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

表1：当社のオートエンコーダーアーキテクチャの概略。図示された値はエンコーダーの構造に対応しています。デコーダーはこれの逆バージョンとして構築されています。エンコーダーとデコーダーの出力層を除く各層の活性化関数として、勾配0.2のリーキー整流線形単位（ReLU）が適用され、出力層では線形活性化関数が使用されています。

再構築誤差をピクセル単位の`²比較またはSSIMで評価します。後者の場合、輝度、コントラスト、構造マップも表示されます。`²距離では、欠陥とエッジの再構築における不正確さが誤差マップで同等に重み付けされるため、区別できません。SSIMは画像比較のために3つの異なる統計的特徴量を計算し、局所的なパッチ領域で動作するため、再構築における小さな局所化誤差に敏感ではありません。さらに、ピクセル強度の大幅な違いではなく、構造の変化として現れる欠陥を検出します。この特定の単純な例で追加された欠陥の場合、コントラスト関数が最大の残差を生成します。

4. EXPERIMENTS

4.1. Datasets

産業シナリオにおける無監督欠陥セグメンテーション用のデータセットが不足しているため、私たちは2種類の織物テクスチャからなる新規データセットを提案し、一般公開しています[1]。各テクスチャにつき、トレーニングと検証用に欠陥のない画像100枚、および切断、粗面化、汚染などさまざまな欠陥を含む画像50枚を提供しています。すべての欠陥に対するピクセル単位の正確なグラウンドトゥルース注釈も提供されています。すべての画像は512×512ピクセルのサイズで、単一チャネルのグレースケール画像として取得されています。欠陥ありと欠陥なしのテクスチャの例は図4に示されています。さらに、当手法をナノファイバー材料のデータセット（Carrera et al., 2017）で評価しました。このデータセットには、トレーニングと検証用に1024 × 700ピクセルの欠陥なしグレースケール画像5枚、評価用に欠陥あり画像40枚が含まれています。このデータセットのサンプル画像は図1に示されています。

4.2. トレーニングおよび評価手順

すべてのデータセットにおいて、セクション3.1で説明するように、各オートエンコーダーを対応する損失関数と評価指標で訓練します。各アーキテクチャは、与えられた訓練画像からランダムに切り出された128×128サイズの欠陥のないパッチ10,000枚で訓練されます。テクスチャのよりグローバルな文脈を捕捉するため、切り出し前に画像を256×256サイズにダウンサンプリングしました。

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

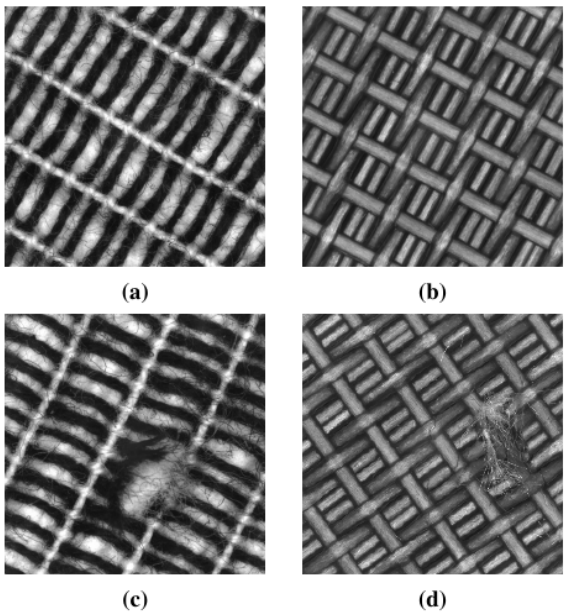

図 4: 2 種類の織物のテクスチャデータセットから提供された画像例。（a）および（b）は、トレーニングに使用できる欠陥のないテクスチャの例です。(c) および (d) は、両方のデータセットの欠陥の例です。詳細については本文をご覧ください。

各ネットワークは、初期学習率2×10^{-4}と重み減衰10^{-5}を設定したADAM（Kingma and Ba, 2015）オプティマイザーを使用して、200エポックで訓練されます。すべてのテストアーキテクチャで共有されるオートエンコーダーネットワークの正確なパラメータ化は、表1に示されています。実験における潜在空間の次元は、テクスチャ画像ではd = 100、ナノファイバーでは構造的複雑さのためd = 500に設定されています。VAEでは、近似事後分布Q(z|x)からN = 6の潜在サンプルをデコードし、各ピクセルの再構築確率を評価します。特徴マッチングオートエンコーダーは、ImageNet（Russakovsky et al., 2015）で事前訓練されたAlexNet（Krizhevsky et al., 2012）の最初の3つの畳み込み層と、重み係数λ = 1で正則化されています。SSIM では、特に明記しない限り、ウィンドウサイズは K = 11 に設定され、$\alpha = \beta = \gamma = 1$ を設定することで 3 つの残差マップが均等に重み付けされます。

評価は、テスト画像上でストライドを行い、トレーニング済みのオートエンコーダを使用して 128 × 128 サイズの画像パッチを再構築し、それぞれの残差マップ R を計算することで行われます。原則として、水平および垂直のストライドを 128 に設定することは可能です。しかし、空間上の位置が異なる場合、オートエンコーダは同じデータに対してわずかに異なる再構築結果を生成するため、ストライドによるアーティファクトが発生します。そのため、ストライドを 30 ピクセルに減らし、再構築されたピクセル値を平均化しました。その結果得られた残差マップに閾値を設定し、欠陥が存在する可能性のある領域候補を取得します。直径 4 の円形の構造要素を用いた開孔処理を形態学的後処理として適用し、幅数ピクセル程度の外れ値領域を削除します（Steger et al., 2018）。評価指標として、受信者動作特性（ROC）を計算します。真陽性率は、データセット全体で欠陥として正しく分類されたピクセルの比率として定義されます。偽陽性率は、欠陥と誤って分類されたピクセルの比率です。

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 $\alpha = \beta = \gamma = 1$.

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.

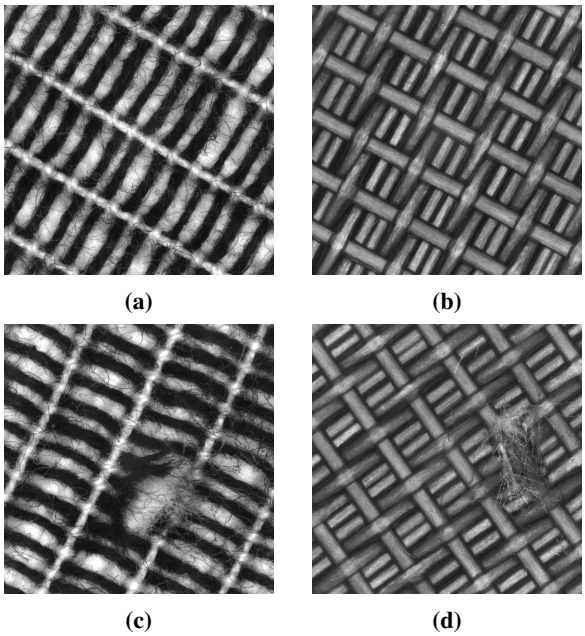

Figure 4: Example images from the contributed texture dataset of two woven fabrics. **(a)** and **(b)** show examples of non-defective textures that can be used for training. **(c)** and **(d)** show exemplary defects for both datasets. See the text for details.

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