

Explainable Deep One-Class Classification

Philipp Liznerski^{*1}, Lukas Ruff^{*2}, Robert A. Vandermeulen^{*2},
 Billy Joe Franks¹, Marius Kloft¹, and Klaus-Robert Müller³

¹Department of Computer Science, TU Kaiserslautern, Germany

²Machine Learning Group, TU Berlin, Germany

³Google Research Brain Team, ML Group TU Berlin, MPII, and Korea University

Abstract

Deep one-class classification variants for anomaly detection learn a mapping that concentrates nominal samples in feature space causing anomalies to be mapped away. Because this transformation is highly non-linear, finding interpretations poses a significant challenge. In this paper we present an explainable deep one-class classification method, *Fully Convolutional Data Description* (FCDD), where the mapped samples are themselves also an explanation heatmap. FCDD yields competitive detection performance and provides reasonable explanations on common anomaly detection benchmarks with CIFAR-10 and ImageNet. On MVTec-AD, a recent manufacturing dataset offering ground-truth anomaly maps, FCDD meets the state of the art in an unsupervised setting, and outperforms its competitors in a semi-supervised setting. Finally, using FCDD’s explanations we demonstrate the vulnerability of deep one-class classification models to spurious image features such as image watermarks.

1 Introduction

Anomaly detection (AD) is the task of identifying anomalies in a corpus of data [7, 10, 1]. Powerful new anomaly detectors based on deep learning have made AD more effective and scalable to large, complex datasets such as high-resolution images [36, 3]. While there exists much recent work on deep AD, there exists limited work on making such techniques explainable. Explanations are needed in industrial applications to match safety and security requirements [42, 4, 21], avoid unfair social bias [14], and support human experts in decision making [42, 20]. One typically makes anomaly detection explainable by annotating pixels with an anomaly score and, in some applications, such as finding tumors for breast cancer detection [35], these annotations are the primary goal of the detector.

One approach to deep AD, known as *Deep Support Vector Data Description* (DSVDD) [36], is based on finding a neural network that transforms data such that nominal data is concentrated to a predetermined center and anomalous data lies elsewhere. In this paper we present *Fully Convolutional Data Description* (FCDD), a modification of DSVDD so that the transformed samples are themselves an image corresponding to a downsampled anomaly heatmap. The pixels in this heatmap that are far from the predetermined center correspond to anomalous regions in the input image. FCDD does this by only using convolutional and pooling layers, thereby limiting the receptive field of each output pixel. Our method is based on the one-class classification paradigm [32, 49, 36] which is able to naturally incorporate known anomalies, but it is also effective when simply using synthetic anomalies.

We show that FCDD’s anomaly detection performance is close to the state of the art on the standard AD benchmarks CIFAR-10 and ImageNet while providing transparent explanations. On MVTec-AD, an AD dataset containing ground-truth anomaly heatmaps (cf., Figure 1),

^{*}equal contribution

we demonstrate the correctness of FCDD’s explanations, where we set a new state of the art. In further experiments we find that deep one-class classification models (e.g. DSVDD) are prone to the “Clever Hans” effect [26] where a detector fixates on spurious features such as image watermarks. In general, we find that the generated anomaly heatmaps are less noisy and provide more structure than all baselines, including gradient-based methods [45, 47] and autoencoders [41, 3].

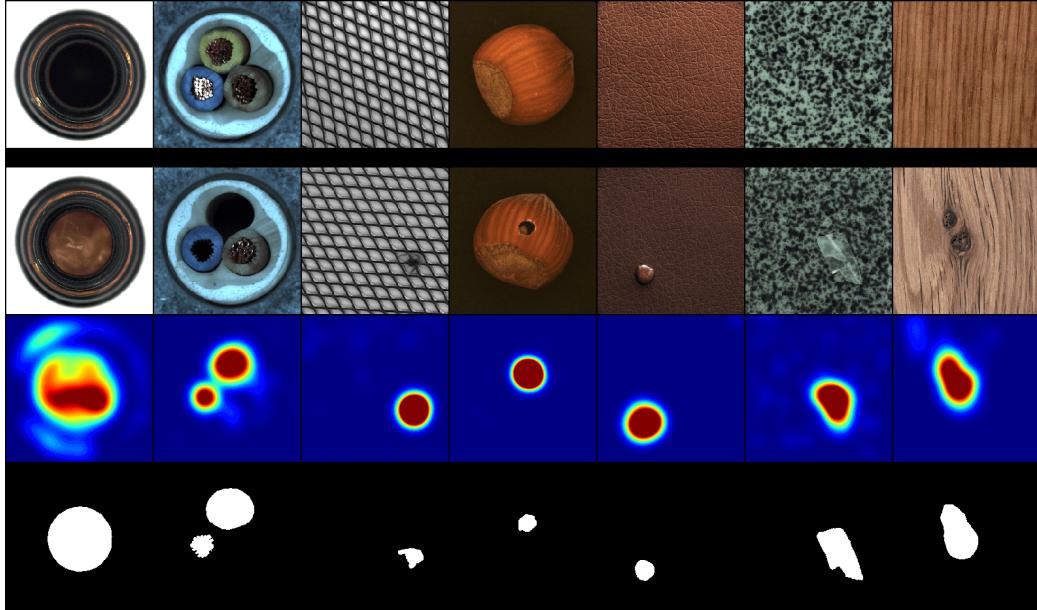


Figure 1: FCDD explanation heatmaps for MVTec-AD [3]. Rows from top to bottom show: (1) nominal samples (2) anomalous samples (3) FCDD anomaly heatmaps (4) ground-truth heatmaps.

2 Related Work

Here we outline related works on deep AD focusing on explanation approaches. Classically deep AD used autoencoders [15, 54, 53, 41]. Trained on a nominal dataset, autoencoders are assumed to not reconstruct anomalous samples well. Thus, the reconstruction error can be used as an anomaly score and the pixel-wise difference as an explanation [3]. Recent works have incorporated attention into reconstruction models that can be used as explanations [51, 28]. In the domain of videos, Sabokrou et al. [40] use a pretrained fully convolutional architecture in combination with a sparse autoencoder to extract 2D features and provide bounding boxes for anomaly localization. One drawback of these methods is that they offer no natural way to incorporate known anomalies during training.

More recently, one-class classification methods for deep AD have been proposed. These methods attempt to separate nominal samples from anomalies in an unsupervised manner by concentrating nominal data in feature space while mapping anomalies to distant locations [36, 6]. In the domain of NLP, DSVDD [36] has been successfully applied to text, which yields a form of interpretation using attention mechanisms [37]. For images, Kauffmann et al. [22] have used a deep Taylor decomposition [31] to create relevance scores. So far this approach has only been applied to the kernel OC-SVM [44] and has not yet been extended to deep methods.

Some of the best performing deep AD methods are based on self-supervision. These methods transform nominal samples, train a network to predict which transformation was used on the input, and provide an anomaly score through the confidence of the prediction [17, 12].

Hendrycks et al. [16] have extended this to incorporate known anomalies as well. No explanation approaches have been considered for these methods so far. Finally we mention that fully convolutional networks have been used to perform semantic segmentation [29, 34]. This approach requires ground-truth segmentation examples for training however, and is thus not applicable to the problem of deep AD.

3 Explaining Deep One-Class Classification

Before introducing our method we first review deep one-class classification. We then introduce fully convolutional architectures and describe their properties. Finally we present our method.

Deep One-Class Classification Deep one-class classification [36, 39] performs anomaly detection by learning a neural network to map nominal samples near a center \mathbf{c} in output space, causing anomalies to be mapped away. For our method we use a *Hypersphere Classifier* (HSC) [38], a recently proposed modification of Deep SAD [39], a semi-supervised version of DSVDD [36]. Let X_1, \dots, X_n denote a collection of samples and y_1, \dots, y_n be labels where $y_i = 1$ denotes an anomaly and $y_i = 0$ denotes a nominal sample. Then the HSC objective is

$$\min_{\mathcal{W}} \quad \frac{1}{n} \sum_{i=1}^n (1 - y_i) h(\phi(X_i; \mathcal{W}) - \mathbf{c}) - y_i \log(1 - \exp(-h(\phi(X_i; \mathcal{W}) - \mathbf{c}))), \quad (1)$$

where $\mathbf{c} \in \mathbb{R}^d$ is a predetermined center, and $\phi : \mathbb{R}^{c \times h \times w} \rightarrow \mathbb{R}^d$ a neural network with weights \mathcal{W} . Here h is the pseudo-Huber loss [19], $h(\mathbf{a}) = \sqrt{\|\mathbf{a}\|_2^2 + 1} - 1$, which is a robust loss that interpolates from quadratic to linear penalization. The HSC loss encourages ϕ to map nominal samples near \mathbf{c} and anomalous samples away from the center \mathbf{c} . Because we use a bias term in last layer of our networks, the location of \mathbf{c} is unimportant so we can simply set it to zero, removing it from the objective.

Fully Convolutional Architecture Our method uses a *fully convolutional network* (FCN) [29, 34] that maps an image to a matrix of features, i.e. $\phi : \mathbb{R}^{c \times h \times w} \rightarrow \mathbb{R}^{1 \times u \times v}$ by using alternating convolutional and pooling layers only, and does not contain any fully connected layers. In this context pooling can be seen as a special kind of convolution with fixed parameters.

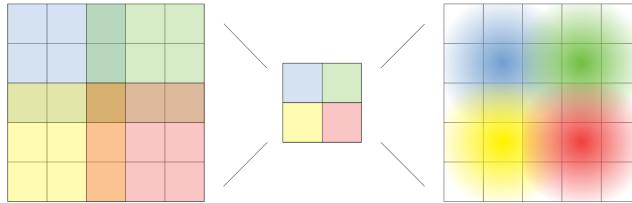


Figure 2: Visualization of a 3×3 convolution followed by a 3×3 transposed convolution with a Gaussian kernel (right), where both use a stride of 2.

A core property of a convolutional layer is that each pixel of its output only depends on a small region of its input, known as the output pixel’s *receptive field*. Since the output of a convolution is produced by moving a filter over the input image, each output pixel has the same relative position as its associated receptive field in the input. For instance, the lower-left corner of the output representation has a corresponding receptive field in the lower-left corner of the input image, etc. (cf., Figure 2 left side). The outcome of several stacked convolutions also has receptive fields of limited size and consistent relative position, though their size grows with the amount of layers. Because of this an FCN is able to preserve spatial information.

Fully Convolutional Data Description Here we introduce our novel explainable AD method *Fully Convolutional Data Description* (FCDD). By taking advantage of FCNs along with the HSC above, we propose a deep one-class method where the output features preserve spatial information and also serve as a downsampled anomaly heatmap. For situations where one would like to have a full-resolution heatmap, we include a methodology for upsampling the low-resolution heatmap based on properties of receptive fields.

FCDD is trained using samples which are labeled as nominal or anomalous. As before, let X_1, \dots, X_n denote a collection of samples with labels y_1, \dots, y_n where $y_i = 1$ denotes an anomaly and $y_i = 0$ denotes a nominal sample. Anomalous samples can simply be a collection of random images which are not from the nominal collection, e.g. one of the many large collections of images which are freely available like 80 Million Tiny Images [50] or ImageNet [9]. The use of such an auxiliary corpus has been recommended in recent works on deep AD, where it is termed *Outlier Exposure* (OE) [16, 17]. When one has access to “true” examples of the anomalous dataset, i.e. something that is likely to be representative of what will be seen at test time, we find that even using a few examples as the corpus of labeled anomalies (~ 5) performs exceptionally well. Furthermore, in the absence of *any* sort of known anomalies, one can generate synthetic anomalies, which we find is also very effective. With an FCN $\phi : \mathbb{R}^{c \times h \times w} \rightarrow \mathbb{R}^{u \times v}$ the FCDD objective utilizes a pseudo-Huber loss on the FCN output matrix $A(X) = (\sqrt{\phi(X; \mathcal{W})^2 + 1} - 1)$, where all operations are applied element-wise. The FCDD objective is then defined as (cf., (1)):

$$\min_{\mathcal{W}} \quad \frac{1}{n} \sum_{i=1}^n (1 - y_i) \frac{1}{u \cdot v} \|A(X_i)\|_1 - y_i \log \left(1 - \exp \left(-\frac{1}{u \cdot v} \|A(X_i)\|_1 \right) \right). \quad (2)$$

Here $\|A(X)\|_1$ is simply the sum of all entries in $A(X)$, which are all positive. FCDD is the utilization of an FCN in conjunction with the novel adaptation of the HSC loss we propose in (2). The objective maximizes $\|A(X)\|_1$ for anomalies and minimizes it for nominal samples, thus we use $\|A(X)\|_1$ as the anomaly score. Entries of $A(X)$ that contribute to $\|A(X)\|_1$ correspond to regions of the input image that add to the anomaly score. Note that $A(X)$ has spatial dimensions $u \times v$ and is smaller than the original image dimensions $h \times w$. One could use $A(X)$ directly as a low-resolution heatmap of the image, however it is often desirable to have full-resolution heatmaps. Because FCDD is not trained with ground-truth heatmaps, it is not possible to train an FCN like that in [34] to upsample the low-resolution heatmap $A(X)$: the network has no examples of how to properly upscale the anomaly heatmap. For this reason we introduce an upsampling scheme based on properties of receptive fields.

Heatmap Upsampling Since we do not have access to ground-truth pixel annotations during training, we cannot learn how to upsample using a deconvolutional type of structure. We derive a principled way to upsample our lower resolution anomaly heatmap. For every output pixel in $A(X)$ there is a unique input pixel which lies at the center of its receptive field. It has been observed before that the effect of the receptive field for an output pixel decays in a Gaussian manner as one moves away from the center of the receptive field [30]. We use this fact to upsample $A(X)$ by using a strided transposed convolution with a fixed Gaussian kernel (cf., Figure 2). We describe this operation and procedure in Algorithm 1 which simply corresponds to a strided transposed convolution. The kernel size is set to the receptive field range of FCDD and the stride to the cumulative stride of FCDD. The variance of the distribution can be picked empirically. Figure 3 shows a complete overview of our FCDD method and the process of generating full-resolution anomaly heatmaps.

4 Experiments

In this section, we experimentally evaluate the performance of FCDD both quantitatively and qualitatively. For a quantitative evaluation, we use the Area Under the ROC Curve (AUC)

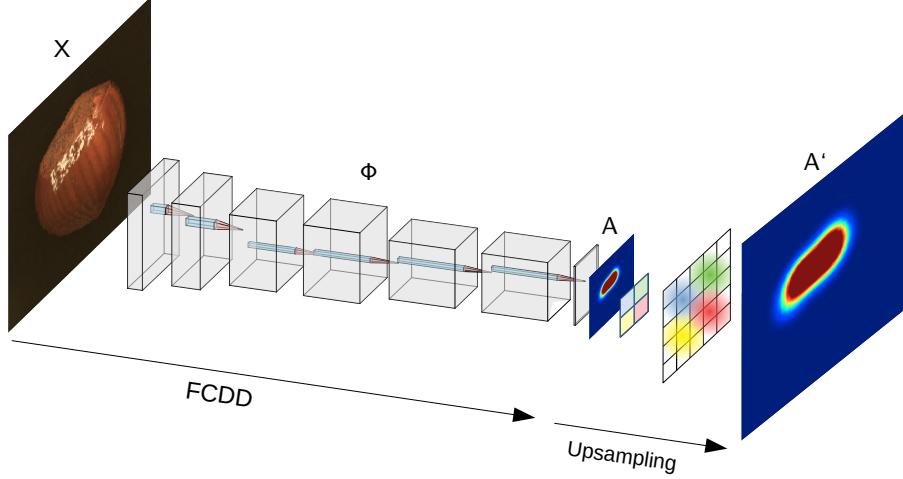


Figure 3: Visualization of the overall procedure to produce full-resolution anomaly heatmaps with FCDD. X denotes the input, ϕ the network, A the produced anomaly heatmap and A' the upsampled version of A using a transposed Gaussian convolution.

Algorithm 1 Receptive Field Upsampling

Input: $A \in \mathbb{R}^{u \times v}$ (low-res anomaly heatmap)
Output: $A' \in \mathbb{R}^{h \times w}$ (full-res anomaly heatmap)
Define: $[G_2(\mu, \sigma)]_{x,y} \triangleq \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x-\mu_1)^2 + (y-\mu_2)^2}{2\sigma^2}\right)$

```

 $A' \leftarrow 0$ 
for all output pixels  $a$  in  $A$  do
     $f \leftarrow$  receptive field of  $a$ 
     $c \leftarrow$  center of field  $f$ 
     $A' \leftarrow A' + a \cdot G_2(c, \sigma)$ 
end for
return  $A'$ 

```

[46] which is the commonly used measure in AD. For a qualitative evaluation, we compare the heatmaps produced by FCDD to existing deep AD explanation approaches. As baselines, we consider gradient-based methods [45] applied to hypersphere classifier (HSC) [38] models with unrestricted network architectures (i.e., networks that also have fully connected layers) as well as autoencoders [3] which enable one to use the pixel-wise reconstruction error as an explanation heatmap directly. We slightly blur the heatmaps of the baselines with the same Gaussian kernel we use for FCDD, which we found results in less noisy, more interpretable heatmaps. We include heatmaps without blurring in Appendix E. We adjust the contrast of the heatmaps per method to help highlight interesting features; see Appendix A for details.

4.1 Standard Anomaly Detection Benchmarks

We first evaluate FCDD on the standard Fashion-MNIST, CIFAR-10, and ImageNet datasets. The common AD benchmark is to utilize these classification datasets in a one-vs-rest setup where the “one” class is used as the nominal class and the rest of the classes are used as anomalies at test time. For training, we only use nominal samples as well as random samples from some auxiliary Outlier Exposure (OE) [16] dataset which is separate from the ground-truth anomaly classes following [16, 17]. We report the mean AUC over all classes for each dataset.

Fashion-MNIST We consider each of the ten Fashion-MNIST [52] classes in a one-vs-rest setup. We train Fashion-MNIST using EMNIST [8] or grayscaled CIFAR-100 [24] as OE. We found that the latter slightly outperforms the former (~ 3 AUC percent points). On Fashion-MNIST, we use a network that consists of three convolutional layers with batch normalization, separated by two downsampling pooling layers.

CIFAR-10 We consider each of the ten CIFAR-10 [24] classes in a one-vs-rest setup. As OE we use CIFAR-100, which does not share any classes with CIFAR-10. We use a model similar to LeNet-5 [27], but decrease the kernel size to three, add batch normalization, and replace the fully connected layers and last max-pool layer with one final 1×1 convolution.

ImageNet We consider 30 classes from ImageNet1k [9] for the one-vs-rest setup following Hendrycks et al. [16]. For OE we use ImageNet22k with ImageNet1k classes removed [16]. We use an architecture for inputs resized to 224×224 , consisting of six convolutional layers with batch normalization and three max-pool layers. The convolutional layers increase in the number of filters to a maximum of 128 filters, while the final layer reduces the amount of filters back to one.

State-of-the-art Methods We report results from state-of-the-art deep anomaly detection methods. Methods that do not incorporate known anomalies are the autoencoder (AE) baseline, DSVDD [36], Geometric Transformation based AD (GEO) [12], and a variant of GEO by Hendrycks et al. [17] (GEO+). Methods that use OE are a Focal loss classifier [17], also GEO+, Deep SAD [39], and HSC.

Table 1: Mean AUC (over all classes and 5 seeds per class) for Fashion-MNIST, CIFAR-10, and ImageNet. Results from existing literature are marked with an asterisk [2, 12, 17, 38].

Dataset	without OE					with OE				
	AE	DSVDD*	GEO*	Geo+*		Focal*	Geo+*	Deep SAD*	HSC*	FCDD
Fashion-MNIST	0.82	0.93	0.94	\times	\times	\times	\times	\times	\times	0.88
CIFAR-10	0.63	0.65	0.86	0.90	0.87	0.96	0.95	0.96	0.92	
ImageNet	0.51	\times	\times	\times	0.56	0.86	0.97	0.97	0.91	

Quantitative Results The mean AUC detection performance on the three AD benchmarks are reported in Table 1. We can see that FCDD, despite using a restricted FCN architecture to improve explainability, achieves a performance that is close to state-of-the-art deep AD methods. We provide detailed results for all individual classes in Appendix D.

Qualitative Results Figures 4 and 5 show the heatmaps for Fashion-MNIST and ImageNet respectively. For a Fashion-MNIST model trained on the nominal class “trousers”, the heatmaps show that FCDD correctly highlights horizontal elements as being anomalous, which makes sense since trousers are vertically aligned. For an ImageNet model trained on the nominal class “acorns”, we observe that colors seem to be most relevant for the model decision. The acorn images mostly have green and brown colors, thus the model discerns other colors to be anomalous. See for example the red leaf in the most anomalous nominal image (last column in (a) and the columns to the right in (b)). We provide further heatmaps for additional classes from all datasets in Appendix E.

Exploitation of Background Information? Figure 6 shows anomaly heatmaps for models trained on the nominal class “ships” of CIFAR-10 when the size of the OE dataset is increased from only 2 OE samples to full OE dataset size. Interestingly, we observe that with

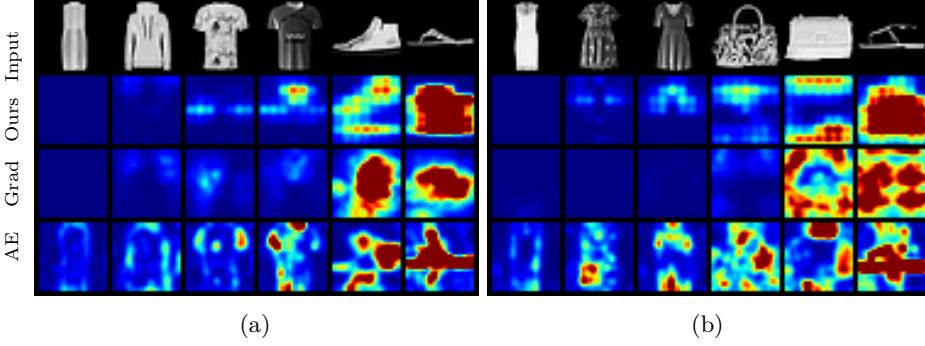


Figure 4: Anomaly heatmaps for anomalous test samples of a Fashion-MNIST model trained on nominal class “trousers”. In (a) EMNIST was used for OE and in (b) CIFAR-100. Columns ordered by increasing anomaly score from left to right.

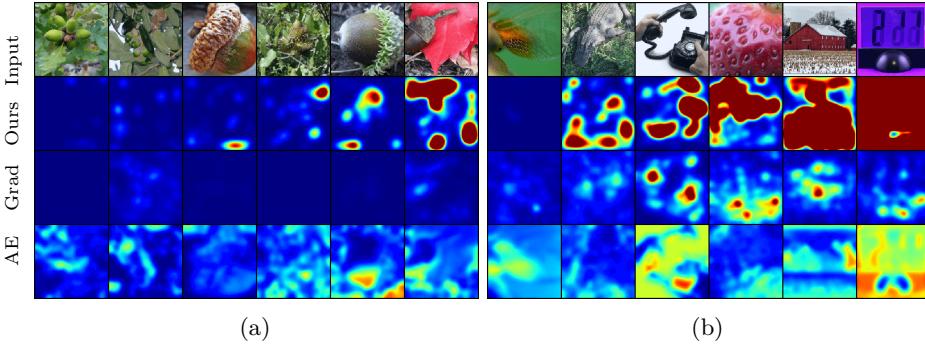


Figure 5: Anomaly heatmaps of an ImageNet model trained on nominal class “acorns”. (a) are nominal and (b) anomalous samples. Columns ordered by increasing anomaly score from left to right.

a larger, more diverse OE dataset the one-class model seems to focus more on the background. This can be misleading, as detection then is based on context (e.g. water) rather than the actual object features (e.g. ship). To further investigate this background hypothesis for CIFAR-10, we conduct another experiment where we simply block the foreground information with centered black boxes during training. Such a model still achieves a mean AUC of 0.86 (compared to 0.92 AUC previously). We also invert this procedure and blacken the border of each image, leaving the center untouched. This achieves a mean AUC of 0.85. We thus conclude the FCDD one-class model considers background at least as important as the object itself for AD on CIFAR-10. We provide examples of this manipulated training procedure in Figure 9 of the Appendix.

Baseline Explanations We found the gradient-based heatmaps to mostly produce centered blobs which lack spatial context (cf., Figure 6) and thus are not useful for explaining. The AE heatmaps, being directly tied to the reconstruction error anomaly score, look reasonable. We again note, however, that it is not straightforward how to include auxiliary OE samples or labeled anomalies into an AE approach, which leaves them with a poorer detection performance (cf., Table 1). Overall we find that the proposed FCDD anomaly heatmaps yield a good and consistent visual interpretation.

4.2 Explaining Defects in Manufacturing

Here we compare the performance of FCDD on the MVTec-AD dataset of defects in manufacturing [3]. This datasets offers annotated ground-truth segmentation maps

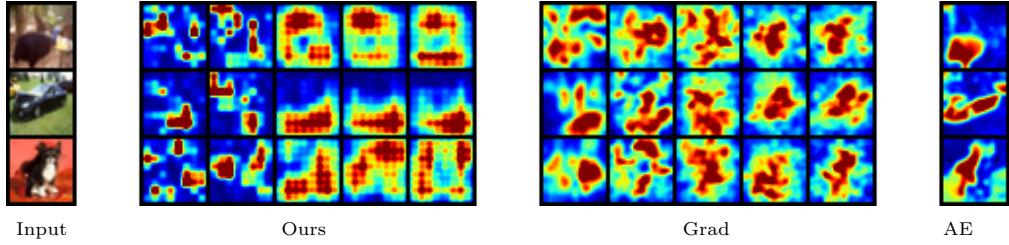


Figure 6: Anomaly heatmaps for three anomalous test samples on a CIFAR-10 model trained on nominal class “ships.” The second, third, and fourth blocks show the heatmaps of FCDD, gradient-based heatmaps of HSC, and AE heatmaps respectively. For Ours and Grad, we grow the number of OE samples from 2, 8, 128, 2048 to full OE. AE is not able to incorporate OE.

for testing, thus allowing a quantitative evaluation of model explanations. MVTec-AD contains 15 object classes of high-resolution RGB images with up to 1024×1024 pixels, where anomalous test samples are further categorized in up to 8 defect types, depending on the class. We follow [3] and compute a per sample AUC from the heatmap pixel scores, taking the given (binary) anomaly segmentation heatmaps as ground-truth pixel labels. We then report the mean over all samples of this “explanation” AUC for a quantitative evaluation. For FCDD, we use the same network architecture as for ImageNet above.

Synthetic Anomalies OE with a natural image dataset like ImageNet is not informative for MVTec-AD since anomalies here are rather subtle defects of the nominal class, rather than being out of class (cf., Figure 1). For this reason, we generate synthetic anomalies using a sort of “confetti noise,” a simple noise model that inserts colored blobs into images and reflects the local nature of anomalies. See Figure 7 for an example.

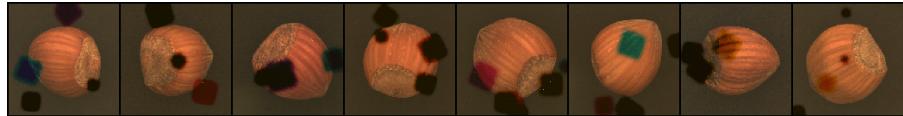


Figure 7: Synthetic anomalies on “hazelnut” using “confetti noise.”

Semi-Supervised FCDD A major advantage of FCDD in comparison to methods based on reconstruction is that it can be readily used in a semi-supervised AD setting [39]. To see the effect of having even only a few labeled anomalies available for training, we pick for each MVTec-AD class just *one* true anomalous sample per defect type at random and add it to the training set. This results in as little as 3–8 anomalous training samples. To also take advantage of the ground-truth heatmaps, we here train a model on a pixel level. Let X_1, \dots, X_n again denote a batch of inputs with corresponding ground-truth heatmaps Y_1, \dots, Y_n , each having $m = h \cdot w$ number of pixels. Let $A(X)$ also again denote the corresponding output anomaly heatmap of X . Then, we can formulate a pixel-wise objective by the following:

$$\min_{\mathcal{W}} \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{m} \sum_{j=1}^m (1 - (Y_i)_j) A'(X_i)_j \right) - \log \left(1 - \exp \left(-\frac{1}{m} \sum_{j=1}^m (Y_i)_j A'(X_i)_j \right) \right). \quad (3)$$

Results Figure 1 in the introduction shows heatmaps of FCDD trained on MVTec-AD. The results of the quantitative explanation are shown in Table 2. We can see that FCDD performs competitively in an unsupervised setting and sets a new state of the art of 0.94

pixel-wise mean AUC in a weak semi-supervised setting where we only use one anomalous sample per defect class. Another advantage of FCDD seems to be that its performance is more reliable, as FCDD shows the lowest standard deviation in mean AUC over classes.

Table 2: Pixel-wise mean AUC scores for all classes of the MVTec-AD dataset [3]. The competitors are Self-Similarity and L2 Autoencoder [3], AnoGAN [43, 3], CNN Feature Dictionaries [33, 3], Inverse Transform Autoencoder [18], Visually Explained Variational Autoencoder [28], and Convolutional Adversarial Variational Autoencoder with Guided Attention [51].

	unsupervised								semi-supervised
	AE-SSIM*	AE-L2*	AnoGAN*	CNNFD*	ITAE*	VEVAE*	CAVGA*	FCDD	FCDD
Bottle	0.93	0.86	0.86	0.78	×	0.87	×	0.80	0.87
Cable	0.82	0.86	0.78	0.79	×	0.9	×	0.80	0.94
Capsule	0.94	0.88	0.84	0.84	×	0.74	×	0.88	0.94
Carpet	0.87	0.59	0.54	0.72	×	0.78	×	0.93	0.99
Grid	0.94	0.9	0.58	0.59	×	0.73	×	0.87	0.94
Hazelnut	0.97	0.95	0.87	0.72	×	0.98	×	0.96	0.97
Leather	0.78	0.75	0.64	0.87	×	0.95	×	0.98	0.99
Metal Nut	0.89	0.86	0.76	0.82	×	0.94	×	0.88	0.98
Pill	0.91	0.85	0.87	0.68	×	0.83	×	0.86	0.95
Screw	0.96	0.96	0.8	0.87	×	0.97	×	0.87	0.93
Tile	0.59	0.51	0.5	0.93	×	0.80	×	0.92	0.97
Toothbrush	0.92	0.93	0.90	0.77	×	0.94	×	0.90	0.89
Transistor	0.90	0.86	0.80	0.66	×	0.93	×	0.80	0.82
Wood	0.73	0.73	0.62	0.91	×	0.77	×	0.89	0.95
Zipper	0.88	0.77	0.78	0.76	×	0.78	×	0.81	0.97
Mean	0.86	0.82	0.74	0.78	0.84	0.86	0.89	0.88	0.94
σ	0.10	0.13	0.13	0.10	×	0.09	×	0.06	0.05

4.3 The Clever Hans Effect

Lapuschkin et al. [25, 26] revealed that roughly one fifth of all horse images in PASCAL VOC [11] contain a watermark in the lower left corner. They showed that a (Fisher vector) classifier recognizes this as the relevant class pattern and fails if the watermark is removed. They call this the “clever hans” effect in memory of the horse Hans, who could correctly answer math problems by reading its master¹. We adapt this experiment for one-class classification and train FCDD with swapped labels, so we can interpret the heatmaps as nominal “horse” class explanations. We use ImageNet as OE.

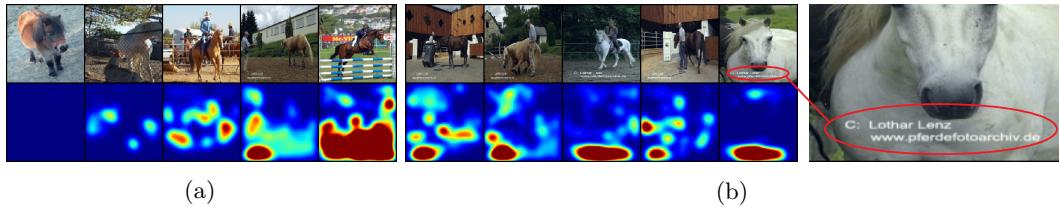


Figure 8: Heatmaps for horses on PASCAL VOC. (a) shows nominal samples ordered from most anomalous to most nominal from left to right. (b) shows examples that indicate that the model is a “clever hans”, i.e. has learned a characterization based on spurious features (watermarks).

Figure 8 (b) shows that a one-class model is indeed also vulnerable to learning a characterization based on spurious features: the watermarks in the lower left corner which have

¹ https://en.wikipedia.org/wiki/Clever_Hans

high scores whereas other regions have low scores. We also observe that the model yields high scores for bars, grids, and fences in Figure 8 (a). This is due to many images in the dataset containing horses jumping over bars or being in fenced areas. In both cases, the horse features themselves do not attain the highest scores. This underscores that explanations are crucial in one-class classification and anomaly detection to make model decisions transparent.

5 Conclusion

We have introduced FCDD, a novel explainable approach to deep AD. We have shown that FCDD yields competitive anomaly scores on common AD benchmark datasets and achieves a new state of the art for anomaly localization on MVTec-AD. Using FCDD’s explanation we have demonstrated that deep one-class classification models are also prone to exploit spurious data features such as image watermarks.

Acknowledgements

MK, PL, and BJF acknowledge support by the German Research Foundation (DFG) award KL 2698/2-1 and by the German Federal Ministry of Science and Education (BMBF) awards 01IS18051A, and 031B0770E. LR acknowledges support by the German Federal Ministry of Education and Research (BMBF) in the project ALICE III (01IS18049B). RV acknowledges support by the Berlin Institute for the Foundations of Learning and Data (BIFOLD) sponsored by the German Federal Ministry of Education and Research (BMBF). KRM was supported in part by the Institute for Information & Communications Technology Promotion and funded by the Korea government (MSIT) (No. 2017-0-01779), and was partly supported by the German Federal Ministry of Education and Research (BMBF) under Grants 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A and 01IS18037A; the German Research Foundation (DFG) under Grant Math+, EXC 2046/1, Project ID 390685689.

References

- [1] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, 3rd edition, 1994.
- [2] L. Bergman and Y. Hoshen. Classification-based anomaly detection for general data. In *ICLR*, 2020.
- [3] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, pages 9592–9600, 2019.
- [4] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause. Safe model-based reinforcement learning with stability guarantees. In *NIPS*, pages 908–918, 2017.
- [5] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010*, pages 177–186. Springer, 2010.
- [6] R. Chalapathy, A. K. Menon, and S. Chawla. Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*, 2018.
- [7] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.
- [8] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik. EMNIST: Extending MNIST to handwritten letters. In *IJCNN*, pages 2921–2926, 2017.
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.
- [10] F. Y. Edgeworth. On discordant observations. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 23(5):364–375, 1887.

- [11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
- [12] I. Golan and R. El-Yaniv. Deep anomaly detection using geometric transformations. In *NeurIPS*, pages 9758–9769, 2018.
- [13] I. Golan and R. El-Yaniv. Deep anomaly detection using geometric transformations. In *NeurIPS*, pages 9758–9769, 2018.
- [14] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social GAN: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, pages 2255–2264, 2018.
- [15] S. Hawkins, H. He, G. Williams, and R. Baxter. Outlier Detection Using Replicator Neural Networks. In *DaWaK*, volume 2454, pages 170–180, 2002.
- [16] D. Hendrycks, M. Mazeika, and T. G. Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.
- [17] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, pages 15637–15648, 2019.
- [18] C. Huang, J. Cao, F. Ye, M. Li, Y. Zhang, and C. Lu. Inverse-transform autoencoder for anomaly detection. *arXiv preprint arXiv:1911.10676*, 2019.
- [19] P. J. Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.
- [20] M. H. Jarrahi. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, 61(4):577–586, 2018.
- [21] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pages 97–117. Springer, 2017.
- [22] J. Kauffmann, K.-R. Müller, and G. Montavon. Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models. *Pattern Recognition*, 101:107198, 2020.
- [23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *ICLR*, 2015.
- [24] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. *Technical report*, 2009.
- [25] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Muller, and W. Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *CVPR*, pages 2912–2920, 2016.
- [26] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.
- [27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [28] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps. Towards visually explaining variational autoencoders. *arXiv preprint arXiv:1911.07389*, 2019.
- [29] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.
- [30] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NIPS*, pages 4898–4906, 2016.
- [31] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

- [32] M. M. Moya, M. W. Koch, and L. D. Hostetler. One-class classifier networks for target recognition applications. In *World Congress on Neural Networks*, pages 797–801, 1993.
- [33] P. Napoletano, F. Piccoli, and R. Schettini. Anomaly detection in nanofibrous materials by CNN-based self-similarity. *Sensors*, 18(1):209, 2018.
- [34] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, pages 1520–1528, 2015.
- [35] G. Quellec, M. Lamard, M. Cozic, G. Coatrieux, and G. Cazuguel. Multiple-instance learning for anomaly detection in digital mammography. *IEEE Transactions on Medical Imaging*, 35(7):1604–1614, 2016.
- [36] L. Ruff, R. A. Vandermeulen, N. Görnitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft. Deep one-class classification. In *ICML*, volume 80, pages 4390–4399, 2018.
- [37] L. Ruff, Y. Zemlyanskiy, R. Vandermeulen, T. Schnake, and M. Kloft. Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *ACL*, pages 4061–4071, 2019.
- [38] L. Ruff, R. A. Vandermeulen, B. J. Franks, K.-R. Müller, and M. Kloft. Rethinking assumptions in deep anomaly detection. *arXiv preprint arXiv:2006.00339*, 2020.
- [39] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft. Deep semi-supervised anomaly detection. In *ICLR*, 2020.
- [40] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, 172:88–97, 2018.
- [41] M. Sakurada and T. Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pages 4–11, 2014.
- [42] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. Toward interpretable machine learning: Transparent deep neural networks and beyond. *arXiv preprint arXiv:2003.07631*, 2020.
- [43] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.
- [44] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the Support of a High-Dimensional Distribution. *Neural computation*, 13(7):1443–1471, 2001.
- [45] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.
- [46] K. A. Spackman. Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 160–163. Elsevier, 1989.
- [47] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328, 2017.
- [48] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.
- [49] D. M. J. Tax. *One-class classification*. PhD thesis, Delft University of Technology, 2001.

- [50] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.
- [51] S. Venkataramanan, K.-C. Peng, R. V. Singh, and A. Mahalanobis. Attention guided anomaly detection and localization in images. *arXiv preprint arXiv:1911.08616*, 2019.
- [52] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- [53] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1933–1941, 2017.
- [54] C. Zhou and R. C. Paffenroth. Anomaly detection with robust deep autoencoders. In *KDD*, pages 665–674, 2017.

A Anomaly Heatmap Visualization

For anomaly heatmap visualization, the FCDD anomaly scores $A'(X)$ need to be rescaled to values in $[0, 1]$. Instead of applying standard min-max scaling that would divide by $\max A'(X)$, we use anomaly score quantiles to adjust the contrast in the heatmaps. For a collection of inputs $\mathcal{X} = \{X_1, \dots, X_n\}$ with corresponding full-resolution anomaly heatmaps $\mathcal{A} = \{A'(X_1), \dots, A'(X_n)\}$, the normalized heatmap $I(X)$ for some $A'(X)$ is computed as

$$I(X)_j = \min \left\{ \frac{A'(X)_j - \min(\mathcal{A})}{q_\eta(\{A' - \min(\mathcal{A}) \mid A' \in \mathcal{A}\})}, 1 \right\},$$

where j denotes the j -th pixel and q_η the η -th percentile over all pixels and examples in \mathcal{A} . The subtraction and min operation are applied on a pixel level, i.e. the minimum is extracted over all pixels and all samples of \mathcal{A} and subtraction is then applied elementwise. Using the η -th percentile might leave some of the values above 1, which is why we finally clamp the pixels at 1.

The specific choice of η and set of samples \mathcal{X} differs per figure. We select them empirically depending on the dataset and the property that should be highlighted. In general, the lower η the more red (anomalous) regions we have in the heatmaps because more values are left above one and vice versa. The choice of \mathcal{X} ranges from just one sample X , s.t. $A'(X)$ is normalized only w.r.t. to its own scores to provide a heatmap of relative anomalies, to the complete dataset showing absolute anomalies. The choice of η and \mathcal{X} is consistent per figure. In the following we list the choices made for the individual figures in the main paper.

MVTec-AD Figure 1 uses $\eta = 0.97$ and sets \mathcal{X} to the set of all samples used in the figure.

Fashion-MNIST Figure 4 uses $\eta = 0.93$ and sets \mathcal{X} to the set of all samples used per subfigure.

CIFAR-10 Figure 6 uses $\eta = 0.93$ and sets \mathcal{X} to X for each heatmap $I(X)$ to show relative anomalies. So each image is normalized with respect to itself only.

ImageNet Figure 5 uses $\eta = 0.97$ and sets \mathcal{X} to the complete train set consisting of all nominal samples and equally many randomly chosen auxiliary anomalies.

Pascal VOC Figure 8 uses $\eta = 0.97$ and sets \mathcal{X} to the set of all samples used per subfigure.

Heatmap Upsampling For the Gaussian kernel heatmap upsampling described in Algorithm 1, we set the default parameters as

$$\sigma = \begin{cases} \log(0.45k + 1) + 0.25 & k < 20 \\ 10 & \text{otherwise} \end{cases},$$

where k is the receptive field extent.

B Details on the Network Architectures

Here we provide the complete network architectures we used on the different datasets. We use leaky ReLU activations between the convolutional layers, and place them after the batch normalization layer, if batch normalization is applied.

Fashion-MNIST

Layer (type:depth-idx)	Output Shape	Param #
<hr/>		
Conv2d: 1-1	[-1, 8, 28, 28]	200
BatchNorm2d: 1-2	[-1, 8, 28, 28]	--
MaxPool2d: 1-3	[-1, 8, 14, 14]	--
Conv2d: 1-4	[-1, 16, 14, 14]	3,200
MaxPool2d: 1-5	[-1, 16, 7, 7]	--
Conv2d: 1-6	[-1, 1, 7, 7]	16
<hr/>		
Total params:	3,416	
Trainable params:	3,416	
Non-trainable params:	0	
<hr/>		
Input size (MB):	0.00	
Forward/backward pass size (MB):	0.07	
Params size (MB):	0.01	
Estimated Total Size (MB):	0.09	
<hr/>		

CIFAR-10

Layer (type:depth-idx)	Output Shape	Param #
<hr/>		
Conv2d: 1-1	[-1, 32, 32, 32]	864
BatchNorm2d: 1-2	[-1, 32, 32, 32]	--
MaxPool2d: 1-3	[-1, 32, 16, 16]	--
Conv2d: 1-4	[-1, 64, 16, 16]	18,432
BatchNorm2d: 1-5	[-1, 64, 16, 16]	--
MaxPool2d: 1-6	[-1, 64, 8, 8]	--
Conv2d: 1-7	[-1, 128, 8, 8]	73,728
Conv2d: 1-8	[-1, 1, 8, 8]	128
<hr/>		
Total params:	93,152	
Trainable params:	93,152	
Non-trainable params:	0	
<hr/>		
Input size (MB):	0.01	
Forward/backward pass size (MB):	0.44	
Params size (MB):	0.36	
Estimated Total Size (MB):	0.81	
<hr/>		

ImageNet, MVTec-AD, and Pascal VOC

Layer (type:depth-idx)	Output Shape	Param #
<hr/>		
Conv2d: 1-1	[-1, 8, 224, 224]	600
BatchNorm2d: 1-2	[-1, 8, 224, 224]	--
MaxPool2d: 1-3	[-1, 8, 112, 112]	--
Conv2d: 1-4	[-1, 32, 112, 112]	6,400
BatchNorm2d: 1-5	[-1, 32, 112, 112]	--
MaxPool2d: 1-6	[-1, 32, 56, 56]	--
Conv2d: 1-7	[-1, 64, 56, 56]	18,432
BatchNorm2d: 1-8	[-1, 64, 56, 56]	--
Conv2d: 1-9	[-1, 128, 56, 56]	73,728
BatchNorm2d: 1-10	[-1, 128, 56, 56]	--
MaxPool2d: 1-11	[-1, 128, 28, 28]	--
Conv2d: 1-12	[-1, 128, 28, 28]	147,456
Conv2d: 1-13	[-1, 1, 28, 28]	128
<hr/>		
Total params:	246,744	
Trainable params:	246,744	
Non-trainable params:	0	

```
Input size (MB): 0.57
Forward/backward pass size (MB): 11.49
Params size (MB): 0.94
Estimated Total Size (MB): 13.01
```

C Training and Optimization

Here we provide training and optimization details for the experiments from Section 4. We apply common pre-processing (e.g. data normalization) and data augmentation steps in our data loading pipeline. To sample auxiliary anomalies in an online manner during training, each nominal sample of a batch has a 50% chance of being replaced by a randomly picked auxiliary anomaly. This leads to balanced training batches for sufficiently large batch sizes. One epoch in our implementation still refers to the original nominal data training set size, so that approximately 50% of the nominal samples have been seen per training epoch. Below, we list further details for the specific datasets.

Fashion-MNIST We train for 400 epochs using a batch size of 128 samples. We optimize the network parameters using SGD [5] with Nesterov momentum ($\mu = 0.9$) [48], weight decay of 10^{-6} and an initial learning rate of 0.01, which decreases the previous learning rate per epoch by a factor of 0.98. The pre-processing pipeline is: (1) Random crop to size 28 with beforehand zero-padding of 2 pixels on all sides (2) random horizontal flipping with a chance of 50% (3) data normalization.

CIFAR-10 We train for 600 epochs using a batch size of 200 samples. We optimize the network using Adam [23] ($\beta = (0.9, 0.999)$) with weight decay 10^{-6} and an initial learning rate of 0.001 which is decreased by a factor of 10 at epoch 400 and 500. The pre-processing pipeline is: (1) Random color jitter with all parameters² set to 0.01 (2) random crop to size 32 with beforehand zero-padding of 4 pixels on all sides (3) random horizontal flipping with a chance of 50% (4) additive Gaussian noise with $\sigma = 0.001$ (5) data normalization.

ImageNet We use the same setup as in CIFAR-10, but resize all images to size 256×256 before forwarding them through the pipeline and change the random crop to size 224 with no padding. Test samples are center cropped to a size of 224 before being normalized.

Pascal VOC We use the same setup as in CIFAR-10, but resize all images to size 224×224 before forwarding them through the pipeline and remove the Random Crop step.

MVTec-AD For MVTec-AD we redefine an epoch to be ten times an iteration of the full dataset because this improves the computational performance of the data pipeline. We train for 200 epochs using SGD with Nesterov momentum ($\mu = 0.9$), weight decay 10^{-5} , and an initial learning rate of 0.001, which decreases per epoch by a factor of 0.985. The pre-processing pipeline is: (1) Resize to 240×240 pixels (2) random crop to size 224 with no padding (3) random color jitter with either all parameters set to 0.04 or 0.0005, randomly chosen (4) 50% chance to apply additive Gaussian noise (5) data normalization. For the upsampling we change the default Gaussian kernel standard deviation from 10 to 12 when using the confetti noise based auxiliary anomaly samples.

D Quantitative Detection Results for Individual Classes

Table 3 shows the class-wise results on Fashion-MNIST for AE, Deep Support Vector Data Description (DSVDD) [36, 2] and Geometric Transformation based AD (GEO) [13].

²<https://pytorch.org/docs/1.4.0/torchvision/transforms.html#torchvision.transforms.ColorJitter>

Table 3: AuROC scores for all classes of Fashion-MNIST [52].

	without OE			with OE FCDD
	AE	DSVDD*	Geo*	
T-Shirt/Top	0.85	0.98	0.99	0.82
Trouser	0.89	0.90	0.98	0.98
Pullover	0.78	0.91	0.91	0.83
Dress	0.87	0.94	0.90	0.90
Coat	0.88	0.89	0.92	0.86
Sandal	0.43	0.92	0.93	0.91
Shirt	0.70	0.83	0.83	0.75
Sneaker	0.95	0.99	0.99	0.99
Bag	0.86	0.92	0.91	0.86
Ankle Boot	0.96	0.99	0.99	0.93
Mean	0.82	0.93	0.94	0.88

In Table 4 the class-wise results for CIFAR-10 are reported. Competitors without OE are AE [36], DSVDD [36], GEO [13] and an adaptation of GEO (GEO+) [17]. Competitors with OE are the focal loss classifier [17], again GEO+ [17], Deep Semi-supervised Anomaly Detection (Deep SAD) [39, 38] and the hypersphere Classifier [38].

Table 4: AuROC scores for all classes of CIFAR-10 [24].

	without OE				with OE				
	AE*	DSVDD*	GEO*	Geo+*	Focal*	Geo+*	Deep SAD*	HSC*	FCDD
Airplane	0.59	0.62	0.75	0.78	0.88	0.90	0.94	0.97	0.91
Automobile	0.57	0.66	0.96	0.97	0.94	0.99	0.98	0.99	0.94
Bird	0.49	0.51	0.78	0.87	0.79	0.94	0.90	0.93	0.87
Cat	0.58	0.59	0.72	0.81	0.80	0.88	0.87	0.90	0.86
Deer	0.54	0.61	0.88	0.93	0.82	0.97	0.95	0.97	0.92
Dog	0.62	0.66	0.88	0.90	0.86	0.94	0.93	0.94	0.91
Frog	0.51	0.68	0.83	0.91	0.93	0.97	0.97	0.98	0.96
Horse	0.59	0.67	0.96	0.97	0.88	0.99	0.97	0.98	0.93
Ship	0.77	0.76	0.93	0.95	0.93	0.99	0.97	0.98	0.94
Truck	0.67	0.73	0.91	0.93	0.92	0.99	0.96	0.97	0.93
Mean	0.59	0.65	0.86	0.90	0.87	0.96	0.95	0.96	0.92

In Table 5 the class-wise results for Imagenet are shown, where competitors are the AE, the focal loss classifier [17], Geo+ [17], Deep SAD [38] and HSC [38]. Results from the literature are marked with an asterisk.

Table 5: AuROC scores for 30 classes of ImageNet [9].

	without OE AE	with OE					
		Focal*	Geo+*	Deep	SAD*	HSC*	FCDD
Acorn	0.53	×	×	0.99	0.99	0.96	
Airliner	0.80	×	×	0.97	1.00	0.98	
Ambulance	0.30	×	×	0.99	1.00	0.99	
American alligator	0.61	×	×	0.93	0.98	0.92	
Banjo	0.33	×	×	0.97	0.98	0.88	
Barn	0.49	×	×	0.99	1.00	0.95	
Bikini	0.49	×	×	0.97	0.99	0.89	
Digital clock	0.30	×	×	0.99	0.97	0.90	
Dragonfly	0.64	×	×	0.99	0.98	0.96	
Dumbbell	0.46	×	×	0.93	0.92	0.82	
Forklift	0.44	×	×	0.91	0.99	0.90	
Goblet	0.49	×	×	0.92	0.94	0.86	
Grand piano	0.24	×	×	1.00	0.97	0.93	
Hotdog	0.53	×	×	0.96	0.99	0.96	
Hourglass	0.42	×	×	0.96	0.97	0.91	
Manhole cover	0.85	×	×	0.99	1.00	0.99	
Mosque	0.59	×	×	0.99	0.99	0.95	
Nail	0.62	×	×	0.93	0.94	0.89	
Parking meter	0.48	×	×	0.99	0.93	0.78	
Pillow	0.60	×	×	0.99	0.94	0.91	
Revolver	0.55	×	×	0.98	0.98	0.89	
Rotary dial telephone	0.47	×	×	0.90	0.98	0.82	
Schooner	0.67	×	×	0.99	0.99	0.93	
Snowmobile	0.47	×	×	0.98	0.99	0.95	
Soccer ball	0.44	×	×	0.97	0.93	0.78	
Stingray	0.61	×	×	0.99	0.99	0.97	
Strawberry	0.35	×	×	0.98	0.99	0.96	
Tank	0.39	×	×	0.97	0.99	0.91	
Toaster	0.45	×	×	0.98	0.92	0.78	
Volcano	0.56	×	×	0.90	1.00	0.97	
Mean	0.51	0.56	0.86	0.97	0.97	0.91	

E Further Qualitative Anomaly Heatmap Results

This section reports further anomaly heatmaps, including examples for the background experiments, unblurred baseline heatmaps, and class-wise heatmaps for all datasets.

Background Exploitation Experiment Figure 9 shows some training images manipulated by blackening the center and background, as described in Section 4.1. The inputs are shown with corresponding FCDD heatmaps. We use $\eta = 0.93$ and set \mathcal{X} to the complete train set consisting of nominal samples and equally many randomly chosen auxiliary anomalies.

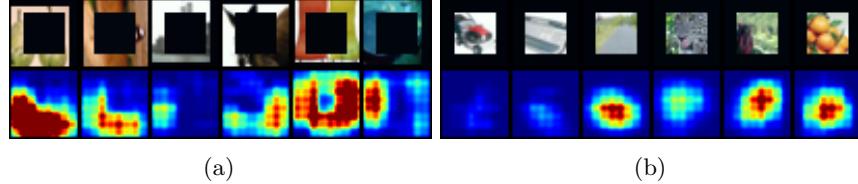


Figure 9: Background exploitation experiments with "cars" considered nominal, (a) shows black centered anomalous images and (b) shows black bordered anomalous images. The images show training samples, thus include augmentation, like random crop.

Unblurred Anomaly Heatmap Baselines Here we show some unblurred baseline heatmaps for the figures in Section 4.1. Figures 10, 11, and 12 show the unblurred heatmaps for Fashion-MNIST, ImageNet, and CIFAR-10 respectively.

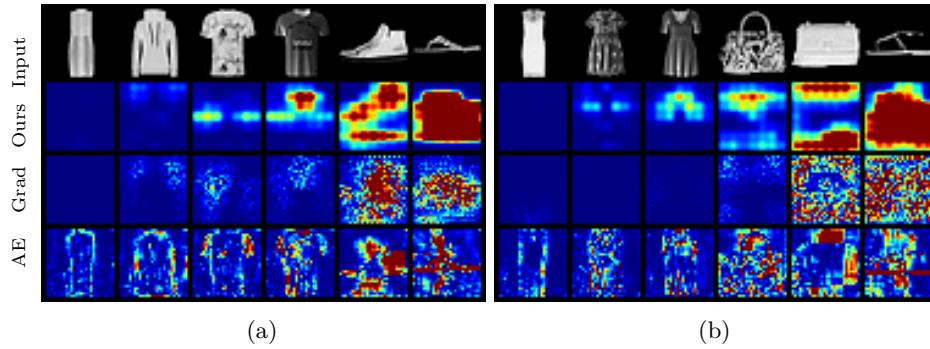


Figure 10: Anomaly heatmaps for anomalous test samples of a Fashion-MNIST model trained on nominal class “trousers”. In (a) EMNIST was used as OE and in (b) CIFAR-100.

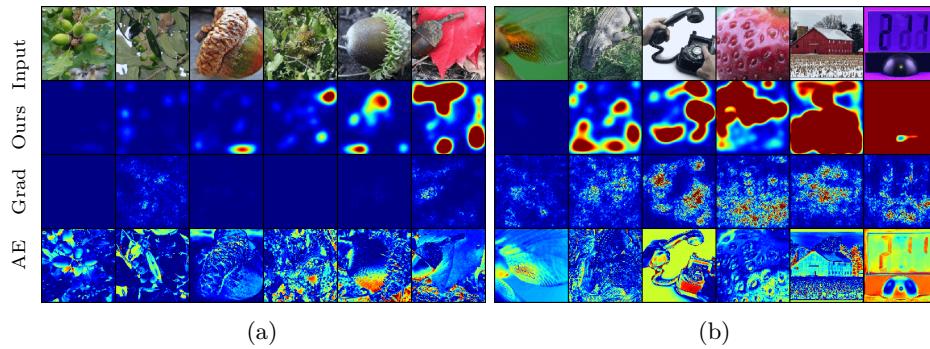


Figure 11: Anomaly heatmaps of an ImageNet model trained on nominal class “acorns”. (a) are nominal and (b) anomalous samples.

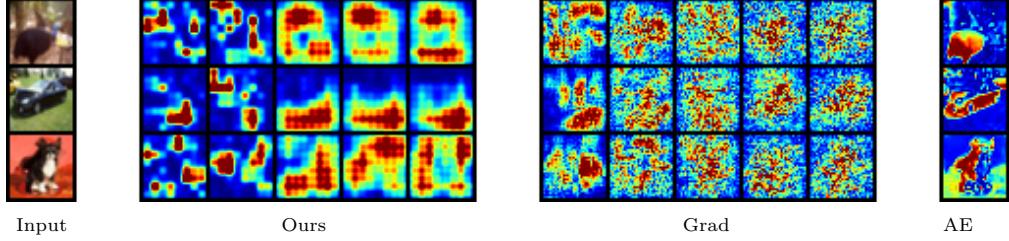


Figure 12: Anomaly heatmaps for three anomalous test samples (Input left) on a CIFAR-10 model trained on nominal class “ships”. The second, third, and fourth blocks show the heatmaps of FCDD (Ours), gradient-based heatmaps of HSC, and AE heatmaps respectively. For Ours and Grad, we grow the number of OE samples from 2, 8, 128, 2048 to full OE. AE is not able to incorporate OE.

Class-wise Anomaly Heatmaps For all datasets we show heatmaps in this section with adjusted contrast curves by setting \mathcal{X} to the set of all samples used per subfigure. Further, we set $\eta = 0.93$ for Fashion-MNIST and CIFAR-10, $\eta = 0.97$ for MVTec-AD and ImageNet. The rows in all heatmaps show the following: (1) Input samples (2) FCDD heatmaps (3) gradient heatmaps used on HSC (4) autoencoder reconstruction heatmaps. Heatmaps for MVTec-AD add a fifth row containing the ground-truth anomaly map. Heatmaps for Fashion-MNIST using auxiliary anomalies from EMNIST are in Figure 13, using CIFAR-100 for OE instead are in Figure 14. CIFAR-10 heatmaps are in Figure 15, and heatmaps for all classes of ImageNet are in Figures 17 and 18. Finally, we present MVTec-AD heatmaps in Figure 16.

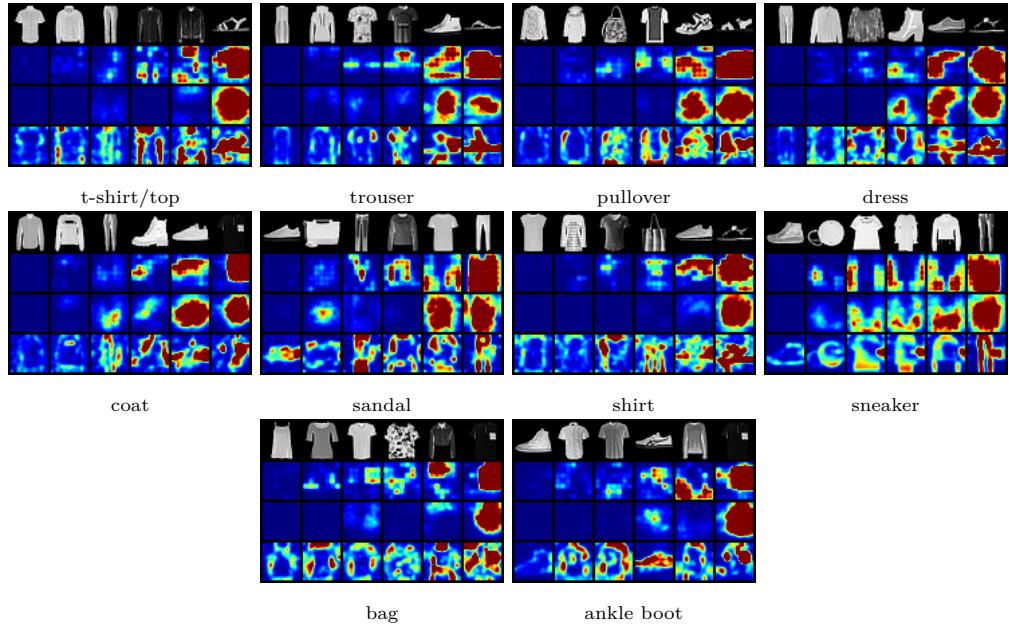


Figure 13: Anomaly heatmaps for anomalous test samples in Fashion-MNIST using EMNIST OE. Columns ordered by increasing anomaly score from left to right. The subcaptions refer to the nominal class that each model is trained on.

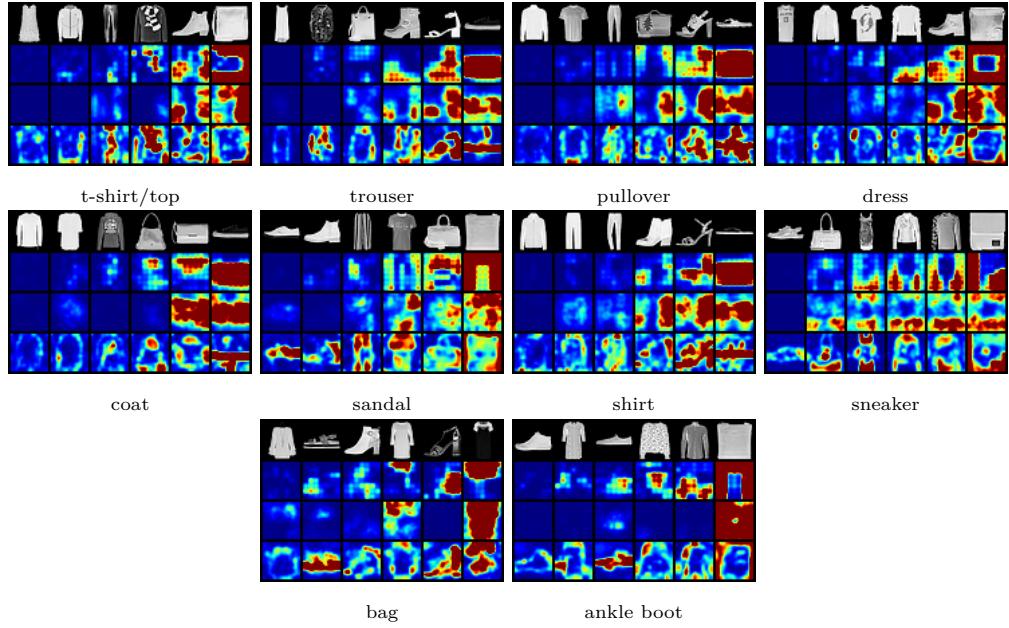


Figure 14: Anomaly heatmaps for anomalous test samples in Fashion-MNIST using CIFAR-100 OE. Columns ordered by increasing anomaly score from left to right. The subcaptions refer to the nominal class that each model is trained on.

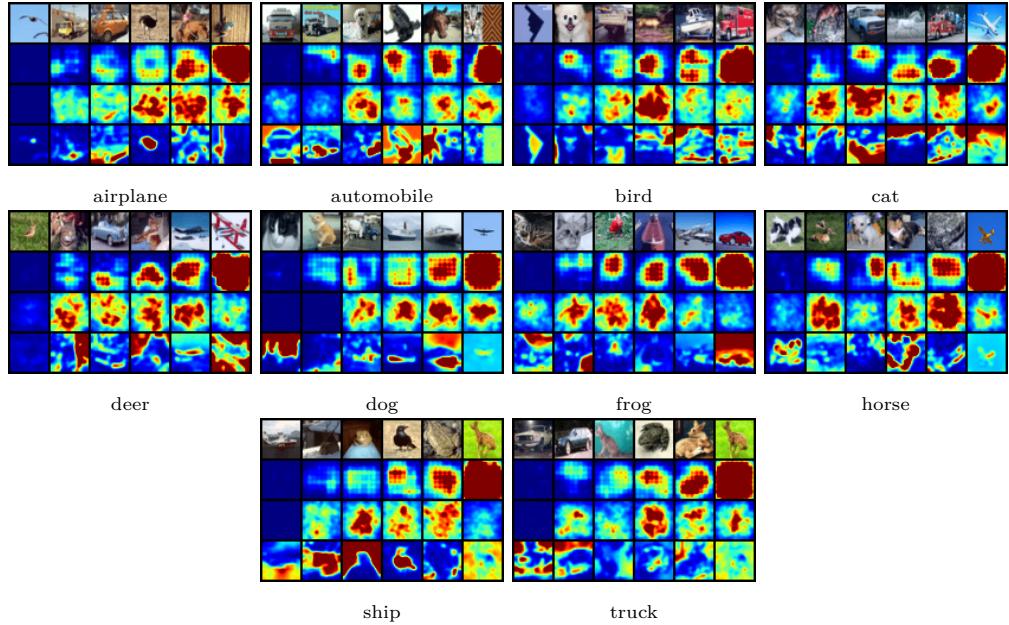


Figure 15: Anomaly heatmaps for anomalous test samples in CIFAR-10. Columns ordered by increasing anomaly score from left to right. The subcaptions refer to the nominal class that each model is trained on.

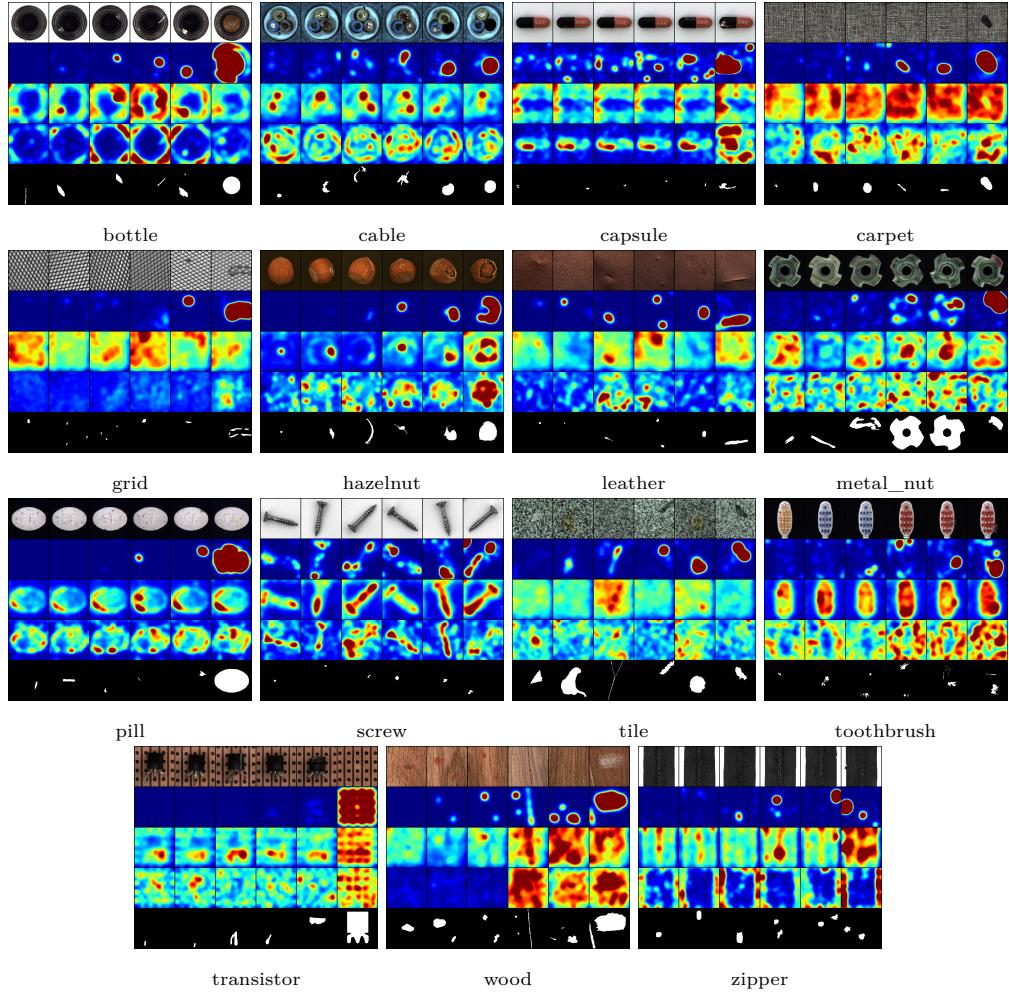


Figure 16: Anomaly heatmaps for anomalous test samples in MVTec-AD. Columns ordered by increasing anomaly score from left to right. The subcaptions refer to the nominal class that each model is trained on.

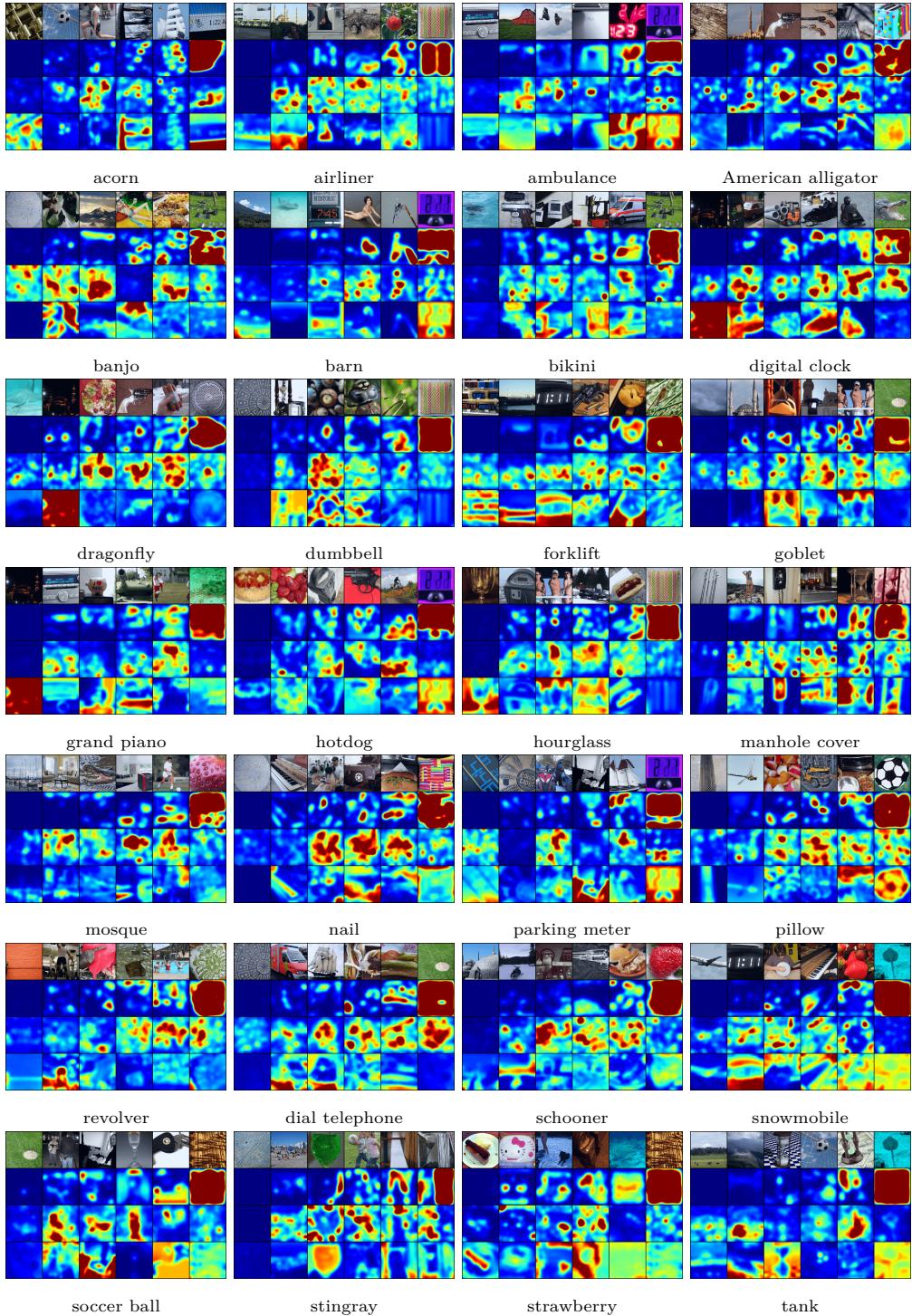


Figure 17: Anomaly heatmaps for anomalous test samples in ImageNet, where classes 1-28 are shown. Columns ordered by increasing anomaly score from left to right. The subcaptions refer to the nominal class that each model is trained on.

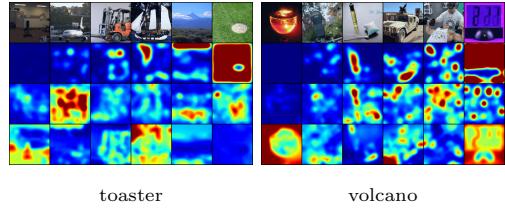


Figure 18: Anomaly heatmaps for anomalous test samples in ImageNet, where classes 29-30 are shown. Columns ordered by increasing anomaly score from left to right. The subcaptions refer to the nominal class that each model is trained on.