ENERGY-BASED ANOMALY DETECTION AND LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This brief sketches initial progress towards a unified energy-based solution for the semi-supervised visual anomaly detection and localization problem. In this setup, we have access to only anomaly-free training data and want to detect and identify anomalies of an arbitrary nature on test data. We employ the density estimates from the energy-based model (EBM) as normalcy scores that can be used to discriminate normal images from anomalous ones. Further, we back-propagate the gradients of the energy score with respect to the image in order to generate a gradient map that provides pixel-level spatial localization of the anomalies in the image. In addition to the spatial localization, we show that simple processing of the gradient map can also provide alternative normalcy scores that either match or surpass the detection performance obtained with the energy value. To quantitatively validate the performance of the proposed method, we conduct experiments on the MVTec industrial dataset. Though still preliminary, our results are very promising and reveal the potential of EBMs for simultaneously detecting and localizing unforeseen anomalies in images.

1 INTRODUCTION

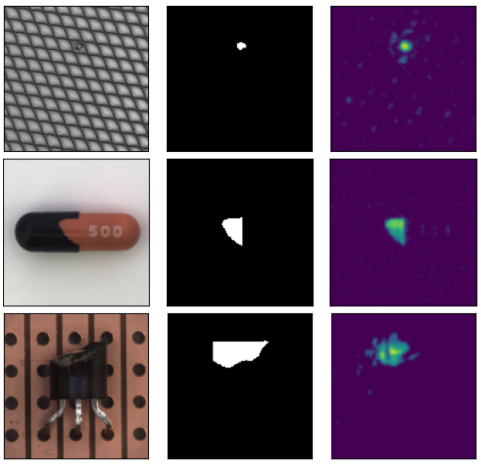


Figure 1: Sample localization outputs for *grid*, *capsule* and *transistor* categories from MVTec AD: From left to right, columns show test image, ground truth mask and predicted localization

The advent of smart manufacturing and the explosion of data collection in factories and other industrial setups have brought considerable attention to the topic of anomaly detection. Its goal is to identify rare, abnormal events from the observation of data. This research area has already inspired a vast body of literature whose central idea is to learn a model of normalcy from the positive data and subsequently flag samples that deviate from the learned model as anomalous. Classic approaches include methods based on distances and nearest-neighbor Knorr et al. (2000); Breunig et al. (2000), clustering He et al. (2003), principal component analysis and its kernel variants Hoffmann (2007), and one-class support vector machines. Recently, deep learning has also been used for (deep) anomaly detection, and existing out-of-distribution (OOD) detection methods were successfully applied to this problem Lee et al. (2018); Ahuja et al. (2019); Hendrycks et al. (2019); Ren et al. (2019). In addition, deep generative models such as variational autoencoders (VAE) and generative adversarial networks (GAN) have also been used to characterize abnormality in a variety of manners, based on gradient Kwon et al. (2020), reconstruction error Chen & Konukoglu (2018), or density estimation An & Cho (2015). We refer to Ruff et al. (2021); Goldstein & Uchida (2016) for an exhaustive review of anomaly detection methods. The overwhelming majority of these methods deal with the detection problem, where one simply needs to discriminate normal samples from anomalous ones. There are numerous usages, medical

imaging being one example, wherein it is very important not only to identify abnormal images, but also to localize the abnormality in such images for diagnosis purposes (e.g. areas with brain lesions in CT images)

The literature on anomaly localization is relatively more scant, but includes recent methods based on deep generative models that provide spatial localization by comparing the reconstructed image with the original input Bergmann et al. (2019b); Schlegl et al. (2017); Chen & Konukoglu (2018). While still budding, there is also a new trend of emerging works that focus on anomaly localization Yi & Yoon (2020); Defard et al. (2020); Cohen & Hoshen (2021).

Energy-based models for visual anomaly detection and localization. The principal idea behind energy-based models (EBM) is to learn an *energy function* $E_\theta(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}$ that parametrizes a density $p_\theta(\mathbf{x})$ as (LeCun et al., 2006):

$$p_\theta(\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{x}))}{\int_{\mathbf{x}} \exp(-E_\theta(\mathbf{x}))},$$

where θ are parameters of the model. An interesting property of EBMs is that, unlike probabilistic models, they don't necessitate normalization nor a tractable likelihood. Hence, any nonlinear regression function can act as an energy function. These attributes provide great flexibility to EBMs, and we have seen many recent works successfully take advantage of it across different data domains and a variety of energy functions and deep networks. However, being devoid of a tractable likelihood requirement comes with its own challenges. Stable and resource-efficient training of EBMs remains an open problem and recent research has focused on training on popular datasets that were compiled for small-scale classification problems (e.g. CIFAR10). In this regard, the application of EBMs to new problems and different data domains is an opportunity to further develop and popularize EBMs.

Of special importance to this work, EBMs have been shown to have better performance in OOD detection than other likelihood-based generative models (Zhai et al., 2016; Grathwohl et al., 2019; Liu et al., 2020; Du & Mordatch, 2019). This makes EBMs a promising alternative for semi-supervised anomaly detection problems. This is especially true for real-world industrial settings where defect data is limited or even unavailable, and the nature of anomalies and defects is not known *a priori*. In those cases, training a model with an explicit discriminative loss term is impractical.

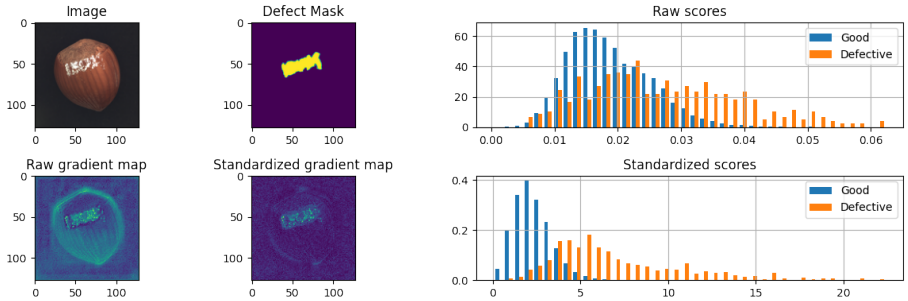
Contributions. This paper proposes a unified EBM method that simultaneously addresses image-level anomaly detection and pixel-level anomaly localization in the semi-supervised setting. Our main contributions can be listed as:

- An EBM method for visual anomaly detection that jointly tackles detection and localization.
- The back-propagation of gradients of the EBM energy w.r.t. the image, resulting in a (gradient) heat map from which we can derive both an innovative detection score and pixel-level localization of anomalies in the input image. To our knowledge, this is the first time that gradients of energy score w.r.t input images are interpreted as spatial anomaly information.
- The application of the method to a real-world industrial case and its quantitative validation with objective metrics. Industrial defect detection stresses the capabilities of generative models as it comes with challenges that can be hardly approximated with synthetic, or research datasets. We provide detailed detection and localization results on the MVTec AD dataset, and compare against reported bechmarks.

2 METHOD

We present the details on our method in this section. Details about the training procedure are provided in Section 3. During inference, an input $\mathbf{x} \in \mathbb{R}^{w \times h \times c}$ is first processed by the network in a typical forward pass to generate an energy scalar. w and h are the width and height, respectively, of the input, and c is the number of channels ($c = 3$ for RGB images, $c = 1$ for grayscale images). The gradient of this scalar w.r.t to the input (not the weights) is calculated by back-propagation to produce a gradient-map, $g(\mathbf{x})$, which has the same shape as that of the input,

$$g(\mathbf{x}) \triangleq \frac{\partial \log p_\theta(\mathbf{x})}{\partial \mathbf{x}}. \tag{1}$$



(a) Sample image, with visualizations of raw and standardized gradient maps

(b) Histograms of raw and standardized scores for good vs defective pixels

Figure 2

Let g_k denote the kth channel of g. A visualization of these gradient-maps shows that although the gradient value is high in the anomalous regions in the image (see Fig. 2a), it is also high in other non-defective regions of the image. Hence, distinguishing between good and defective pixels based on these raw scores may result in significant false positives. We can significantly improve on this by modeling a probability density $p_{\mathbf{x}}(g)$ over the gradient-map's values for each pixel location individually. Note that this density is modelled over values of g at a particular pixel location $\mathbf{x}$. The modeling is performed using the set of gradient-maps, $g^{train}(\mathbf{x})$, derived from the set of training images. During testing, the log-likelihood score of the observed gradient value, $\hat{g}$, at $\mathbf{x}$ is calculated as $l = \log p_{\mathbf{x}}(\hat{g})$ and this score is used as an anomaly score. Hence, the determination of whether a pixel is good or defective is based on how much its gradient value deviates from the typical values observed on the training set images. If the distribution used is a normal distribution, then the log-likelihood, l, is equivalent to standardizing the raw anomaly score by its mean, $\mu_k(\mathbf{x})$, and standard deviation, $\sigma_k(\mathbf{x})$, as follows

$$l_k(\mathbf{x}) = \frac{g_k(\mathbf{x}) - \mu_k(\mathbf{x})}{\sigma_k(\mathbf{x})}.$$

Note that $\mu_k(\mathbf{x})$ and $\sigma_k(\mathbf{x})$ are calculated for each per-pixel location, $\mathbf{x}$, in each channel k separately, from g^{train} values at $\mathbf{x}$. The final anomaly score, $a(\mathbf{x})$ at $\mathbf{x}$ is the L_1 or L_2 norm of the corresponding raw or standardized gradient scores from each channel, i.e.

$$a_{raw}(\mathbf{x}) = \left(\sum_{k=1}^{c} |g_k(\mathbf{x})|^r \right)^{1/r}, \quad a_{std}(\mathbf{x}) = \left(\sum_{k=1}^{c} |l_k(\mathbf{x})|^r \right)^{1/r}, \; r = 1, 2. \tag{2}$$

As can be seen in Fig. 2b, standardized scores show much better separation between good and defective pixels.

Finally, we aggregate the per-pixel anomaly scores to obtain a single score $A(\mathbf{x})$ for an entire image, again by taking the L_1 or L_2 norm as follows

$$\begin{aligned} A_{raw} &= \|g(\mathbf{x})\|_r \triangleq \left(\sum_{\mathbf{x}} |a_{raw}(\mathbf{x})|^r \right)^{1/r} \\ A_{std} &= \|l(\mathbf{x})\|_r \triangleq \left(\sum_{\mathbf{x}} |a_{std}(\mathbf{x})|^r \right)^{1/r} \end{aligned} \tag{3}$$

This score can be used to for the classification of the image as defective or good. Note that A_{raw} is essentially the same as approximate mass score proposed in (Grathwohl et al., 2019).

Table 1: AUROC results for anomaly detection on MVTec (results for other benchmarks taken from Cohen & Hoshen (2021))

	Geom	GANomaly	AE(L2)	ITAE	SPADE	EBM (Ours)
Average	0.67	0.76	0.75	0.84	0.85	0.72

Table 2: AUROC results for anomaly localization on MVTec (results for other benchmarks taken from Bergmann et al. (2019b))

Category	AE (SSIM)	AE (L2)	AnoGAN	CNN Feature Dictionary	EBM (Ours)
Carpet	0.87	0.59	0.54	0.72	0.63
Grid	0.94	0.90	0.58	0.59	0.86
Leather	0.78	0.75	0.64	0.87	0.87
Tile	0.59	0.51	0.50	0.93	0.57
Wood	0.73	0.73	0.62	0.91	0.74
Bottle	0.93	0.86	0.86	0.78	0.72
Cable	0.82	0.86	0.78	0.79	0.56
Capsule	0.94	0.88	0.84	0.84	0.64
Hazelnut	0.97	0.95	0.87	0.72	0.78
Metal Nut	0.89	0.86	0.76	0.82	0.65
Pill	0.91	0.85	0.87	0.68	0.75
Screw	0.96	0.96	0.80	0.87	0.87
Toothbrush	0.92	0.93	0.90	0.77	0.68
Transistor	0.90	0.86	0.80	0.66	0.74
Zipper	0.88	0.77	0.78	0.76	0.55

3 EXPERIMENTS

3.1 TRAINING

In this work, we take an MCMC-based maximum likelihood (ML) approach to training our model. As our energy function approximator, we use an all convolutional neural network similar to the one Nijkamp et al. (2019) used (see Table 3 in the Appendix), but instead of *Leaky-ReLU*, we use *ELU* (Clevert et al., 2016) activation functions between all layers to stabilize training. To optimize for likelihood, we use Contrastive Divergence (CD) (Hinton, 2002).

To synthesize the negative samples needed by the CD objective, we employ a Stochastic Gradient Langevin Dynamics (SGLD) MCMC sampler (Welling & Teh, 2011). We initialize MCMC chains from a *uniform* distribution and run a fixed number of MCMC steps (100). For a more resource-efficient training, we experimented with reducing the number of steps and using a *replay buffer* to persist MCMC chains (Tieleman, 2008) instead of initalizing from noise at each training iteration, but we could not manage to find a stable training regime on all categories and had sporadic energy spikes. We leave the solution of these problems to later work and share only the results of *non-persistent* trainings for consistency over all dataset categories. During sampling, we don't accumulate gradients w.r.t. model parameters for later use at model parameter update step.

3.2 RESULTS

We test our approach on the problem of visual anomaly detection on the recently introduced MVTec dataset (Bergmann et al., 2019a), which is a dataset for benchmarking anomaly detection methods with a focus on industrial inspection. There are fifteen different object and texture categories in the dataset, and we train a separate EBM for each category. The training is performed using defect-free training images only and the model is not exposed to any defective images during training. Training of EBMs is known to be computationally demanding and hence the input images are downsized, to 128×128 to keep the training time manageable.

For a test input, we generate both the per-pixel anomaly scores, $a(\mathbf{x})$, and the per-frame anomaly scores, $A(\mathbf{x})$ as described in Section 2. These scores can be used to distinguish between good and defective data, at both the pixel level and the frame level. This effectively creates a binary classifier, whose performance is characterized by the receiver operating characteristics (ROC) curve. We tested performance with both the raw and standardized scores, and found that the standardized scores generally outperform the raw scores and hence, we report results using standardized scores. For the detection task, we report the area under the ROC curve (AUROC) in Table 1, and for the localization task, we report the AUROC in Table 2. For a comparison between raw and standardized scores, please refer to section A.2 in the Appendix. We note that although the EBM achieves respectable scores on both tasks, there is room for improvement.

4 CONCLUSION

This work presented an energy-based method for the semi-supervised anomaly detection and localization problem. We propose an alternative scoring function for image-level anomaly detection and spatial anomaly maps for pixel-level localization; both are derived from back-propagating gradients of the energy score w.r.t. the test image. We experiment our approach on the MVTec industrial dataset and benchmark against the state of the art. While the method is still in the early stages of development, initial results are encouraging with indications of how to obtain further improvements. These include using larger models with higher capacity to model the energy function, and exploring regimens to perform training in a more stable and less resource-intensive manner.

REFERENCES

Nilesh A. Ahuja, Ibrahima J. Ndiour, Trushant Kalyanpur, and Omesh Tickoo. Probabilistic modeling of deep features for out-of-distribution and adversarial detection. In *Bayesian Deep Learning workshop, NeurIPS*, 2019.

Jinwon An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. 2015.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592, Long Beach, CA, USA, June 2019a. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00982. URL https://ieeexplore.ieee.org/document/8954181/.

Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2019b. doi: 10.5220/0007364503720380. URL http://dx.doi.org/10.5220/0007364503720380.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, 2000.

Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders, 2018.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv:1511.07289 [cs]*, February 2016. URL http://arxiv.org/abs/1511.07289. arXiv: 1511.07289.

Niv Cohen and Yedid Hoshen. Sub-Image Anomaly Detection with Deep Pyramid Correspondences. *arXiv:2005.02357 [cs]*, February 2021. URL http://arxiv.org/abs/2005.02357. arXiv: 2005.02357.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization, 2020.

Yilun Du and Igor Mordatch. Implicit Generation and Generalization in Energy-Based Models. *arXiv:1903.08689 [cs, stat]*, November 2019. URL http://arxiv.org/abs/1903.08689. arXiv: 1903.08689.

Markus Goldstein and Seiichi Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE*, 11(4), 04 2016.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One. *arXiv:1912.03263 [cs, stat]*, December 2019. URL http://arxiv.org/abs/1912.03263. arXiv: 1912.03263.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster based local outliers. *Pattern Recognition Letters*, 2003:9–10, 2003.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019.

Geoffrey E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, August 2002. ISSN 0899-7667, 1530-888X. doi: 10.1162/089976602760128018. URL http://www.mitpressjournals.org/doi/10.1162/089976602760128018.

Heiko Hoffmann. Kernel pca for novelty detection. *Pattern Recognition*, 40(3):863–874, 2007.

Edwin M. Knorr, Raymond T. Ng, and Vladimir Tucakov. Distance-based outliers: Algorithms and applications. *The VLDB Journal*, 8(3–4):237–253, 2000.

G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib. Novelty detection through model-based characterization of neural networks. In *IEEE International Conference on Image Processing (ICIP)*, pp. 3179–3183, 2020. doi: 10.1109/ICIP40778.2020.9190706.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A Tutorial on Energy-Based Learning. pp. 59, 2006.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. 2020. URL http://arxiv.org/abs/2010.03759. version: 3.

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning Non-Convergent Non-Persistent Short-Run MCMC Toward Energy-Based Model. *arXiv:1904.09770 [cs, stat]*, November 2019. URL http://arxiv.org/abs/1904.09770. arXiv: 1904.09770.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pp. 14707–14718, 2019.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Gregoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Muller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, pp. 1–40, 2021. ISSN 1558-2256. doi: 10.1109/jproc.2021.3052449. URL http://dx.doi.org/10.1109/JPROC.2021.3052449.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, 2017.

Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pp. 1064–1071, Helsinki, Finland, 2008. ACM Press. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390290. URL http://portal.acm.org/citation.cfm?doid=1390156.1390290.

Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. pp. 8, 2011.

Jihun Yi and Sungroh Yoon. Patch SVDD: Patch-level SVDD for Anomaly Detection and Segmentation. *arXiv:2006.16067 [cs]*, July 2020. URL http://arxiv.org/abs/2006.16067. arXiv: 2006.16067.

Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. 2016. URL http://arxiv.org/abs/1605.07717.

A APPENDIX

A.1 CNN TOPOLOGY

Table 3: CNN topology

kernel	**stride**	**padding**	f_{in}	f_{out}	**activation**
(3x3)	1	1	$n_c = 1, 3$	32	ELU
(4x4)	2	1	32	64	ELU
(4x4)	2	1	64	128	ELU
(4x4)	2	1	128	256	ELU
(4x4)	2	1	256	256	ELU
(4x4)	2	1	256	256	ELU
(4x4)	1	0	256	1	N/A

A.2 RAW VS STANDARDIZED SCORES

We compare AUROC scores for both detection and localization using raw gradient-map scores and standardized scores. In general, we observe that standardized scores perform better than raw scores. For a few categories, the raw scores do provide better AUROC.

Table 4: Comparing Raw and Standardized AUROC scores for detection

	Energy	Raw gradients	Standardized gradients
Average	0.56	0.69	0.72

Table 5: Comparing Raw and Standardized AUROC scores for localization

Category	Raw	Standardized	Difference
Carpet	0.53	0.63	0.10
Grid	0.86	0.86	0.00
Leather	0.43	0.86	0.43
Tile	0.50	0.57	0.07
Wood	0.73	0.74	0.01
Bottle	0.70	0.72	0.02
Cable	0.46	0.56	0.10
Capsule	0.44	0.64	0.20
Hazelnut	0.73	0.78	0.05
Metal Nut	0.71	0.65	-0.06
Pill	0.71	0.74	0.03
Screw	0.88	0.87	-0.01
Toothbrush	0.82	0.68	-0.14
Transistor	0.74	0.74	0.00
Zipper	0.64	0.55	-0.09

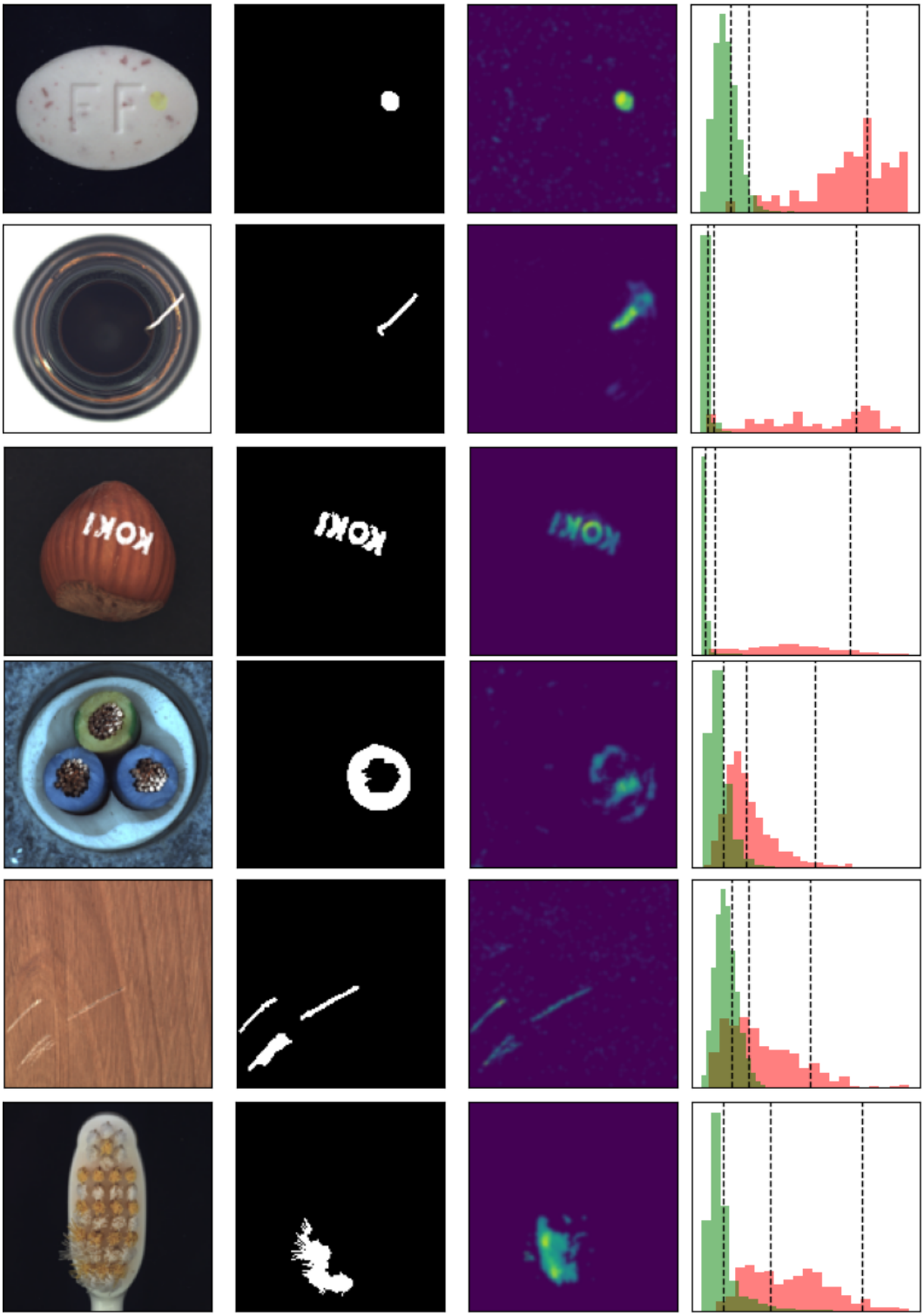


Figure 3: From left to right: sample images with anomalies, their corresponding ground truth masks, anomaly score maps, and histograms (good vs. defective pixels). Vertical dashed lines are *three-sigma rule* thresholds