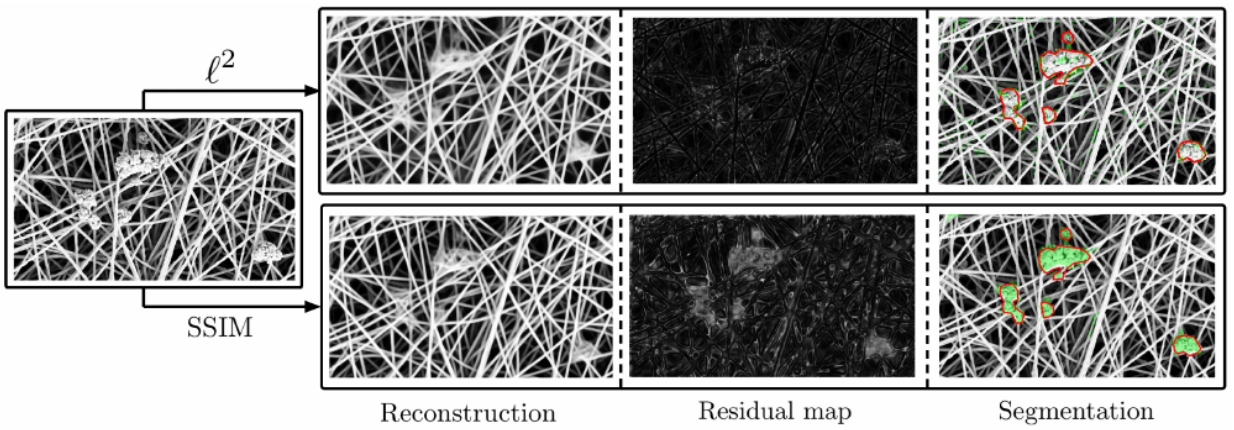

図1：ナノファイバー材料の欠陥画像が、一般的に使用されるピクセル単位の`2 -距離または構造的類似性に基づく知覚的類似性メトリクス（SSIM）を最適化するオートエンコーダーによって再構築されます。`2 -オートエンコーダーは欠陥を適切に再構築できませんが、元の入力と再構築のピクセル単位の比較では、欠陥セグメンテーションを可能にするような有意な残差は得られません。SSIMを使用した残差マップは、オートエンコーダーがもたらす視覚的に目立つ変化に重点を置き、欠陥の正確なセグメンテーションを可能にします。

追加のモデル事前知識（手動で作成された特徴量や事前学習済みネットワークなど）に依存する非監督型欠陥セグメンテーション手法。

ここでは、後者の問題に取り組む方法を概観するに留めます。
＃2. 関連研究

トレーニングデータから逸脱する異常を検出することは、機械学習における長年の課題です。Pimentel et al.（2014）は、この分野の包括的な概要を提供しています。コンピュータビジョンでは、このタスクの2つのバリエーションを区別する必要があります。まず、分類シナリオでは、新規サンプルが完全に異なるオブジェクトクラスとして出現し、外れ値として予測される必要があります。次に、異常が既知の構造からの微妙な偏差として現れ、これらの偏差のセグメンテーションが望まれるシナリオがあります。分類問題に関しては、複数のアプローチが提案されています（Perera and Patel, 2018; Sabokrou et al., 2018）。
Napoletanoら（2018）は、分類タスクで事前学習されたCNNから特徴量を抽出します。特徴量はトレーニング中に辞書でクラスタリングされ、抽出された特徴量が学習されたクラスタ中心から大きく乖離する場合、異常な構造が特定されます。このアプローチの汎用性は保証されていません。事前学習されたネットワークが新しいタスクに有用な特徴量を抽出できない可能性があり、クラスタリングに選択すべきネットワークの特徴量が不明確だからです。この方法による結果は、当研究でも使用するNanoTWICEデータセットにおける現在の最先端成果です。彼らは、Carreraら（2017）の以前の結果を改善しています。Carreraらは、正常なデータの疎な表現を生成する辞書を作成しました。同様の疎な表現を用いた異常検出のアプローチには、（Boracchiら、2014；Carreraら、2015、2016）があります。
Schlegl et al.（2017）は、網膜の光コヒーレンストモグラフィ画像を用いて GAN をトレーニングし、ピクセルごとの`2 -再構成誤差と識別器損失を最小化する潜在的なサンプルを検索することで、網膜液などの異常を検出しています。適切な潜在サンプルを見つけるために実行しなければならない最適化ステップの数が多いため、このアプローチは非常に時間がかかります。したがって、このアプローチは、時間に制約のないアプリケーションにのみ有用です。最近、Zenati ら（2018）は、より高速な推論のために、双方向 GAN（Donahue ら、2017）を使用して、欠落しているエンコーダネットワークを追加することを提案しました。しかし、GANはモード崩壊に陥りやすく、つまり、非欠陥画像の分布のすべてのモードがモデルによって捕捉される保証はありません。さらに、敵対的学習の損失関数は通常収束まで訓練できないため（Arjovsky and Bottou, 2017）、オートエンコーダーよりも訓練が困難です。代わりに、定期的な最適化間隔後に訓練結果を手動で判断する必要があります。
Baur ら（2018）は、オートエンコーダーアーキテクチャと`1 -距離に基づくピクセル単位の誤差指標を用いた欠陥セグメンテーションのフレームワークを提案しています。彼らの損失関数の欠点を回避するため、入力データの事前アラインメントを要求し、再構築画像の視覚的品質を向上させるための敵対的損失を追加することで、再構築品質を改善しています。しかし、構造化されていないデータを取り扱う多くのアプリケーションでは、事前アラインメントは不可能です。さらに、トレーニング中に追加の敵対的損失を最適化しながら、評価時にピクセル単位の比較に基づいて欠陥をセグメント化すると、敵対的トレーニングが再構築に与える影響が不明確なため、より悪い結果になる可能性があります。
他のアプローチでは、変分オートエンコーダー（VAE；Kingma and Welling, 2014）の潜在空間の構造を考慮し、外れ値検出のための測定基準を定義しています。An and Cho（2015）は、推定されたエンコーディング分布から複数のサンプルを抽出し、デコード出力の変動を測定することで、各画像ピクセルの再構築確率を定義しています。Soukup と Pinetz（2018）は、デコーダの出力を完全に無視し、代わりに、事前分布とエンコーダ分布間の新規性測定値として KL 発散を計算しています。これは、欠陥のある入力は、事前分布とは大きく異なる平均値と分散値として現れるという仮定に基づいています。同様に、ヴァシレフら（2018）は、潜在空間の挙動のみを考慮するか、あるいは

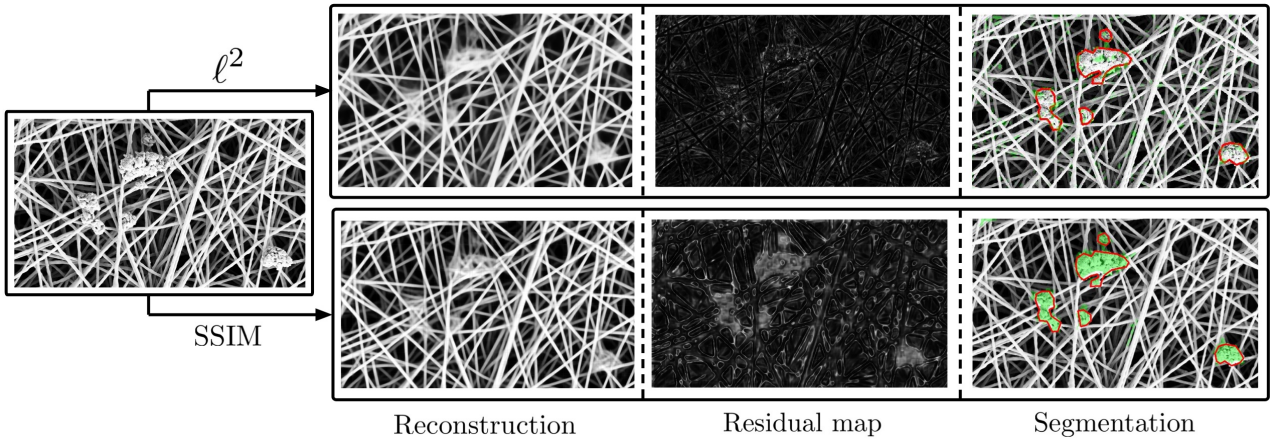

Figure 1: A defective image of nanofibrous materials is reconstructed by an autoencoder optimizing either the commonly used pixel-wise ℓ^2-distance or a perceptual similarity metric based on structural similiarity (SSIM). Even though an ℓ^2-autoencoder fails to properly reconstruct the defects, a per-pixel 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