Layer	Output Size	Parameters		
		Kernel	Stride	Padding
Input	128x128x1			
Conv1	64x64x32	4x4	2	1
Conv2	32x32x32	4x4	2	1
Conv3	32x32x32	3x3	1	1
Conv4	16x16x64	4x4	2	1
Conv5	16x16x64	3x3	1	1
Conv6	8x8x128	4x4	2	1
Conv7	8x8x64	3x3	1	1
Conv8	8x8x32	3x3	1	1
Conv9	1x1xd	8x8	1	0

Table 1: General outline of our autoencoder architecture. The depicted values correspond to the structure of the encoder. The decoder is built as a reversed version of this. Leaky rectified linear units (ReLUs) with slope 0.2 are applied as activation functions after each layer except for the output layers of both the encoder and the decoder, in which linear activation functions are used.

evaluating the reconstruction error with a per-pixel ℓ^2-comparison or SSIM. For the latter, the luminance, contrast, and structure maps are also shown. For the ℓ^2-distance, both the defects and the inaccuracies in the reconstruction of the edges are weighted equally in the error map, which makes them indistinguishable. Since SSIM computes three different statistical features for image comparison and operates on local patch regions, it is less sensitive to small localization inaccuracies in the reconstruction. In addition, it detects defects that manifest themselves in a change of structure rather than large differences in pixel intensity. For the defects added in this particular toy example, the contrast function yields the largest residuals.

4. EXPERIMENTS

4.1. Datasets

Due to the lack of datasets for unsupervised defect segmentation in industrial scenarios, we contribute a novel dataset of two woven fabric textures, which is available to the public[1]. We provide 100 defect-free images per texture for training and validation and 50 images that contain various defects such as cuts, roughened areas, and contaminations on the fabric. Pixel-accurate ground truth annotations for all defects are also provided. All images are of size 512×512 pixels and were acquired as single-channel gray-scale images. Examples of defective and defect-free textures can be seen in Figure 4. We further evaluate our method on a dataset of nanofibrous materials (Carrera et al., 2017), which contains five defect-free gray-scale images of size 1024×700 for training and validation and 40 defective images for evaluation. A sample image of this dataset is shown in Figure 1.

4.2. Training and Evaluation Procedure

For all datasets, we train the autoencoders with their respective losses and evaluation metrics, as described in Section 3.1. Each architecture is trained on 10 000 defect-free patches of size 128×128, randomly cropped from the given training images. In order to capture a more global context of the textures, we down-scaled the images to size 256×256 before cropping. Each network is trained for 200 epochs using the ADAM (Kingma and Ba, 2015) optimizer with an initial learning rate of 2×10^{-4} and a weight decay set to 10^{-5}. The exact parametrization of the autoencoder network shared by all tested architectures is given in Table 1. The latent space dimension for our experiments is set to $d = 100$ on the texture images and to $d = 500$ for the nanofibres due to their higher structural complexity. For the VAE, we decode $N = 6$ latent samples from the approximate posterior distribution $Q(\mathbf{z}|\mathbf{x})$ to evaluate the reconstruction probability for each pixel. The feature matching autoencoder is regularized with the first three convolutional layers of an AlexNet (Krizhevsky et al., 2012) pretrained on ImageNet (Russakovsky et al., 2015) and a weight factor of $\lambda = 1$. For SSIM, the window size is set to $K = 11$ unless mentioned otherwise and its three residual maps are equally weighted by setting $\alpha = \beta = \gamma = 1$.

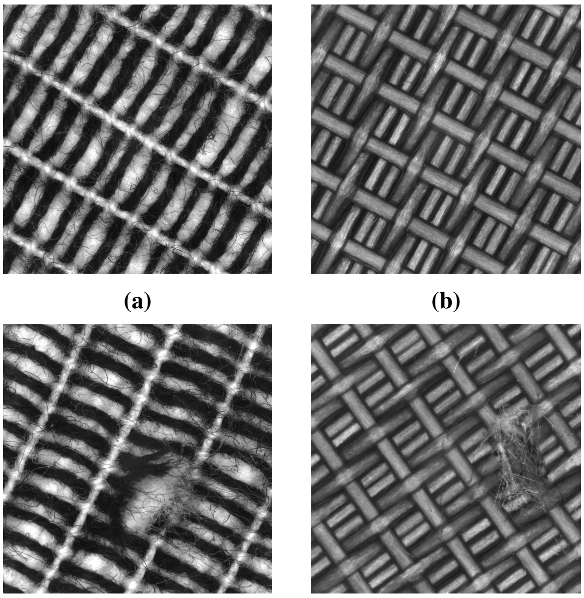
(a) (b) (c) (d)

Figure 4: Example images from the contributed texture dataset of two woven fabrics. **(a)** and **(b)** show examples of non-defective textures that can be used for training. **(c)** and **(d)** show exemplary defects for both datasets. See the text for details.

The evaluation is performed by striding over the test images and reconstructing image patches of size 128×128 using the trained autoencoder and computing its respective residual map R. In principle, it would be possible to set the horizontal and vertical stride to 128. However, at different spatial locations, the autoencoder produces slightly different reconstructions of the same data, which leads to some striding artifacts. Therefore, we decreased the stride to 30 pixels and averaged the reconstructed pixel values. The resulting residual maps are thresholded to obtain candidate regions where a defect might be present. An opening with a circular structuring element of diameter 4 is applied as a morphological post-processing to delete outlier regions that are only a few pixels wide (Steger et al., 2018). We compute the receiver operating characteristic (ROC) as the evaluation metric. The true positive rate is defined as the ratio of pixels correctly classified as defect across the entire dataset. The false positive rate is the ratio of pixels misclassified as defect.

[1] The dataset is available at `https://www.mvtec.com/company/research/publications`