MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection: Supplementary Material

Paul Bergmann Michael Fauser David Sattlegger Carsten Steger

MVTec Software GmbH

www.mvtec.com

{paul.bergmann, fauser, sattleger, steger}@mvtec.com

Abstract

We provide the following supplementary material:

- *Average region overlap plots for every dataset category and all evaluated methods when using different anomaly thresholds.*
- *Qualitative anomaly segmentation results for each method.*
- *A video showing example images of the dataset, the diversity of anomalies, and the quality of annotations.*

1. Average Region Overlap Curves

Although the determination of a threshold for anomaly maps without any defective training samples is a vital component for achieving a truly unsupervised anomaly detection, it might occur that the threshold is highly under- or overestimated. Consequently, the obtained segmentation results might underperform in comparison to a more appropriately set threshold. To facilitate a comparison of the evaluated methods independent of the estimated threshold, Figures 1–3 show the average region overlap together with the false positive rates that are obtained for different threshold values. The average region overlap measures the average true positive rate per ground-truth region.

Figure 1 shows that the CNN Feature Dictionary performs the most consistently across all textures. The other methods work well on some textures and fail on at least two others. Figure 2 shows curves for objects where an evaluation with the variation model was possible. Figure 3 shows evaluation results for the remaining objects of the dataset. Either the L2 or the SSIM autoencoder achieves the highest average region overlap with the ground truth for most of the objects when fixing a false positive rate of 5%. Since the CNN Feature Dictionary does not take into account spatial information of extracted patches, it does not perform as well for objects as it does for textures. Overall, there is still much room for improvement for each of the evaluated methods.

2. Qualitative Segmentation Results

We further provide qualitative results for each evaluated method on a specific texture or object in Figure 4. The top row shows an anomaly for which the respective method yields reasonable segmentation results. The bottom row shows a different type of anomaly on the same object or texture for which the method fails. This illustrates the large diversity of anomalies in our dataset, which allows for a more thorough evaluation of methods across different domains on the same object or texture.

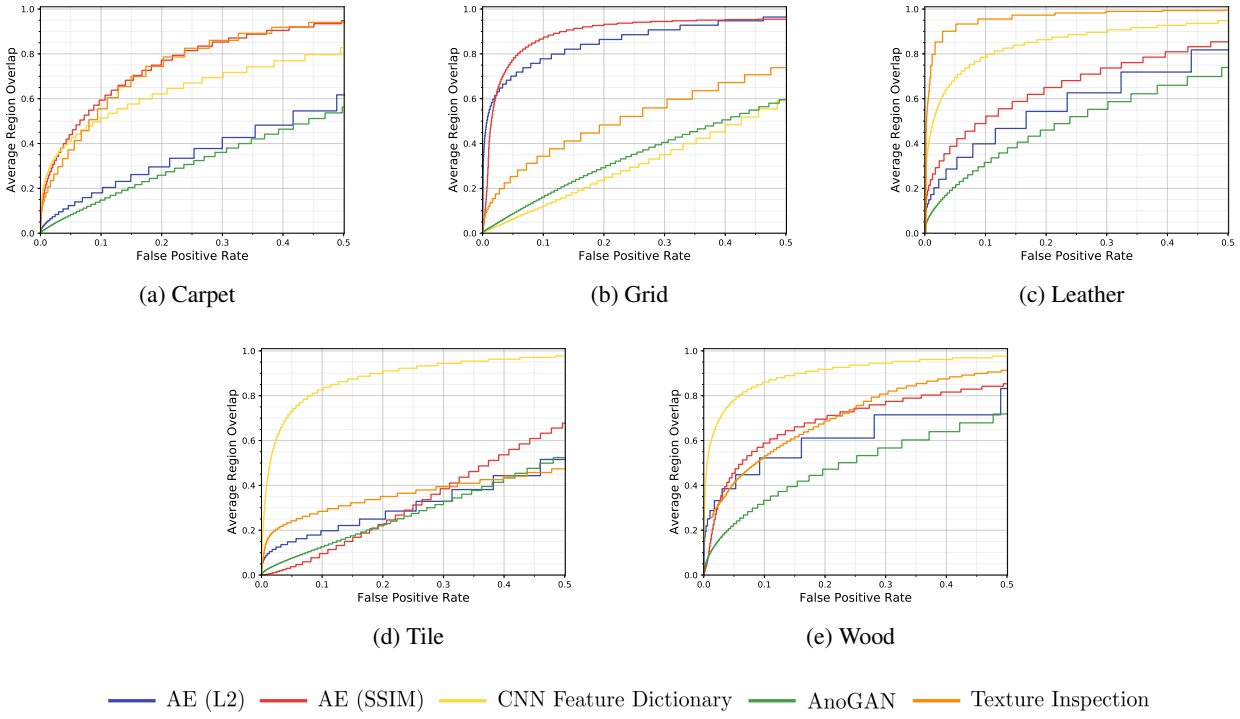
The SSIM Autoencoder, for instance, reliably detects the crack in the zipper, while it fails to detect more subtle anomalies in the fabric. The L2 Autoencoder shows similar behavior while at the same time yielding many false positives due to small reflections that are hard to model in the reconstruction and yield large per-pixel residuals.

Since AnoGAN also evaluates per-pixel residuals, the crack on the capsule with large color difference to the original pill can be segmented, while more structural defects, such as the scratched number imprint, are hard to detect.

The CNN Feature Dictionary is the only method that achieves reasonable results on *tile*. However, it cannot segment the transparent glue strip. It does well on more salient defects such as the crack through the material.

The texture inspection model performs well on structural anomalies such as the hole in the fabric material. It fails to detect the less obvious metal part on the fabric.

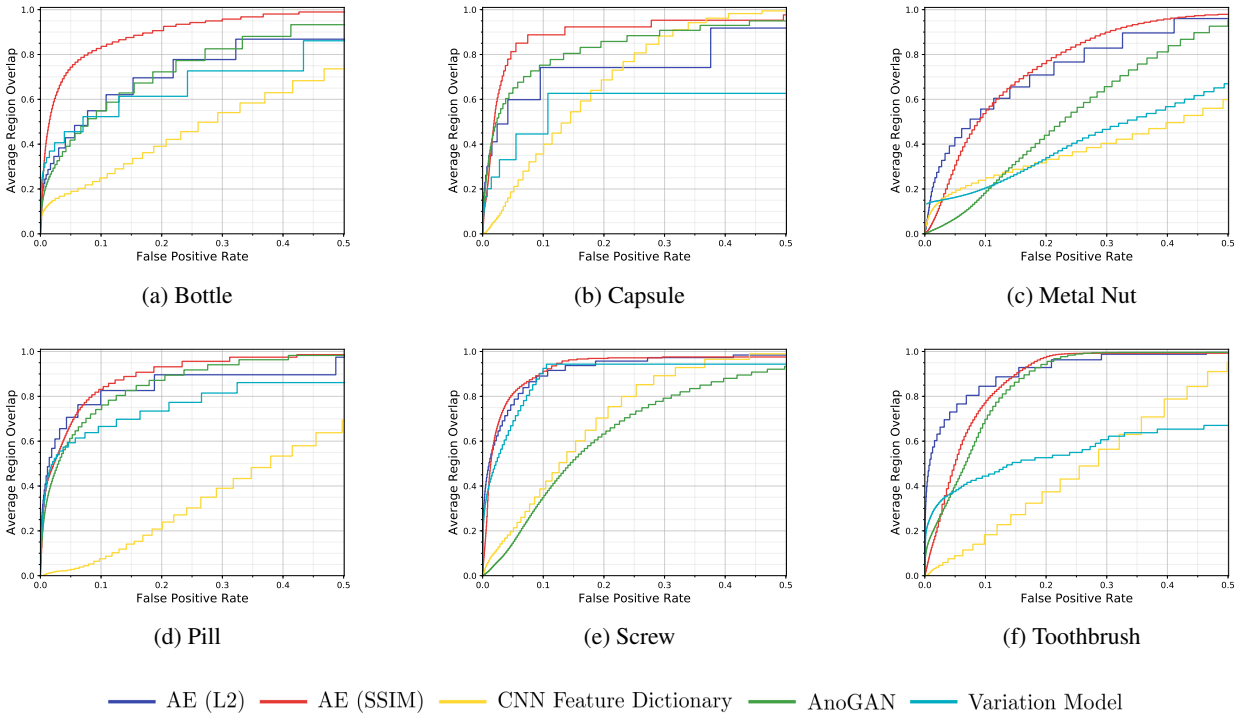
Whenever a pixel-precise alignment of objects is possible and defects deviate noticeably from the norm in terms of gray-value, the variation model is able to detect anomalies reliably, e.g., for the scratch in the screw's head in Figure 4 (top row). However, the model is sensitive to slight misalignments and often yields large anomaly scores around wrongly aligned edges.



(a) Carpet (b) Grid (c) Leather

(d) Tile (e) Wood

Figure 1: Average region overlap for the textures of the MVTec AD dataset. No method manages to perform well in all categories.



(a) Bottle (b) Capsule (c) Metal Nut

(d) Pill (e) Screw (f) Toothbrush

Figure 2: Average region overlap for all objects of the MVTec AD dataset for which an additional evaluation with the variation model was possible.

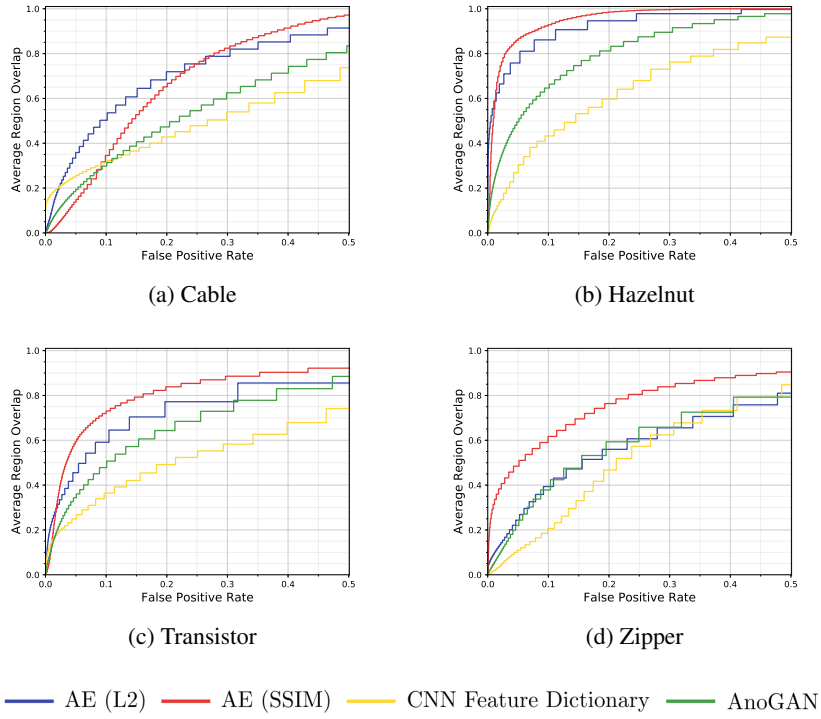


Figure 3: Average region overlap for all objects of the MVTec AD dataset for which an additional evaluation with the variation model was not possible.

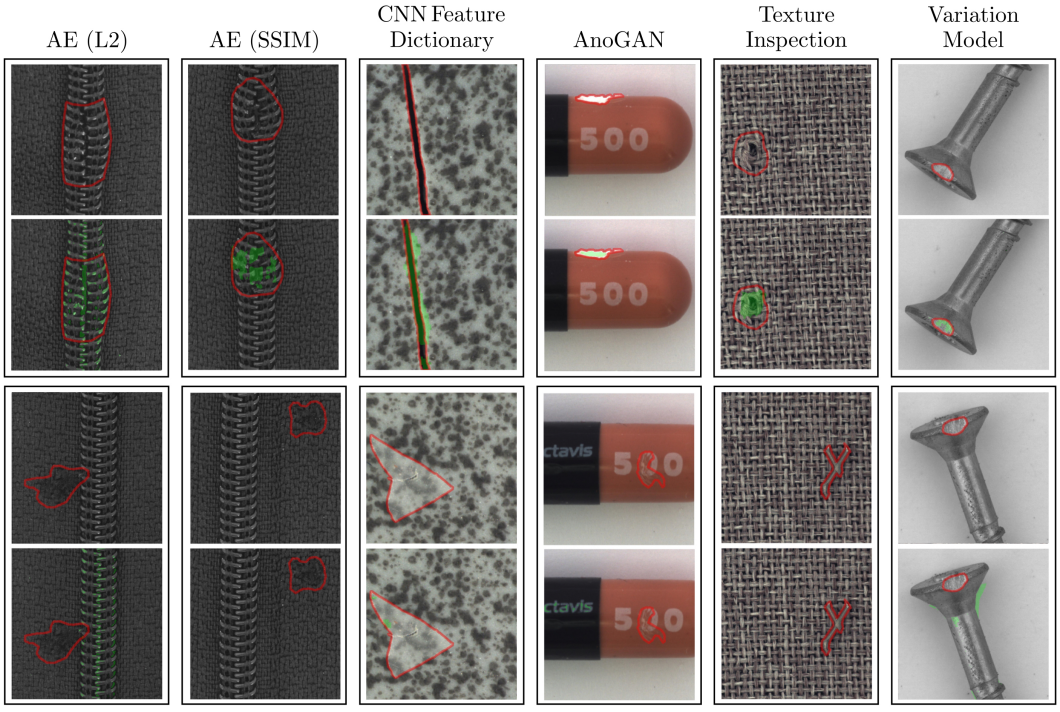


Figure 4: Qualitative anomaly segmentation results for all evaluated architectures. The top row shows an example for which the respective method worked well. The second row shows a failure case of the respective method on the same category for a different type of anomaly.