

Figure 5: Qualitative comparison between reconstructions, residual maps, and segmentation results of an ℓ^2 -autoencoder and an SSIM autoencoder on two datasets of woven fabric textures. The ground truth regions containing defects are outlined in red while green areas mark the segmentation result of the respective method.

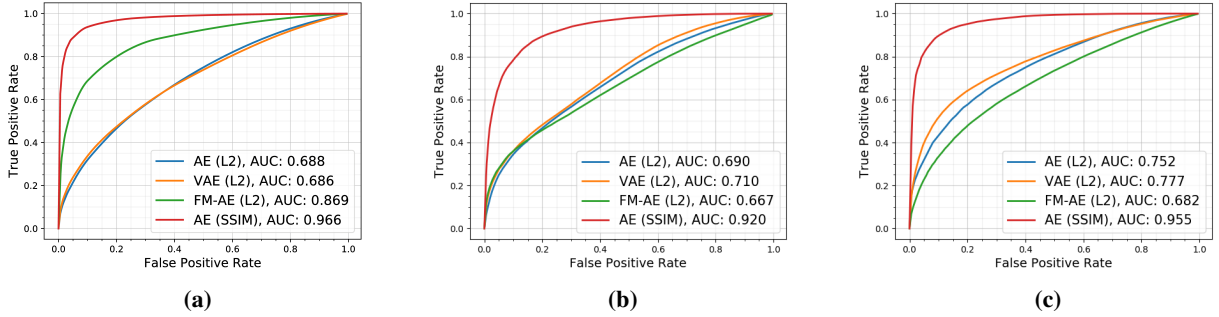


Figure 6: Resulting ROC curves of the proposed SSIM autoencoder (red line) on the evaluated datasets of nanofibrous materials (a) and the two texture datasets (b), (c) in comparison with other autoencoding architectures that use per-pixel loss functions (green, orange, and blue lines). Corresponding AUC values are given in the legend.

4.3. Results

Figure 5 shows a qualitative comparison between the performance of the ℓ^2 -autoencoder and the SSIM autoencoder on images of the two texture datasets. Although both architectures remove the defect in the reconstruction, only the SSIM residual map reveals the defects and provides an accurate segmentation result. The same can be observed for the NanoTWICE dataset, as shown in Figure 1.

We confirm this qualitative behavior by numerical results. Figure 6 compares the ROC curves and their respective AUC values of our approach using SSIM to the per-pixel architectures. The performance of the latter is often only marginally better than classifying each pixel randomly. For the VAE, we found that the reconstructions obtained by different latent samples from the posterior does not vary greatly. Thus, it could not improve on the deterministic framework. Employing feature matching only improved the segmentation result for the dataset of nanofibrous materials, while not yielding a benefit for the two texture datasets. Using SSIM as the loss and evaluation metric outperforms all other tested architectures significantly. By merely changing the loss function, the achieved AUC improves from 0.688 to 0.966 on the dataset of nanofibrous materials, which is comparable to the state-of-the-art by [Napoletano et al. \(2018\)](#), where values of up to 0.974 are reported. In contrast to this method, autoencoders do not rely on any model priors

such as handcrafted features or pretrained networks. For the two texture datasets, similar leaps in performance are observed.

Since the dataset of nanofibrous materials contains defects of various sizes and smaller sized defects contribute less to the overall true positive rate when weighting all pixel equally, we further evaluated the overlap of each detected anomaly region with the ground truth for this dataset and report the p -quantiles for $p \in \{25\%, 50\%, 75\%\}$ in Figure 7. For false positive rates as low as 5%, more than 50% of the defects have an overlap with the ground truth that is larger than 91%. This outperforms the results achieved by [Napoletano et al. \(2018\)](#), who report a minimal overlap of 85% in this setting.

We further tested the sensitivity of the SSIM autoencoder to different hyperparameter settings. We varied the latent space dimension d , SSIM window size k , and the size of the patches that the autoencoder was trained on. Table 2 shows that SSIM is insensitive to different hyperparameter settings once the latent space dimension is chosen to be sufficiently large. Using the optimal setup of $d = 500$, $k = 11$, and patch size 128×128 , a forward pass through our architecture takes 2.23 ms on a Tesla V100 GPU. Patch-by-patch evaluation of an entire image of the NanoTWICE dataset takes 3.61 s on average, which is significantly faster than the runtimes reported by [Napoletano et al. \(2018\)](#). Their approach requires between