Q_1 and Q_2 are modified by the anomaly generation process to produce Q_{1A} and Q_{2A} that contains simulated feature-level anomalies. Q_{1A} and Q_{2A} are then input into the subspace restriction module that attempts to restore the extracted features to an anomaly free representation Q_{1S}, Q_{2S}. Note that the anomaly generation process is performed only during training so at inference Q_1 and Q_2 are directly input into the subspace restriction module since they may already contain an anomaly during testing.

For clarity, the steps that are only performed during training are marked with blue and steps only done at inference are marked with orange in Figure 5, while the trainable modules are marked with red in Figure 5. The Subspace restriction module is trained with an $L1$ loss to reconstruct anomaly-features Q_1 and Q_2 from Q_{1A} and Q_{2A}, respectively.

The Object specific decoder is trained to reconstruct the anomaly free appearance from the reconstructed features Q_{1S} and Q_{2S}. The pretrained general appearance decoder reconstructs the anomaly appearance I_G and D_G from Q_{1A} and Q_{2A} at training or Q_1 and Q_2 at inference. Then, I_G,D_G,I_S and D_S are concatenated and input into the Anomaly detection module. The Anomaly detection module is trained to localize the simulated anomalies during training and real anomalies at test time. It directly outputs an anomaly segmentation mask M_{out} and is trained using the Focal loss [11]. Following [22, 23], the image-level anomaly score is estimated by first smoothing M_{out} with a Gaussian filter and then taking the maximum value of the smoothed mask.

During training, anomalies are generated by modifying the quantized feature maps Q_1 and Q_2 as follows. First an anomaly map M is generated by thresholding and binarizing a Perlin noise map, following previous works [22, 23]. M is then resized to fit to the spatial dimensions of Q_1 and Q_2. Feature vectors of Q_1 and Q_2 in regions correspond to positive values of M are replaced with feature vectors randomly sampled from codebooks VQ_1 and VQ_2 respectively, generating modified feature maps Q_{1A} and Q_{2A} that contain simulated anomalies.

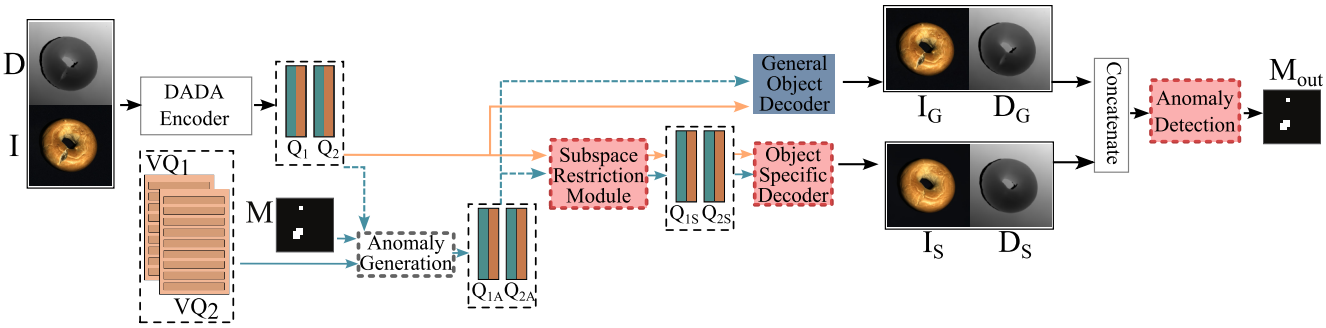

Figure 5. The architecture and second stage training process of 3DSR. The modules marked in red are trainable during the second stage of training.

4. Experiments

Datasets. 3DSR is evaluated on two recent 3D anomaly detection benchmarks, the MVTec3D dataset [2] and the Eyecandies dataset [4], each containing 10 different object classes. The MVTec3D [2] anomaly detection benchmark contains 4147 scans obtained by a high-resolution industrial 3D sensor that also acquires RGB data. Of the 4147 scans, 894 are anomalous containing various defects that are visible in either RGB or 3D data. 3DSR is also evaluated on the Eyecandies dataset [4], a difficult rendered dataset with RGB and 3D anomalies. It contains 10000 anomaly-free examples for training and 500 examples for testing of which 250 are anomalous.

Evaluation problem setup. The evaluation is divided into 3 different problem setups, where only the depth (3D setup), RGB images (RGB setup) or both the depth and RGB images (3D+RGB) are used. Since 3DSR is a 3D and RGB+3D method, we provide DSR [22] results instead in the RGB setup.

Evaluation metrics. Anomaly localization and image-level anomaly detection capabilities of each method are evaluated. For the image-level detection the standard image-level AUROC metric is used. For anomaly localization the PRO metric [1] is used. Additionally, the mean pixel-level AUROC is used for localization evaluation.

4.1. Implementation details

In the first stage of training, the DADA module is trained using the simulated depth data for 3D and the ImageNet dataset [6] for RGB supervision. DADA is trained using a batch size of 64 for $100K$ iterations using a learning rate of 0.0002. The vector quantization codebooks contains 2048 embeddings of dimension 256. In the second stage, 3DSR is trained on the MVTec3D dataset [2] or the Eyecandies dataset [4]. In both datasets the 3D data is given in the form of a sorted point cloud. The data is preprocessed by first normalizing the depth map to values between 0 and 1. Then, the missing values in each depth image are replaced with the average of all the valid pixels in a 3×3 neighborhood. If there are no surrounding valid pixels, the value is set to 0. A foreground mask is obtained by classifying points as either foreground or background by it's distance to the background plane. The background plane equation is obtained from valid points at the edges of the sorted point cloud. During training, anomalies are generated only on the foreground object. 3DSR is trained on an individual object class of each dataset as is standard for surface anomaly detection methods [14, 17, 22, 23]. It is trained with a batch size of 16 for $30K$ iterations with a learning rate of 0.0002.

4.2. Results

Anomaly detection results in terms of image-level AUROC are shown in Table 1. 3DSR significantly outperforms competing methods in the 3D settings, where only depth information is used. The 3D image-level AUROC is improved by approximately 5 percentage points. This demonstrates the informative depth representations learned by DADA and validates the 3DSR pipeline.

In classes where depth information is vital for anomaly detection, such as Tire and Foam, the improvement in the 3D setting is even greater. Additionally, 3DSR outperforms competing methods in the 3D+RGB setup by 3.3 percentage points demonstrating the ability of DSR to efficiently utilize information from both the depth and RGB modalities again emphasizing the powerful joint representations of 3D and RGB learned by DADA using the proposed data simulation process. In Table 1 the first, second and third best performing methods are marked for each object class. Note that 3DSR achieves first place on 7 out of 10 classes for the 3D setup and 5 out of 10 classes on the 3D+RGB setup, while staying in the top 3 in every object class.

Anomaly localization results are shown in Table 2 in terms of the AUPRO metric [1]. 3DSR achieves second place in segmentation results. It achieves comparable results to M3DM in the 3D problem setup while outperforming it in the RGB+3D setup. A comparison of 3DSR and state-of-the-art methods in terms of mean pixel-level

Table 1. Anomaly detection results on the MVTec3D dataset for the 3D, RGB and 3D+RGB problem setups. The results are listed as image-level AUROC scores (higher is better). The results of evaluated methods are ranked and the first, second and third place are marked.

	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
3D	Voxel AE [2]	69.3	42.5	51.5	79.0	49.4	55.8	53.7	48.4	63.9	58.3	57.1
	Depth GAN [2]	53.0	37.6	60.7	60.3	49.7	48.4	59.5	48.9	53.6	52.1	52.3
	Depth AE [2]	46.8	73.1②	49.7	67.3	53.4	41.7	48.5	54.9	56.4	54.6	54.6
	FPFH [9]	82.5	55.1	95.2	79.7	88.3③	58.2	75.8	88.9	92.9	65.3③	78.2
	3D-ST [3]	86.2	48.4	83.2	89.4③	84.8	66.3	76.3	68.7	95.8②	48.6	74.8
	AST_{3D} [17]	88.1③	57.6	96.5②	95.7②	67.9	79.7②	99.0①	91.5③	95.6③	61.1	83.3③
	$M3DM_{3D}$ [19]	94.1②	65.1③	96.5②	96.9①	90.5②	76.0③	88.0③	97.4①	92.6	76.5②	87.4②
	$3DSR_{3D}$	94.5①	83.5①	96.9①	85.7	95.5①	88.0①	96.3②	93.4③	99.8①	88.8 ①	92.2①
RGB	PatchCore [14]	87.6	88.0	79.1	68.2	91.2	70.1	69.5	61.8	84.1	70.2	77.0
	DifferNet [15]	85.9	70.3	64.3	43.5	79.7	79.0	78.7	64.3③	71.5	59.0	69.6
	PADiM [5]	97.5①	77.5	69.8	58.2	95.9	66.3	85.8	53.5	83.2	76.0	76.4
	CS-Flow [16]	94.1	93.0①	82.7	79.5②	99.0②	88.6③	73.1	47.1	98.6②	74.5	83.0
	AST_{RGB} [17]	94.7②	92.8③	85.1③	82.5①	98.1③	95.1①	89.5③	61.3	99.2①	82.1②	88.0②
	$M3DM_{RGB}$ [19]	94.4③	91.8	89.6②	74.9	95.9	76.7	91.9②	64.8②	93.8	76.7③	85.0③
	DSR_{RGB} [22]	84.4	93.0①	96.4①	79.4③	99.8①	90.4②	93.8①	73.0①	97.8③	90.0①	89.8①
3D+RGB	Voxel AE [2]	51.0	54.0	38.4	69.3	44.6	63.2	55.0	49.4	72.1	41.3	53.8
	Depth GAN [2]	53.8	37.2	58.0	60.3	43.0	53.4	64.2	60.1	44.3	57.7	53.2
	Depth AE [2]	64.8	50.2	65.0	48.8	80.5	52.2	71.2	52.9	54.0	55.2	59.5
	PatchCore+FPFH [9]	91.8	74.8	96.7	88.3	93.2	58.2	89.6	91.2③	92.1	88.6③	86.5
	AST [17]	98.3②	87.3②	97.6②	97.1③	93.2③	88.5③	97.4②	98.1①	100①	79.7	93.7③
	M3DM [19]	99.4①	90.9①	97.2③	97.6②	96.0②	94.2②	97.3③	89.9	97.2③	85.0③	94.5②
	3DSR	98.1③	86.7③	99.6①	98.1①	100①	99.4①	98.6①	97.8②	100①	99.5①	97.8①

Table 2. Anomaly localization results on the MVTec3D dataset for the 3D, RGB and 3D+RGB problem setups. The results are listed as AUPRO scores (higher is better). The results of evaluated methods are ranked and the first, second and third place are marked.

	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
3D	Depth AE [2]	14.7	6.9	29.3	21.7	20.7	18.1	16.4	6.6	54.5	14.2	20.3
	Depth GAN [2]	11.1	7.2	21.2	17.4	16.0	12.8	0.3	4.2	44.6	7.5	14.3
	Voxel AE [2]	26.0	34.1	58.1	35.1	50.2	23.4	35.1	65.8	1.5	18.5	34.8
	FPFH [9]	97.3①	87.9①	98.2②	90.6①	89.2②	73.5②	97.7①	98.2①	95.6②	96.1①	92.4①
	M3DM [19]	94.3②	81.8③	97.7③	88.2②	88.1③	74.3①	95.8③	97.4③	95.0③	92.9②	90.6③
	3DSR	92.2③	87.2②	98.4①	85.9③	94.0①	71.4③	97.0②	97.8②	97.7①	85.8③	90.7②
RGB	CFlow [7]	85.5	91.9	95.8③	86.7	96.9	50.0	88.9	93.5	90.4	91.9③	87.1
	PatchCore [14]	90.1	94.9③	92.8	87.7	89.2	56.3	90.4	93.2	90.8	90.6	87.6
	PADiM [2]	98.0①	94.4	94.5	92.5①	96.1③	79.2③	96.6②	94.0③	93.7③	91.2	93.0③
	M3DM [19]	95.2②	97.2①	97.3②	89.1②	93.2	84.3②	97.0①	95.6①	96.8①	96.6②	94.2②
	DSR [22]	92.3③	97.0②	97.9①	85.9	97.9①	89.4①	94.3③	95.1②	96.4②	98.0①	94.4①
3D+RGB	Depth AE [2]	43.2	15.8	80.8	49.1	84.1	40.6	26.2	21.6	71.6	47.8	48.1
	Depth VM [2]	38.8	32.1	19.4	57.0	40.8	28.2	24.4	34.9	26.8	33.1	33.5
	Voxel AE [2]	46.7	75.0	80.8	55.0	76.5	47.3	72.1	91.8	1.9	17.0	56.4
	3D-ST [3]	95.0	48.3	98.6①	92.1	90.5	63.2	94.5	98.8①	97.6②	54.2	83.3
	PatchCore + FPFH [9]	97.6①	96.9②	97.9③	97.3①	93.3③	88.8③	97.5③	98.1②	95.0	97.1③	95.9③
	M3DM [19]	97.0②	97.1①	97.9③	95.0②	94.1②	93.2②	97.7②	97.1	97.1③	97.5②	96.4②
	3DSR	96.4③	96.6③	98.1②	94.2③	98.0①	97.3①	98.1①	97.7③	97.9①	97.9①	97.2①

and image-level AUROC on the RGB+3D setup is shown in Table 3, where 3DSR outperforms both AST [17] and M3DM [19] on both image-level and pixel-level AUROC.

Table 4 shows the comparison between 3DSR and the previous top performing method M3DM [19] on the Eyecandies [4] dataset in terms of the image-level AUROC on the 3D and 3D+RGB problem setups. 3DSR outperforms M3DM [19] on the 3D anomaly detection setup and achieves an image-level AUROC score that is 3 percentage points higher than that of M3DM [19]. On the 3D+RGB anomaly detection setup 3DSR achieves state-of-the-art performance, slightly outperforming M3DM [19] in the mean AUROC score, while achieving a lower vari-

Method	I-AUROC	P-AUROC
PatchCore + FPFH [9]	86.5	99.2
AST [17]	93.7	97.6
M3DM [19]	94.5	99.2
3DSR	**97.8**	**99.5**

Table 3. Mean image-level AUC and pixel-level AUC values on the 3D+RGB setup on the MVTec3D dataset.

Method	Candy cane	Chocolate cookie	Chocolate praline	Confetto	Gummy Bear	Hazelnut truffle	Licorice sandwich	Lollipop	Marshmallow	Peppermint candy	Mean
$3DSR_{3D}$	60.0	76.8	74.2	77.0	76.1	74.9	81.1	83.1	81.1	91.7	**77.6**
$M3DM_{3D}$	48.2	58.9	80.5	84.5	78.0	53.8	76.6	82.7	80.0	82.2	72.5
DSR_{RGB} [22]	70.6	96.5	95.0	96.6	87.0	79.0	88.5	85.7	99.8	99.2	**89.8**
$M3DM_{RGB}$	64.8	94.9	94.1	100	87.8	63.2	93.3	81.1	998	100	87.9
$3DSR_{3D+RGB}$	65.1	99.8	90.4	97.8	87.5	86.1	96.5	89.9	99.0	97.1	**90.9**
$M3DM_{3D+RGB}$	62.4	95.8	95.8	100	88.6	75.8	94.9	83.6	100	100	89.7

Table 4. Comparison between M3DM [19] and 3DSR on the Eyecandies dataset in terms of image-level AUROC (higher is better).

ance in scores across classes. The results suggest that in the Eyecandies dataset [4] most anomalies that are perceptible in 3D are also visible in RGB, since similar results are achieved in the RGB and the 3D+RGB problem setups.

In Figure 6, qualitative examples for the MVTec3D [2] and Eyecandies [4] datasets are shown. The first two rows contain the RGB and depth images, respectively. The third row shows the 3DSR output mask overlaid on the RGB image. The ground truth anomaly masks are shown in row 4. In classes such as Cookie (Column 4), Peach (Column 7) and Potato (Column 8), anomalies are subtle in the RGB image, but are visible in the depth image and are detected by 3DSR. In Column 6, the anomaly on the foam is only visible in the RGB space and is also detected by 3DSR. In other columns the anomalies are visible in both RGB and depth images. The last two columns show examples from the Eyecandies dataset [4]. Column 11 contains a 3D anomaly and column 12 contains an anomaly only visible in the RGB image. 3DSR segments all of these examples accurately.

5. Ablation study

The results of the ablation study are shown in Table 5. First, the naive approach of training the VQVAE of DSR [22] on RGB+depth image pairs of all classes in the MVTec3D [2] training set is evaluated in experiment DSR_{naive}. This results in poor anomaly detection performance showing the need for better RGB+3D representations enabled by the proposed contributions.

The contribution of the simulated training data is evaluated in the $3DSR_{no\_affine}$ and $3DSR_{no\_perlin}$ experiments. In the $3DSR_{no\_perlin}$ experiment the use of Perlin noise maps for training DADA is omitted and the depth training data is replaced with ImageNet images converted to grayscale. Grayscale images do not sufficiently model the properties of depth images and the image-level AUROC score drops significantly by approximately 8 percentage points. In $3DSR_{no\_affine}$ only the perlin noise map scaled from 0 to 1 is used without the additional scaling with α and β as described in Section 3.2. This leads to an approximately 3 percentage point drop in the image-level AUROC. Experiments $3DSR_{no\_affine}$ and $3DSR_{no\_perlin}$ demonstrate the effectiveness of using Perlin noise as a simulation source of industrial depth and the benefit of using an affine transformation of the simulated data to model the characteristics of the data.

The contribution of the DADA module is evaluated in experiments $3DSR_{VQVAE}$, $3DSR_{weighted}$. In these experiments, a VQVAE from [22] is used to learn the joint RGB and depth representations, where the RGB and depth representations are not separated by grouped convolutions. In $3DSR_{VQVAE}$ the proposed DADA module is replaced with a VQVAE [13] model which causes an approximately 2 percentage point drop in image-level AUROC. This experiment shows the benefit of separating the RGB and depth data in the DADA architecture and shows that this separation leads to an improved downstream anomaly detection. Experiment $3DSR_{weighted}$ also uses a vector quantized autoencoder from [22] but the λ_D and λ_I values in the loss in Equation (1) are set to the best-performing values $\lambda_D = 10$ and $\lambda_I = 1$, increasing the loss contribution of depth image reconstruction. This leads to a 1.3 percentage point drop in image-level AUROC suggesting that replacing DADA with VQ-VAE and a simple loss reweighing is not sufficient. The change in λ_I,λ_D also accounts for the difference between $3DSR_{weighted}$ and $3DSR_{VQVAE}$, showing the impact of the choice of λ_I,λ_D.

Method	I-AUROC	P-AUROC	PRO
DSR_{naive}	87.6	96.5	92.3
$3DSR_{no\_perlin}$	90.0	98.3	93.3
$3DSR_{no\_affine}$	94.8	99.2	95.9
$3DSR_{VQVAE}$	95.8	99.3	96.3
$3DSR_{weighted}$	96.5	99.4	96.7
3DSR	**97.8**	**99.5**	**97.2**

Table 5. Ablation study results.

Inference efficiency. The performance in terms of frames-per-second (FPS) is evaluated on an NVIDIA RTX A4500 GPU and shown in Table 6. Compared to other recent methods such as M3DM [19], 3DSR is very fast and can run in real-time on a GPU due to the efficiency of the DADA and the anomaly segmentation module. The previous best method M3DM [19] that uses point-cloud data requires heavy preprocessing and two large transformer net-

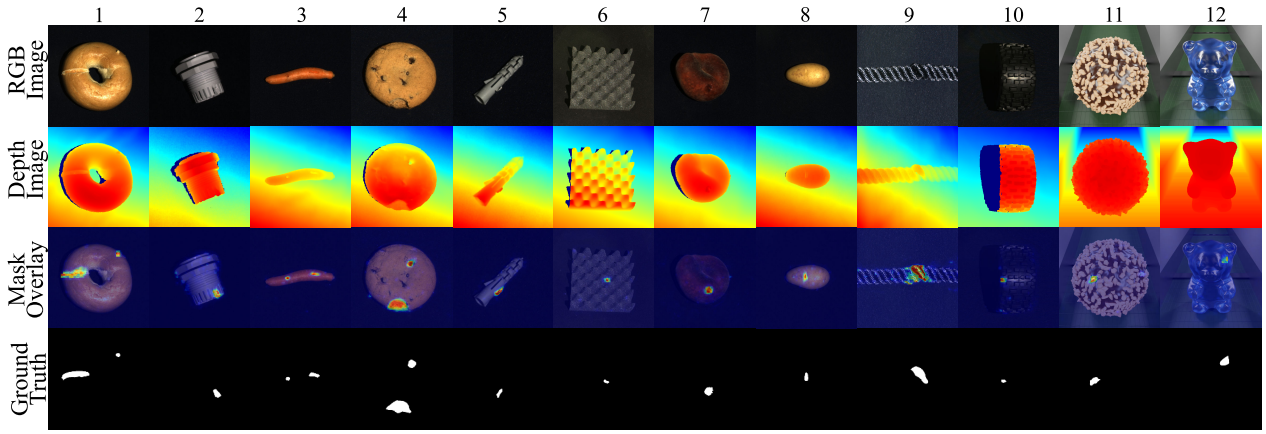

Figure 6. Qualitative results of 3DSR on the MVTec3D and Eyecandies benchmarks.

works. 3DSR is an order of magnitude faster than M3DM and is also almost twice as fast as AST [17] in terms of FPS.

Method	AST [17]	M3DM [19]	3DSR
FPS	18②	0.6③	33①

Table 6. Method performance in terms of frames-per-second (FPS) on the NVIDIA RTX A4500 GPU.

Limitations. Figure 7 shows Eyecandies dataset [4] examples that are particularly difficult for 3DSR. In row one, the anomaly is a small dent at the object edge. It is not visible in the RGB image or distinguishable from the natural edge in the depth map making it difficult to detect. In rows 2 and 3, the anomalies are depth-based and result in minor changes to the object's surface. They are hardly visible in the RGB images. Nonetheles, 3DSR is able to segment them with a lower confidence. In row 4, the anomaly is a deformation visible in the RGB image, however due to the object's transparency, such anomalies are difficult to detect. Note that the Eyecandies dataset [4] contains challenging anomalies that closely resemble the object's normal appearance, making them difficult to detect, even for humans.

6. Conclusion

We proposed a 3D anomaly detection method 3DSR, which is capable of detecting 3D anomalies in industrial depth data and can even utilize depth and RGB data to further improve the anomaly detection performance. Our first contribution is the novel Depth-aware Discrete Autoencoder (DADA) that separately encodes 3D and RGB data during training thus learning better representations of individual modalities. The second contribution is the simulated depth generation process for learning robust representations of industrial 3D data. The new method

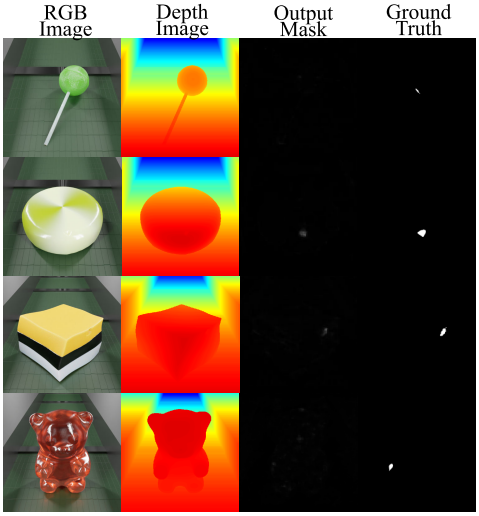

Figure 7. Difficult cases on the Eyecandies dataset.

3DSR (third contribution), achieves state-of-the-art results on the MVTec3D [2] and Eyecandies [4] datasets. On the MVTec3D anomaly detection dataset [2], 3DSR surpasses competing methods significantly in the 3D and 3D+RGB anomaly detection setups demonstrating a strong 3D anomaly detection capability validating the proposed contributions. 3DSR is faster than competing methods and is an order of magnitude faster than the previous best method M3DM [19] which uses a point-cloud-based 3D information extraction. The proposed depth simulation may also help transfer the recent progress in RGB anomaly detection methods to the 3D domain by improving the representations extracted by backbone networks by training on simulated data with self-supervision.

Acknowledgement *This work was supported by Slovenian research agency programs J2-3169, J2-2506, P2-0214 and 23-20MR.R588*

References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec 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. 1, 2, 5

[2] Paul Bergmann., Xin Jin., David Sattlegger., and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 202–213. INSTICC, SciTePress, 2022. 1, 2, 3, 5, 6, 7, 8

[3] Paul Bergmann and David Sattlegger. Anomaly detection in 3d point clouds using deep geometric descriptors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2613–2623, 2023. 1, 2, 6

[4] Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *Proceedings of the Asian Conference on Computer Vision*, pages 3586–3602, 2022. 3, 5, 6, 7, 8

[5] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 2, 6

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4, 5

[7] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107, 2022. 6

[8] Geoffrey E Hinton, Alex Krizhevsky, and Ilya Sutskever. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25(1106-1114):1, 2012. 3

[9] Eliahu Horwitz and Yedid Hoshen. An empirical investigation of 3d anomaly detection and segmentation. *arXiv preprint arXiv:2203.05550*, 2022. 1, 2, 6

[10] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 2

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[12] Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985. 4

[13] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 14866–14876. Curran Associates, Inc., 2019. 2, 3, 4, 7

[14] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. *arXiv preprint arXiv:2106.08265*, 2021. 1, 2, 5, 6

[15] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1907–1916, 2021. 1, 2, 6

[16] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1088–1097, 2022. 2, 6

[17] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2592–2602, 2023. 1, 2, 5, 6, 8

[18] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009. 2

[19] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8041, 2023. 1, 2, 6, 7, 8

[20] Minghui Yang, Peng Wu, and Hui Feng. Memseg: A semi-supervised method for image surface defect detection using differences and commonalities. *Engineering Applications of Artificial Intelligence*, 119:105835, 2023. 1, 2

[21] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021. 2

[22] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr–a dual subspace re-projection network for surface anomaly detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 539–554. Springer, 2022. 2, 4, 5, 6, 7

[23] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8330–8339, October 2021. 2, 4, 5

[24] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 2