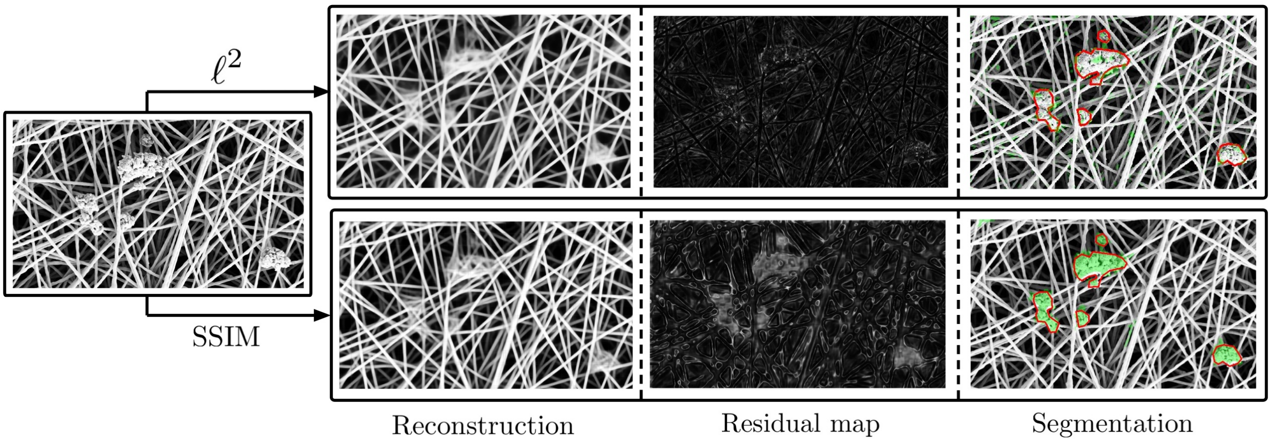

Figure 1: A defective image of nanofibrous materials is reconstructed by an autoencoder optimizing either the commonly used pixel-wise ℓ^2-distance or a perceptual similarity metric based on structural similiarity (SSIM). Even though an ℓ^2-autoencoder fails to properly reconstruct the defects, a per-pixel comparison of the original input and reconstruction does not yield significant residuals that would allow for defect segmentation. The residual map using SSIM puts more importance on the visually salient changes made by the autoencoder, enabling for an accurate segmentation of the defects.

unsupervised defect segmentation approaches that rely on additional model priors such as handcrafted features or pretrained networks.

2. RELATED WORK

Detecting anomalies that deviate from the training data has been a long-standing problem in machine learning. Pimentel et al. (2014) give a comprehensive overview of the field. In computer vision, one needs to distinguish between two variants of this task. First, there is the classification scenario, where novel samples appear as entirely different object classes that should be predicted as outliers. Second, there is a scenario where anomalies manifest themselves in subtle deviations from otherwise known structures and a segmentation of these deviations is desired. For the classification problem, a number of approaches have been proposed (Perera and Patel, 2018; Sabokrou et al., 2018). Here, we limit ourselves to an overview of methods that attempt to tackle the latter problem.

Napoletano et al. (2018) extract features from a CNN that has been pretrained on a classification task. The features are clustered in a dictionary during training and anomalous structures are identified when the extracted features strongly deviate from the learned cluster centers. General applicability of this approach is not guaranteed since the pretrained network might not extract useful features for the new task at hand and it is unclear which features of the network should be selected for clustering. The results achieved with this method are the current state-of-the-art on the NanoTWICE dataset, which we also use in our experiments. They improve upon previous results by Carrera et al. (2017), who build a dictionary that yields a sparse representation of the normal data. Similar approaches using sparse representations for novelty detection are (Boracchi et al., 2014; Carrera et al., 2015, 2016).

Schlegl et al. (2017) train a GAN on optical coherence tomography images of the retina and detect anomalies such as retinal fluid by searching for a latent sample that minimizes the per-pixel ℓ^2-reconstruction error as well as a discriminator loss. The large number of optimization steps that must be performed to find a suitable latent sample makes this approach very slow. Therefore, it is only useful in applications that are not time-critical. Recently, Zenati et al. (2018) proposed to use bidirectional GANs (Donahue et al., 2017) to add the missing encoder network for faster inference. However, GANs are prone to run into mode collapse, i.e., there is no guarantee that all modes of the distribution of non-defective images are captured by the model. Furthermore, they are more difficult to train than autoencoders since the loss function of the adversarial training typically cannot be trained to convergence (Arjovsky and Bottou, 2017). Instead, the training results must be judged manually after regular optimization intervals.

Baur et al. (2018) propose a framework for defect segmentation using autoencoding architectures and a per-pixel error metric based on the ℓ^1-distance. To prevent the disadvantages of their loss function, they improve the reconstruction quality by requiring aligned input data and adding an adversarial loss to enhance the visual quality of the reconstructed images. However, for many applications that work on unstructured data, prior alignment is impossible. Furthermore, optimizing for an additional adversarial loss during training but simply segmenting defects based on per-pixel comparisons during evaluation might lead to worse results since it is unclear how the adversarial training influences the reconstruction.

Other approaches take into account the structure of the latent space of variational autoencoders (VAEs; Kingma and Welling, 2014) in order to define measures for outlier detection. An and Cho (2015) define a reconstruction probability for every image pixel by drawing multiple samples from the estimated encoding distribution and measuring the variability of the decoded outputs. Soukup and Pinetz (2018) disregard the decoder output entirely and instead compute the KL divergence as a novelty measure between the prior and the encoder distribution. This is based on the assumption that defective inputs will manifest themselves in mean and variance values that are very different from those of the prior. Similarly, Vasilev et al. (2018) define multiple novelty measures, either by purely considering latent space behavior or by