Reimagining Anomalies: What if Anomalies Were Normal?

Philipp Liznerski[1*], Saurabh Varshneya[1*], Ece Calikus[2], Puyu Wang[1], Alexander Bartscher[1], Sebastian Josef Vollmer[1,3], Sophie Fellenz[1], Marius Kloft[1]

[1]RPTU University Kaiserslautern-Landau, Germany
[2]Uppsala University, Sweden
[3]German Research Center for Artificial Intelligence (DFKI), Germany

Abstract

Deep learning-based methods have achieved a breakthrough in image anomaly detection, but their complexity introduces a considerable challenge to understanding why an instance is predicted to be anomalous. We introduce a novel explanation method that generates multiple alternative modifications for each anomaly, capturing diverse concepts of anomalousness. Each modification is trained to be perceived as normal by the anomaly detector. The method provides a semantic explanation of the mechanism that triggered the detector, allowing users to explore "what-if scenarios." Qualitative and quantitative analyses across various image datasets demonstrate that applying this method to state-of-the-art detectors provides high-quality semantic explanations.

Code — https://github.com/liznerski/counterfactual-xad
Appendix — https://arxiv.org/abs/2402.14469

1 Introduction

Anomaly detection identifies patterns that deviate from normal behavior, the so-called *anomalies*. These anomalies can correspond to crucial actionable information in various domains such as medicine, manufacturing, and environmental monitoring (Chandola, Banerjee, and Kumar 2009).

Recently, deep learning has shown tremendous success in anomaly detection (AD), reducing error rates to approximately 1% in numerous image benchmarks (Reiss et al. 2021; Ruff et al. 2021). However, detectors based on deep learning lack the out-of-the-box interpretability of their traditional counterparts, making it difficult to understand the reasoning behind their predictions (Liznerski et al. 2021). Their lack of transparency is particularly concerning in sectors where safety is crucial and in situations where building trust is essential (Samek et al. 2020). Understanding modern anomaly detectors is a major challenge in contemporary AD and a necessary step before using AD in decision-making systems (Ruff et al. 2021).

Although feature-attribution techniques such as anomaly heatmaps (Roth et al. 2022) have been explored, they do not explain the underlying semantics relevant to the decision-making of the detectors. In domains beyond AD, counterfactual explanation (CE) has emerged as a popular alternative. CE generates synthetic samples that change the model's prediction with minimal alterations to the original sample (Ghandeharioun et al. 2021).

In this paper, we propose the use of CE to explain image anomaly detectors. While prevalent approaches identify anomalous regions within images, the presented technique generates a set of counterfactual examples for each anomaly, capturing diverse disentangled aspects (see Figure 1). Our goal is to explain why a detector flags an instance as anomalous. The framework provides semantic CEs that reveal when detectors rely on spurious or biased cues (Appendix A), e.g., labeling portraits as anomalous due to skin

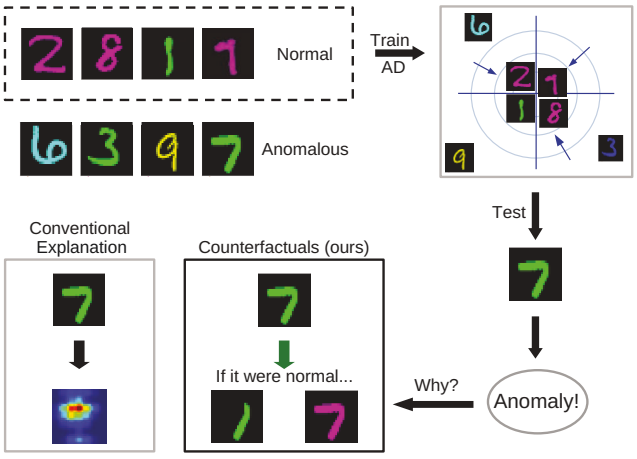


Figure 1: The figure illustrates the benefit of counterfactual explanation of anomaly detectors over traditional methods, using a dataset of handwritten digits in various colors. The normal data (top left) consist of pink digits and instances of the digit one in any color. An example anomaly—a green seven—is shown on the right. Conventional explanation methods localize the anomaly in the image and highlight it on a heatmap (bottom left). In contrast, the proposed method transforms the anomaly into multiple counterfactuals, addressing the crucial question: "How must the anomaly be altered to appear normal to the detector?"

*These authors contributed equally.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tone or background context. These insights support accountable and lawful deployment of AD systems.

2 Related Work

In the past decade, research has increased on improving the interpretability of neural networks. This increase is driven by the growing use of ML in decision-making systems, where transparency of predictions is crucial and even legally mandated in many countries (Neuwirth 2022).

Explanation of image AD. Research in explainable image AD has primarily focused on feature attribution methods, pinpointing image areas that influence predictions. Feature attribution methods trace an importance score from the model output back to the pixels (Selvaraju et al. 2017; Zhang et al. 2018a) or alter parts of the image and measure the impact on the model output (Fong and Vedaldi 2017; Dhurandhar et al. 2018). Some of these approaches have been applied to AD (Liznerski et al. 2021; Li et al. 2021). Several methods generate explanations using generative models or autoencoders, where the pixel-wise reconstruction error yields an anomaly heatmap (Bergmann et al. 2019; Dehaene et al. 2020). Others use fully convolutional architectures (Liznerski et al. 2021) or transfer learning (Defard et al. 2021; Roth et al. 2022). All these methods identify regions within an image that influence the detector's prediction; however, they do not explain the detectors at a higher semantic level (Adebayo et al. 2018; Varshneya et al. 2024).

Counterfactual explanation of neural networks on images. CE methods (Guidotti 2022) identify the necessary changes in the input to alter the model prediction in a specific way. Such explanations can provide profound insights that enhance comprehension of model behavior and align more closely with human cognitive processes (Pearl 2009). Existing CE algorithms are designed primarily for supervised learning on tabular data (Wachter, Mittelstadt, and Russell 2017; Mothilal, Sharma, and Tan 2020; Guidotti 2022). A few studies have also explored the application of CE to image classification (Goyal et al. 2019; Ghandeharioun et al. 2021; Abid, Yuksekgonul, and Zou 2022; Singla et al. 2023). DISSECT (Ghandeharioun et al. 2021) is notable for its ability to generate multiple CEs with disentangled high-level concepts. Recent work started exploring CE for supervised image AD. Studies by Sanchez et al. (2022); Wolleb et al. (2022); Siddiqui et al. (2024); Ahamed, Xu, and Rahmim (2024); Fontanella et al. (2024) utilize diffusion models guided by text prompts or learnable conditions to generate normal counterparts of abnormal medical images. These approaches rely on supervised learning, framing the AD problem as a classification task. They fine-tune pretrained diffusion models or use classifier-guidance, utilizing both normal and ground-truth anomalies. Although these approaches are promising for explaining model decisions with counterfactuals, they are applicable only to supervised settings, making them unsuitable for unsupervised AD models.

Counterfactual Explanations of AD. Virtually all CE methods for AD have been applied to tabular data (Angiulli et al. 2023; Datta, Chen, and Ramakrishnan 2022; Han et al. 2023) and time series (Sulem et al. 2022; Cheng et al. 2022). These methods use knowledge graphs or structural causal models to generate CEs for categorical features (Datta, Chen, and Ramakrishnan 2022; Han et al. 2023) or take advantage of temporal aspects (Sulem et al. 2022; Cheng et al. 2022). Some of these methods have been applied to fairness (Han et al. 2023) and algorithmic recourse (Datta, Chen, and Ramakrishnan 2022). None of these methods is applicable to image data. Very recently, Ji et al. (2024) introduced AR-Pro, a CE approach to explain anomalies in images. However, this method performs defect repair by modifying localized faulty regions with mask supervision, merely transforming defective regions into their normal versions. Our framework instead explains semantic anomalies without localization, identifying conceptual attributes (background, color, bias) that drive a detector's decision. As shown in Appendix B, it also handles defect-type anomalies as a special case: on industrial data similar to AR-Pro it produces repair-like CEs but generalizes to multi-concept and semantic anomalies.

3 Counterfactual Explanation of Image AD

We formally present a novel framework for generating CEs of image AD. We first define the general setup and then explain how to use GANs and diffusion models to produce CEs. To the best of our knowledge, this approach is the first to explain semantic image AD using CE.

3.1 What if the Anomaly were Normal?

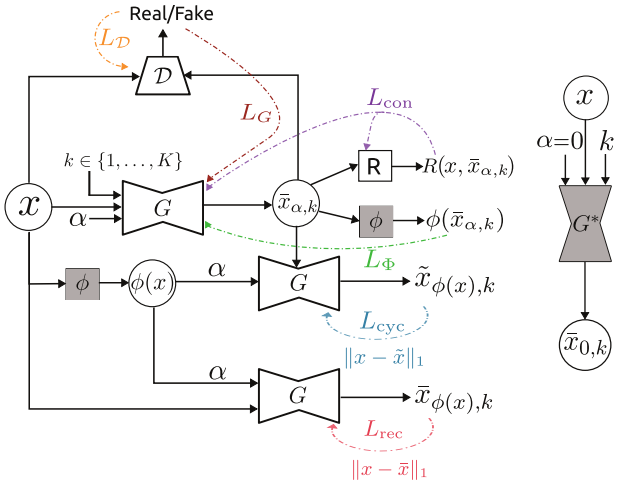
Our aim is to provide explanations for a given anomaly detector $\phi : \mathbb{R}^D \to [0, 1]$ that maps an image $x \in \mathbb{R}^D$ to an anomaly score $\alpha \in [0, 1]$. We define a CE for the detector ϕ and perceived anomaly $\boldsymbol{x}^* \in \mathbb{R}^D$ (i.e., $\phi(\boldsymbol{x}^*) \gg 0$) as a modified sample $\bar{\boldsymbol{x}}^*$ with $\phi(\bar{\boldsymbol{x}}^*) \approx 0$ and $\|\bar{\boldsymbol{x}}^* - \boldsymbol{x}^*\|_1 \leq \epsilon$ for an $\epsilon \geq 0$. In other words, a CE must be normal according to ϕ, while being minimally changed w.r.t. the original anomaly $\boldsymbol{x}^*$. Thus, CEs address the question: "What if the anomaly $\boldsymbol{x}$ were normal?", explaining the behavior of the anomaly detector at a high semantic level.

To produce such CEs for deep AD, we need to train a generator $G : \mathbb{R}^D \to \mathbb{R}^D$ to yield $G(\boldsymbol{x}^*) = \bar{\boldsymbol{x}}^*$. However, normal images can differ from anomalies in multiple ways, and thus multiple CEs may be required to adequately explain an anomaly. We want the generator to consider multiple categorical concepts $k \in \{1, \ldots, K\}$. Thus, the generator is now of the form $G : \mathbb{R}^D \times \{1, \ldots, K\} \to \mathbb{R}^D$ and is supposed to produce $G(\boldsymbol{x}^*, k) = \bar{\boldsymbol{x}}^*_k$ with $\|\bar{\boldsymbol{x}}^*_k - \bar{\boldsymbol{x}}^*_{k'}\|_1 \geq \epsilon'$ for any $k \neq k'$.

The same data $\{(\boldsymbol{x}_0, y_0), \ldots, (\boldsymbol{x}_n, y_n)\}$ used to train ϕ can also train G. Note that in AD, training labels y_i are typically unknown, and most samples are assumed normal.

3.2 Deep Generative Models for CE of Image AD

In practice, it has been found beneficial to train the generator to produce sequences of CEs with increasing impact on a classifier's output (Ghandeharioun et al. 2021). The proposed approach is based on this idea. We modify the generator $G : \mathbb{R}^D \times [0, 1] \times \{1, \ldots, K\} \to \mathbb{R}^D$ to consider a target



(a) Training (b) Test

Figure 2: Schematic overview of the proposed CE framework for explaining a black box image AD model ϕ. (a) shows the training losses and their impact on the different models. (b) shows the inference, where the target anomaly score is zero. Gray nodes (i.e., the trained AD model ϕ and generator G^*) represent models that are not optimized.

score α, aiming for the trained G to produce a sample with an anomaly score of approximately α.

GANs for CE of image AD. Here, we first introduce a GAN-based (Goodfellow et al. 2020) model for producing CEs of image AD. In particular, we train G as a concept-disentangled GAN. We define a discriminator $\mathcal{D} : \mathbb{R}^D \to \mathbb{R}$ and a concept classifier $R : \mathbb{R}^D \times \mathbb{R}^D \to [0, 1]^K$. $\mathcal{D}$ is trained to distinguish between generated $\bar{\boldsymbol{x}}_{\alpha,k} = G(\boldsymbol{x}, \alpha, k)$ and true samples from the dataset, encouraging *realistic* outcomes. R classifies the concept k for a sample $\bar{\boldsymbol{x}}_{\alpha,k}$, encouraging the generated samples to be *concept-disentangled* on a semantic level. Further losses encourage the generator to incur *minimal changes* on the original sample $\boldsymbol{x}$ and to yield target anomaly scores α (i.e., $\phi(\bar{\boldsymbol{x}}_{\alpha,k}) \approx \alpha$). The proposed objective summarizes to

$$\min_{G,R} \max_{\mathcal{D}} \mathbb{E}_{\boldsymbol{x} \sim P_X} \mathbb{E}_{\alpha,k} \big[\lambda_{gan} \left(L_{\mathcal{D}}(\mathcal{D}) + L_G(G) \right) - L_\phi(G) \\ \lambda_\phi + \lambda_{rec} \left(L_{rec}(G) + L_{cyc}(G) \right) + \lambda_r L_{con}(G, R) \big], \quad (1)$$

where p_X denotes the training data distribution, and $\lambda_{gan}, \lambda_\phi, \lambda_{rec}$, and λ_r are some constant factors. Figure 2 shows a schematic overview of the framework.

In the following, we will explain the various losses, starting with $L_\phi(G)$, which encourages for $\bar{\boldsymbol{x}}_{\alpha,k}$ an anomaly score of α. $L_\phi(G)$ can be any loss that measures the divergence of target anomaly scores α and true anomaly scores $\phi(\bar{\boldsymbol{x}}_{\alpha,k})$; for example, the L_2 distance or the KL-divergence. We assume the detector ϕ to be bounded (w.l.o.g, $0 \leq \phi(\boldsymbol{x}) \leq 1$). In our experiments, we found that a continuous binary cross entropy loss produces the best results:

$$L_\phi(G) = \alpha \log \big(\phi(\bar{\boldsymbol{x}}_{\alpha,k})\big) + \big(1 - \alpha\big) \log \big(1 - \phi(\bar{\boldsymbol{x}}_{\alpha,k})\big).$$

The losses $L_{\mathcal{D}}(\mathcal{D})$ and $L_G(G)$ can be any discriminative and generative GAN losses, respectively. We specifically experimented with spectral normalization (Miyato et al. 2018) and the hinge loss (Miyato and Koyama 2018):

$$L_G(G) = -\min(0, -1 + \mathcal{D}(\bar{\boldsymbol{x}}_{\alpha,k})),$$

$$L_{\mathcal{D}}(\mathcal{D}) = \min(0, -1 + \mathcal{D}(\boldsymbol{x})) + \min(0, -1 - \mathcal{D}(\bar{\boldsymbol{x}}_{\alpha,k})).$$

The reconstruction loss makes G recreate $\boldsymbol{x}$ for every concept k, when conditioned on $\boldsymbol{x}$ and its "true" score $\phi(\boldsymbol{x})$:

$$L_{rec}(G) = \|\boldsymbol{x} - G(\boldsymbol{x}, \phi(\boldsymbol{x}), k)\|_1 .$$

This ensures that G remains unchanged when the sample already has the targeted anomaly score, overall encouraging minimal changes.

Similarly, the "cycle consistency loss" (Zhu et al. 2017)

$$L_{cyc}(G) = \|\boldsymbol{x} - \tilde{\boldsymbol{x}}_{\alpha,k}\|_1 ,$$

where $\tilde{\boldsymbol{x}}_{\alpha,k} = G(\bar{\boldsymbol{x}}_{\alpha,k}, \phi(\boldsymbol{x}), k)$, encourages G to recreate the sample $\boldsymbol{x}$ when targeting its true anomaly score $\phi(\boldsymbol{x})$ and being conditioned on any generated sample $\bar{\boldsymbol{x}}_{\alpha,k}$ based on $\boldsymbol{x}$. It encourages minimal changes because the generator needs to be able to revert any change of $\boldsymbol{x}$.

The concept loss drives G to produce disentangled CEs:

$$L_{con}(G, R) = \mathbb{C}\left(k, R\big(\boldsymbol{x}, \bar{\boldsymbol{x}}_{\alpha,k}\big)\right) + \mathbb{C}\left(k, R\big(\bar{\boldsymbol{x}}_{\alpha,k}, \tilde{\boldsymbol{x}}_{\alpha,k}\big)\right),$$

where $\mathbb{C}$ denotes the cross entropy loss.

In summary, the losses encourage the generated samples $\bar{\boldsymbol{x}}_{\alpha,k}$ to be semantically distinguishable for different concepts k while having an anomaly score of α according to ϕ and undergoing minimal changes with respect to the original $\boldsymbol{x}$. This results in a disentangled set of K counterfactual explanations for an anomaly $\boldsymbol{x}^*$ with $\{G(\boldsymbol{x}^*, 0, 1), \ldots, G(\boldsymbol{x}^*, 0, K)\}$. The generator can produce pseudo anomalies $G(\boldsymbol{x}, \alpha, K)$ when $\phi(\boldsymbol{x}) \approx 0$ and $\alpha \gg 0$, which help G in learning how to turn anomalies into normal samples, when included in L_ϕ.

Diffusion Models for CE of Image AD. This section proposes an approach for producing CEs of high-resolution anomaly detectors using state-of-the-art diffusion models. In particular, we incorporate DiffEdit (Couairon et al. 2023). DiffEdit modifies the LAION-5B pre-trained text-conditional latent diffusion model known as Stable Diffusion (Rombach et al. 2022) to semantically edit images. Let $A_{\mathcal{E}} : \mathbb{R}^D \to \mathbb{R}^\Delta$ and $A_\Omega : \mathbb{R}^\Delta \to \mathbb{R}^D$ denote the encoder and decoder of the autoencoder used in Stable Diffusion. From a high-level perspective, the DiffEdit model can be defined as $\psi : \mathbb{R}^{\Delta \times T} \to \mathbb{R}^\Delta$ where T denotes the output dimension of the word embedding model. For an image $\boldsymbol{x} \in \mathbb{R}^D$, we retrieve a semantically modified version $\hat{\boldsymbol{x}}$ controlled by the text prompt t via $\hat{\boldsymbol{x}} = A_\Omega(\psi(A_{\mathcal{E}}(\boldsymbol{x}), t))$. For more details, we refer to the paper by Couairon et al. (2023). We incorporate DiffEdit into the proposed framework by training the generator on its latent output. That is, we redefine the generator $G(\boldsymbol{x}, \alpha, k) = A_\Omega\left(G'(\psi(A_{\mathcal{E}}(\boldsymbol{x}), t), \alpha, k)\right)$ with $G' : \mathbb{R}^\Delta \times [0, 1] \times \{1, \ldots, K\} \to \mathbb{R}^\Delta$. The text prompt t is set to the normal class label (e.g., "cat" for cats being normal). We train the generator G (i.e., the parameters of G') as before.

4 Theoretical Analysis

The objective of the proposed method is intertwined with several interacting losses. Here, we provide a theoretical analysis on the performance of the optimization problem. Let $V(\mathcal{D}, G) = \lambda_{gan}\mathbb{E}_{\boldsymbol{x}\sim p_X}\mathbb{E}_{\alpha,k}[L_{\mathcal{D}}(\mathcal{D})]$ and $U(\mathcal{D}, (G, R)) = \mathbb{E}_{\boldsymbol{x}\sim p_X}\mathbb{E}_{\alpha,k}[\lambda_{gan}L_G(G) - \lambda_\phi L_\phi(G) + \lambda_{rec}(L_{rec}(G)+L_{cyc}(G))+\lambda_r L_{con}(G, R)]$. V trains $\mathcal{D}$, and U trains G and R.

Definition 4.1. *We say $(\mathcal{D}^*, (G^*, R^*))$ is a Nash equilibrium of the system if $V(\mathcal{D}, G^*) \leq V(\mathcal{D}^*, G^*)$ for any $\mathcal{D}$ and $U(\mathcal{D}^*, (G^*, R^*)) \leq U(\mathcal{D}^*, (G, R))$ for any G, R.*

Theorem 4.2. *Assume G and R have enough capacity. Let $(\mathcal{D}^*, (G^*, R^*))$ be a Nash equilibrium of the system. (I) If $\lambda_\phi = \lambda_{rec} = \lambda_{con} = 0$, then $\mathbb{E}_{\alpha,k}[p_{G^*(\alpha,k)}] = p_X$ and $V(\mathcal{D}^*, G^*) = -2\lambda_{gan}$. (II) If $\lambda_\phi = 0$ and ϕ is nearly flat, then $\mathbb{E}_{\alpha,k}[p_{G^*(\alpha,k)}] \approx p_X$ and $V(\mathcal{D}^*, G^*) \approx -2\lambda_{gan}$. If we assume ϕ is flat, then $\mathbb{E}_{\alpha,k}[p_{G^*(\alpha,k)}] = p_X$ and $V(\mathcal{D}^*, G^*) = -2\lambda_{gan}$.*

Part (I) of Theorem 4.2 shows that when training only with the modified GAN-based objectives $L_{\mathcal{D}}$ and L_G, the generator indeed converges to the training data distribution p_X, similar to the original objective in Goodfellow et al. (2020). Part (II) shows that when including the losses that encourage minimal changes, we still obtain the distribution p_X with flatness assumptions on the detector ϕ. It follows that L_ϕ is the main antagonist that causes divergence from the training data distribution to produce reasonable CEs.

Proof of Theorem 4.2. Appendix C provides the full proofs. Here, we show very brief sketches. The main idea for proving part (I) of the theorem is to divide the min-max problem into a min part and a max part and then analyze them separately. The maximizer $\mathcal{D}^*$ of this problem has the explicit form $\mathcal{D}^*(\boldsymbol{x}) = 1$ if $\mathbb{E}_{\alpha,k}[p_{G^*\alpha,k}](\boldsymbol{x}) \leq p_X(\boldsymbol{x})$ and $\mathcal{D}^*(\boldsymbol{x}) = 0$ otherwise. Plugging this into $V(\mathcal{D}^*, G^*)$, we get $V(\mathcal{D}^*, G^*) \geq -2\lambda_{gan}$. According to the property of the Nash equilibrium, we know $\mathbb{E}_{\boldsymbol{x},\alpha,k}[L_G(G^*)] \leq \mathbb{E}_{\boldsymbol{x},\alpha,k}[L_G(G)]$ holds for any G. Specifically considering G as the "ideal" generator with a density function $p_{G(\alpha,k)} = p_X$, we can establish $V(\mathcal{D}^*, G^*) \leq -2\lambda_{gan}$. For part (II), the analysis of V and $\mathcal{D}^*$ are the same as for the max problem. For the min problem, from the flatness of ϕ, we can show that $\int p_X(\boldsymbol{x})\mathcal{D}^*(\boldsymbol{x}) - \mathbb{E}_{\alpha,k}[p_{G^*(\alpha,k)}](\boldsymbol{x})\mathcal{D}^*(\boldsymbol{x})d\boldsymbol{x} \leq \epsilon$. Here, $\epsilon > 0$ is proportional to the flatness of ϕ. If ϕ is almost flat, $\epsilon \approx 0$ and then $V(\mathcal{D}^*, G^*) \approx -2\lambda_{gan}$. Hence, $\mathbb{E}_{\alpha,k}[p_{G^*(\alpha,k)}] \approx p_X$. If ϕ is flat, then $\epsilon = 0$. It implies $V(\mathcal{D}^*, G^*) = -2\lambda_{gan}$ and $\mathbb{E}_{\alpha,k}[p_{G^*(\alpha,k)}] = p_X$. □

Theorem 4.3. *Assume G and R have enough capacity, $\lambda_\phi > 0$, and Prob$(\phi(\boldsymbol{x}) = 0 \cup \phi(\boldsymbol{x}) = 1) > 0$, $\boldsymbol{x} \sim p_X$. Let $(\mathcal{D}^*, (G^*, R^*))$ be a Nash equilibrium of the system. Then $p_{G^*(\alpha,k)} \neq p_X$ for any α, k and $V(\mathcal{D}^*, G^*) \neq -2\lambda_{gan}$.*

Theorem 4.3 shows that L_ϕ indeed causes a divergence from the training data distribution, implying that the generator learns to map samples to anomalous data regimes when $\alpha > 0$. Empirically, our experiments show that with $\alpha = 0$, the generator consistently creates normal samples.

Proof of Theorem 4.3. We prove the theorem by contradiction. For a Nash equilibrium $(\mathcal{D}^*, (G^*, R^*))$, it holds $U(\mathcal{D}^*, (G^*, R^*)) \leq U(\mathcal{D}^*, (G, R))$ for any G, R. We show that if $(\mathcal{D}^*, (G^*, R^*))$ is a Nash equilibrium with $p_{G^*(\alpha,k)} = p_X$ and $V(\mathcal{D}^*, G^*) = -2\lambda_{gan}$, then there exists a generator G' such that $U(\mathcal{D}^*, (G', R^*)) < U(\mathcal{D}^*, (G^*, R^*))$. This violates $U(\mathcal{D}^*, (G^*, R^*)) \leq U(\mathcal{D}^*, (G', R^*))$. Hence, $p_{G^*(\alpha,k)} \neq p_X$ and $V(\mathcal{D}^*, G^*) \neq -2\lambda_{gan}$. We choose G' as satisfying $\phi(G'(\boldsymbol{x}, \alpha, k)) = \alpha$ and being Lipschitz continuous w.r.t. the first and second argument. By noting that $\mathbb{E}_{\boldsymbol{x},\alpha,k}L_\phi(G^*) = -\infty$ when Prob$(\phi(\boldsymbol{x}) = 0 \cup \phi(\boldsymbol{x}) = 1) > 0$ and $U(\mathcal{D}^*, (G', R^*))$ is uniformly bounded, we conclude $U(\mathcal{D}^*, (G', R^*)) < U(\mathcal{D}^*, (G^*, R^*))$. □

5 Experiments

We empirically assess the capabilities of CEs for deep AD, providing qualitative and quantitative evidence of the superiority of the proposed CEs over traditional methods.

Deep Anomaly Detection Methods. We specifically study three state-of-the-art *semantic image AD* methods. *DSVDD* (Ruff et al. 2018) trains a neural network to enclose the (mostly normal) unlabeled training data by a minimal volume hypersphere with the distance to the sphere's center becoming the anomaly score. Hendrycks, Mazeika, and Dietterich (2019) showed that Outlier Exposure (OE)—using a large unstructured collection of natural images as example anomalies during training—consistently outperforms previous AD methods, while still being unsupervised. A neural network learns to differentiate normal data from OE samples with a *Binary Cross Entropy (BCE)* loss. Liznerski et al. (2022) introduced *HSC* as a modification of the DSVDD loss to enable it to take advantage of OE. Since the CE generator requires bounded anomaly scores, we slightly adjust some of the objectives without impacting the performance. A more detailed description can be found in Appendix J.

Datasets. We evaluate on *MNIST* (Deng 2012), *Colored-MNIST*, *CIFAR-10* (Krizhevsky, Hinton et al. 2009), and *GTSDB* (Houben et al. 2013). Furthermore, we introduce ImageNet-Neighbors (INN), a subset of ImageNet-1k (Russakovsky et al. 2015) designed for AD tasks. INN comprises multiple AD setups; in each setup, one ImageNet-1k class is considered normal, and the ten most semantically similar classes, based on the Wu-Palmer similarity metric (Wu and Palmer 1994), are defined as ground-truth test anomalies. We follow previous work for using disjoint OE data for the OE-based AD methods. Details are in Appendix K.

Experimental Setup. Following previous work on semantic image AD (Ruff et al. 2018; Golan and El-Yaniv 2018; Hendrycks et al. 2019; Tack et al. 2020; Ruff et al. 2021; Liznerski et al. 2022), we convert the multi-class classification datasets into AD benchmarks. This is achieved by defining a subset of the classes to be normal and using the remaining classes as ground-truth anomalies during testing. When only one class is considered normal, this approach is known as one vs. rest. We also explore a variation in which multiple classes are normal. This emulates a multifaceted

normal class that includes different notions of normality. Finally, we consider the INN setup where we have particular ground-truth anomalies per normal class. For INN, we generate CEs with diffusion models, while for all other datasets we use the pure GAN-based model.

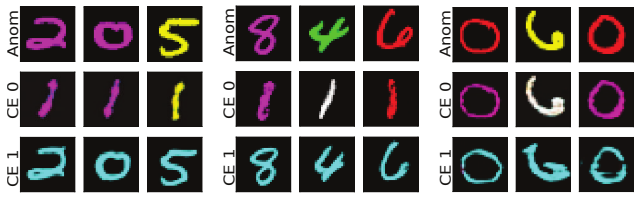
Overall, we consider over 80 different AD setups. Details of the specific AD setups are provided in Appendix O. Our quantitative analysis reports results averaged over all scenarios and multiple seeds per dataset. Detailed quantitative results for each scenario are in Appendix O and further qualitative results in Appendix P. Hyperparameters and architectures are explained in detail in Appendix M, and there is a small hyperparameter sensitivity analysis in Appendix I.

5.1 Qualitative Results

We present qualitative examples of CEs, demonstrating the benefit of using CE for semantic image AD over traditional explanation methods. We refer to Appendix P for more examples using further normal scenarios with similar findings.

CEs Explain why Images are Predicted Anomalous

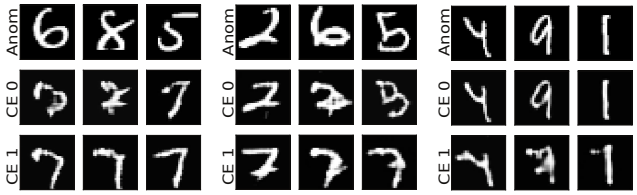
Colored-MNIST. Figure 3 shows CEs when the normal class is formed from instances of the digit one and digits colored cyan. We observe that the CEs generated to explain the BCE detector align well with our expectation. The proposed method transforms the anomalies into ones without changing the color, or their color is changed to cyan without changing the digit. Both modifications are minimal alterations of the anomaly, transforming its appearance to normality in two distinct ways. The CEs of HSC also correspond to normal samples. However, in one case, both the color and the digit is changed, resulting in unnecessary changes. We found that this behavior represents a local optimum of the proposed objective, highlighting the inherent difficulty of the unsupervised generation of CEs. The CEs created to explain the DSVDD detector perform least effectively. They appear normal for one concept but often fail for the other. This behavior may be attributed to DSVDD's limited ability to detect anomalies. See Appendix D for details.



(a) BCE (OE) (b) HSC (OE) (c) DSVDD

Figure 3: Examples of CEs for Colored-MNIST, with cyan digits and the digit one serving as the normal class. The first row shows anomalous images and the next two rows their corresponding CEs using two concepts. The CEs of BCE and HSC appear normal and realistic for each concept.

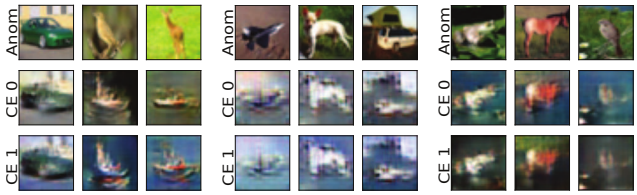
MNIST. In Figure 4, a seven is considered normal. The CEs of BCE and HSC are meaningful: the anomalies are transformed into plausible variations of seven. Since the normal class consists of a single, monochromatic digit, the generator primarily learns to manipulate the presence or absence of the horizontal bar characteristic of a seven to distinguish between concepts. Therefore, the CEs indicate that the detectors do not heavily rely on the horizontal bar to rate anomalousness properly.



(a) BCE (OE) (b) HSC (OE) (c) DSVDD

Figure 4: CEs for MNIST with seven as the normal class. The first row shows anomalies, the other two rows CEs using two different concepts. CEs of BCE and HSC are variations of seven and thus represent intuitive counterfactuals.

CIFAR-10. Especially for BCE, the CEs in Figure 5 represent intuitive normal samples (ships) that retain the anomalous object's color to incur minimal changes on the anomaly. As there is only one single normal class, the CEs primarily disentangle the concepts by changing the background. Ships are typically depicted floating on water, which may vary in color. The CEs reveal that HSC and DSVDD predominantly focus on the background to detect anomalies, as all CEs are perceived as normal (see Section 5.2), although the CEs' foregrounds are often similar to the original anomaly. This aligns with prior findings on background exploitation in CIFAR-10 (Ding and Pang 2024).



(a) BCE (OE) (b) HSC (OE) (c) DSVDD

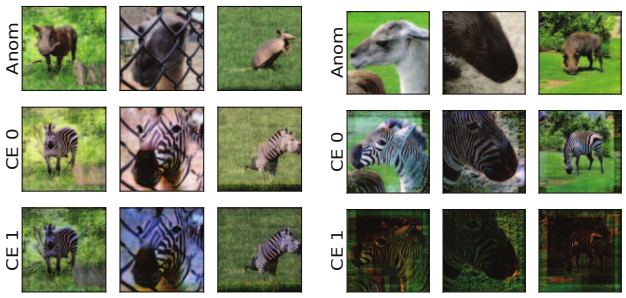
Figure 5: Example CEs for CIFAR-10 when ships are normal. The rows show anomalies and CEs for two concepts, respectively. The CEs of BCE display normal ships, varying the background for successful disentanglement while keeping the object's color to avoid unnecessary changes.

GTSDB. Figure 6 shows the proposed CEs when speed signs are taken as a normal class. The CEs of BCE and HSC show well-disentangled normal traffic signs, obtained from anomalous ones. For instance, the CE of BCE changes the "80km/h restriction ends" sign into a "80km/h limit" sign—a minimal intervention to make the sample appear normal. Note that all triangular anomalies become circles. The CEs show that the shape is an important feature for the detector to rate anomalousness.

(a) BCE (OE) (b) HSC (OE) (c) DSVDD

Figure 6: CEs for GTSDB with speed signs being normal. The first row shows anomalies, the other rows present disentangled CEs, which appear as different normal speed signs.

ImageNet-Neighbors. Figure 7 shows CEs for the INN dataset when zebras are normal. The ground-truth anomalies are "similar" animals, ranging from horses and boars to armadillos. The CEs depict zebras while keeping the general pose and background of the anomalous animal. For disentanglement, the CEs vary the color scheme, which apparently the detectors perceive as normal. The CEs for the second concept for HSC are dark and, while still showing zebras, perturb the image with green and orange patterns. Interestingly, our experiments show that HSC assigns lower anomaly scores to the CEs for this second concept.



(a) BCE (OE) (b) HSC (OE)

Figure 7: Examples of CEs for INN, where zebras are considered normal. The first row shows anomalous images, the other two rows present CEs using two different concepts.

MVTec-AD. *Our experiments focus on semantic image-AD* (Ruff et al. 2021) *rather than low-level AD* where anomalies are defects instead of out-of-class. Figure 8 shows CEs for such a scenario with the MVTec-AD dataset (Bergmann et al. 2019). While the CEs are of good quality

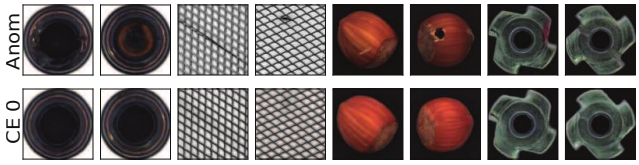
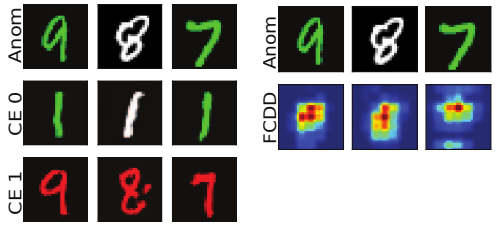


Figure 8: CEs of the BCE detector for MVTec-AD and ImageNet-21k as OE. The top row shows anomalies as defects. The bottom row shows corresponding CEs, resembling normal, healthy images.

and correct, they do not provide valuable insight, as here the behavior of an anomaly detector is less opaque and requires no counterfactual explanation. We include further reasoning for this and more results in Appendix B.

CEs Explain why Images are Predicted Anomalous—*even when Feature Attribution Fails*. We demonstrate the advantage of the proposed CEs over conventional explanations that attribute features. Figure 9 shows (a) CEs generated with our method and (b) heatmaps for the corresponding anomalies generated with FCDD (Liznerski et al. 2021). FCDD's heatmaps explain only spatial aspects of the AD: It highlights the horizontal bar in digit seven, the circle in digit nine, and all of digit eight. These spatial aspects are also explained by the CEs created for the first concept, where anomalies are turned into the digit one. However, FCDD's heatmaps fail to identify the color as anomalous, whereas our CEs capture this aspect with their second concept, where the anomalies are colored red, making them look normal. This shows that CEs can provide more holistic explanations.



(a) Counterfactuals (b) FCDD's Heatmaps

Figure 9: The first row shows anomalies from Colored-MNIST, with red digits and the digit one forming the normal class. The other rows show (a) corresponding CEs and (b) anomaly heatmaps from FCDD. The CEs explain the detector that perceives anomalies turned red or into one as normal, while heatmaps just highlight the difference to one.

5.2 Quantitative Results

This section presents a quantitative analysis of the CEs, assessing their normality, realism, and disentanglement in terms of metrics based on AuROC, FID, and accuracy. The metrics are described in detail in Appendix L.

The CEs Appear as Normal. An important attribute for any CE in AD is that it must be perceived as normal by the anomaly detector. To evaluate this quality criterion, we compare the anomaly scores of the normal test samples with those of the generated CEs in terms of AuROC. Ideally, the AuROC should approach 50%, indicating that CEs and normal samples are indistinguishable. As shown in Table 1, the AuROC is indeed very close to 50% on CIFAR-10, GTSDB, and Colored-MNIST (here abbreviated as C-MNIST), underlining that the detector perceives the CEs as normal. Only on MNIST and INN, some of the CEs appear anomalous. This might be due to the enforced disentanglement that produces diverse samples despite a limited variety of possible normal variations.

	Datasets	Methods		
		BCE OE	HSC OE	DSVDD
Single normal class	MNIST	72.0 ± 4.0	80.8 ± 5.3	75.2 ± 9.2
	CIFAR-10	47.5 ± 10.0	49.9 ± 4.4	54.6 ± 3.4
	INN	69.1 ± 18.1	67.9 ± 13.2	$\times$
Multiple normal classes	C-MNIST	55.6 ± 1.5	55.8 ± 4.7	61.5 ± 4.3
	MNIST	78.1 ± 4.1	82.1 ± 3.8	73.4 ± 6.5
	CIFAR-10	49.0 ± 8.5	44.4 ± 6.7	50.7 ± 3.3
	GTDSB	50.2 ± 8.0	48.6 ± 14.4	53.1 ± 4.8

Table 1: AuROC of normal test data vs. CEs. The CEs appear entirely normal for values $\leq 50\%$.

The CEs are Realistic. To assess the realism, we compute the FID (Heusel et al. 2017) between the CEs and normal test samples, and then normalize it by dividing by the FID between normal and anomalous test samples. The normalized FID_N is 100% if the CEs are as realistic as the anomalies. We found that a FID_N of 50 to 100% is a reasonable target for expressive CEs. If the CEs became too similar to the normal data distribution, they would not be valid counterfactuals, as they would not retain non-anomalous features from the anomalies. Table 2 displays the normalized FID_N scores. The CEs for BCE and HSC are mostly as realistic as the anomalies. On MNIST, INN, and Colored-MNIST, the CEs are even more realistic.

	Datasets	Methods		
		BCE OE	HSC OE	DSVDD
Single normal class	MNIST	43 ± 8.1	68 ± 14.6	100 ± 8.8
	CIFAR-10	116 ± 20.8	300 ± 90.0	116 ± 12.0
	INN	85.0 ± 28.6	85.4 ± 24.6	$\times$
Multiple normal classes	C-MNIST	56 ± 12.4	95 ± 30.5	83 ± 8.7
	MNIST	78 ± 26.0	96 ± 25.0	100 ± 10.7
	CIFAR-10	103 ± 27.9	254 ± 69.7	110 ± 10.0
	GTDSB	110 ± 101	95 ± 73.5	131 ± 118

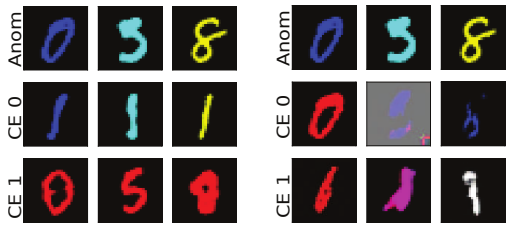
Table 2: FID_N for the CEs. Most of the CEs are as realistic as the anomalies, which are also realistic since they follow the general data distribution (e.g., are digits for MNIST).

The CEs Capture Multiple Disentangled Concepts. In Appendix E, we report the accuracy of the concept classifier. The classifier almost always achieves an accuracy of more than 90%. Exceptions are for DSVDD and MNIST, HSC and MNIST, and BCE and CIFAR-10, where it scores roughly 80%. This provides evidence that, for each anomaly, our method generates concept-disentangled CEs.

The CEs Are Minimal Modifications. To assess the minimality of the modifications present in the CEs, we report the LPIPS (Zhang et al. 2018b) and MSE between test anomalies and their corresponding CEs in Appendix F. The results indicate that the modifications are reasonably limited, as both metrics attain comparatively low values.

5.3 The CEs Reveal a Classifier Bias in Deep AD

The hypothesis of a "classification bias," suggesting supervised classifiers underperform when trained with limited and biased anomaly subsets (Ruff et al. 2020), remains insufficiently investigated. To test this hypothesis, we train a supervised classifier on Colored-MNIST to distinguish between a normal set (red digits and the digit one) and a subset of the ground-truth anomalies (all blue anomalies). This simulates a realistic scenario in which one has no access to all variations of the ground-truth anomalies. The classifier bias becomes apparent as the AuROC of normal test vs. ground-truth anomalies decreases from 98% for BCE with OE (unsupervised) to 75% for supervised BCE. Our CEs illuminate this phenomenon in Figure 10. The CEs of the AD method in (a) indicate that anomalies should be transformed into red or digit one to appear normal. For the supervised classifier in (b), only for the blue anomalous zero, which is seen during training, the CEs show normal versions of the anomaly. For unseen anomalies, such as the yellow eight, the CEs do not show intuitive normal images. This suggests that the classifier is biased towards blue anomalies and fails to generalize to colors not present in the training data. The experiment underlines the need for specialized AD methods (e.g., using OE or semi-supervision) because they are less prone to bias.



(a) BCE with OE (b) BCE with blue

Figure 10: The first row shows anomalies for C-MNIST with red digits and the digit one as the normal class. The other rows show CEs of BCE trained with OE in (a) and a classifier trained with only blue anomalies in (b). The generator's inability to generate normal-looking CEs for non-blue anomalies suggests that the classifier (b) is biased.

6 Conclusion

This paper introduced a novel method that can interpret image anomaly detectors at a semantic level. This is achieved by modifying anomalies until they are perceived as normal by the detector, creating instances known as counterfactuals. We found that counterfactuals can provide a deeper, more nuanced understanding of image anomaly detectors, far beyond the traditional feature-attribution level. Extensive experiments across various image benchmarks and deep anomaly detectors demonstrated the efficacy of the proposed approach, particularly also where conventional techniques fail. This research marks a paradigm shift and a significant departure from the more superficial interpretation of anomaly detectors using feature attribution. This may be a substantial milestone in the pursuit of more transparent and accountable AD systems.

Acknowledgments

MK and SF acknowledge support by the BMFTR award 01IS24071A, by the DFG through FOR 5359 (ID 459419731), TRR 375 (ID 511263698), SPP 2298 (Id 441826958), and SPP 2331 (441958259, 553345933, 466468799), and by the Carl-Zeiss Foundation through the initiatives AI-Care and Process Engineering 4.0. The work of PW is partially supported by the Alexander von Humboldt Foundation. SV acknowledges support by the BMFTR award 01IW23005.

References

Abid, A.; Yuksekgonul, M.; and Zou, J. 2022. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *International Conference on Machine Learning*, 66–88. PMLR.

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31.

Ahamed, S.; Xu, Y.; and Rahmim, A. 2024. IgCONDA-PET: Implicitly-guided counterfactual diffusion for detecting anomalies in PET images. *arXiv preprint arXiv:2405.00239*.

Angiulli, F.; Fassetti, F.; Nisticó, S.; and Palopoli, L. 2023. Counterfactuals explanations for outliers via subspaces density contrastive loss. In *International Conference on Discovery Science*, 159–173. Springer.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9592–9600.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3): 1–58.

Cheng, H.; Xu, D.; Yuan, S.; and Wu, X. 2022. Fine-grained Anomaly detection in sequential data via counterfactual explanations. *arXiv preprint arXiv:2210.04145*.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2023. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*.

Datta, D.; Chen, F.; and Ramakrishnan, N. 2022. Framing algorithmic recourse for anomaly detection. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 283–293.

Defard, T.; Setkov, A.; Loesch, A.; and Audigier, R. 2021. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, 475–489. Springer.

Dehaene, D.; Frigo, O.; Combrexelle, S.; and Eline, P. 2020. Iterative energy-based projection on a normal data manifold for anomaly localization. In *International Conference on Learning Representations*.

Deng, L. 2012. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Dhurandhar, A.; Chen, P.-Y.; Luss, R.; Tu, C.-C.; Ting, P.; Shanmugam, K.; and Das, P. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*.

Ding, C.; and Pang, G. 2024. Improving Open-World Classification with Disentangled Foreground and Background Features. In *ACM Multimedia 2024*.

Fong, R. C.; and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision*, 3429–3437.

Fontanella, A.; Mair, G.; Wardlaw, J.; Trucco, E.; and Storkey, A. 2024. Diffusion models for counterfactual generation and anomaly detection in brain images. *IEEE Transactions on Medical Imaging*.

Ghandeharioun, A.; Kim, B.; Li, C.-L.; Jou, B.; Eoff, B.; and Picard, R. W. 2021. DISSECT: disentangled simultaneous explanations via concept traversals. In *International Conference on Learning Representations*.

Golan, I.; and El-Yaniv, R. 2018. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, 9758–9769.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Goyal, Y.; Wu, Z.; Ernst, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*, 2376–2384. PMLR.

Guidotti, R. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 1–55.

Han, X.; Zhang, L.; Wu, Y.; and Yuan, S. 2023. Achieving counterfactual fairness for anomaly detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 55–66. Springer.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. G. 2019. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*.

Hendrycks, D.; Mazeika, M.; Kadavath, S.; and Song, D. 2019. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, 15637–15648.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*.

Houben, S.; Stallkamp, J.; Salmen, J.; Schlipsing, M.; and Igel, C. 2013. Detection of traffic aigns in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*.

Ji, X.; Xue, A.; Wong, E.; Sokolsky, O.; and Lee, I. 2024. AR-Pro: counterfactual explanations for anomaly repair with formal properties. In *Annual Conference on Neural Information Processing Systems*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. CutPaste: Self-supervised learning for anomaly detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9664–9674.

Liznerski, P.; Ruff, L.; Vandermeulen, R. A.; Franks, B. J.; Kloft, M.; and Müller, K.-R. 2021. Explainable deep one-class classification. In *International Conference on Learning Representations*.

Liznerski, P.; Ruff, L.; Vandermeulen, R. A.; Franks, B. J.; Müller, K.-R.; and Kloft, M. 2022. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. *Transactions on Machine Learning Research*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*.

Miyato, T.; and Koyama, M. 2018. cGANs with projection discriminator. In *International Conference on Learning Representations*.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Conference on Fairness, Accountability, and Transparency*, 607–617.

Neuwirth, R. J. 2022. *The EU artificial intelligence act: regulating subliminal AI systems*. Taylor & Francis.

Pearl, J. 2009. *Causality*. Cambridge university press.

Reiss, T.; Cohen, N.; Bergman, L.; and Hoshen, Y. 2021. PANDA: Adapting pretrained features for anomaly detection and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2806–2814.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

Ruff, L.; Kauffmann, J. R.; Vandermeulen, R. A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, T. G.; and Müller, K.-R. 2021. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5): 756–795.

Ruff, L.; Vandermeulen, R. A.; Görnitz, N.; Binder, A.; Müller, E.; Müller, K.-R.; and Kloft, M. 2020. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*.

Ruff, L.; Vandermeulen, R. A.; Görnitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *International Conference on Machine Learning*, volume 80, 4390–4399.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252.

Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C. J.; and Müller, K.-R. 2020. Toward interpretable machine learning: Transparent deep neural networks and beyond. *arXiv preprint arXiv:2003.07631*.

Sanchez, P.; Kascenas, A.; Liu, X.; O'Neil, A. Q.; and Tsaftaris, S. A. 2022. What is healthy? Generative counterfactual diffusion for lesion localization. In *MICCAI Workshop on Deep Generative Models*, 34–44. Springer.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 618–626.

Siddiqui, A. A.; Tirunagari, S.; Zia, T.; and Windridge, D. 2024. VALD-MD: Visual attribution via latent diffusion for medical diagnostics. *arXiv preprint arXiv:2401.01414*.

Singla, S.; Eslami, M.; Pollack, B.; Wallace, S.; and Batmanghelich, K. 2023. Explaining the black-box smoothly—a counterfactual approach. *Medical Image Analysis*, 84: 102721.

Sulem, D.; Donini, M.; Zafar, M. B.; Aubet, F.-X.; Gasthaus, J.; Januschowski, T.; Das, S.; Kenthapadi, K.; and Archambeau, C. 2022. Diverse counterfactual explanations for anomaly detection in time series. *arXiv preprint arXiv:2203.11103*.

Tack, J.; Mo, S.; Jeong, J.; and Shin, J. 2020. CSI: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, volume 33, 11839–11852.

Varshneya, S.; Ledent, A.; Liznerski, P.; Balinskyy, A.; Mehta, P.; Mustafa, W.; and Kloft, M. 2024. Interpretable Tensor Fusion. *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Wolleb, J.; Bieder, F.; Sandkühler, R.; and Cattin, P. C. 2022. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, 35–45. Springer.

Wu, Z.; and Palmer, M. 1994. Verb semantics and lexical selection. In *Annual Meeting of the Association for Computational Linguistics*, 133–138.

Zhang, J.; Bargal, S. A.; Lin, Z.; Brandt, J.; Shen, X.; and Sclaroff, S. 2018a. Top-down neural attention by excitation backprop. In *International Journal of Computer Vision*, volume 126, 1084–1102. Springer.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018b. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE conference on computer vision and pattern recognition*, 586–595.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2223–2232.