Transfer-Based Semantic Anomaly Detection

Lucas Deecke [1] **Lukas Ruff** [2] **Robert A. Vandermeulen** [3] **Hakan Bilen** [1]

Abstract

Detecting semantic anomalies is challenging due to the countless ways in which they may appear in real-world data. While enhancing the robustness of networks may be sufficient for modeling simplistic anomalies, there is no good known way of preparing models for all potential and unseen anomalies that can potentially occur, such as the appearance of new object classes. In this paper, we show that a previously overlooked strategy for anomaly detection (AD) is to introduce an explicit inductive bias toward representations transferred over from some large and varied semantic task. We rigorously verify our hypothesis in controlled trials that utilize intervention, and show that it gives rise to surprisingly effective auxiliary objectives that outperform previous AD paradigms.

1. Introduction

The goal of anomaly detection (AD) is the identification of unusual samples within data (Edgeworth, 1887; Chandola et al., 2009; Pang et al., 2020a; Ruff et al., 2021). For data types that are semantically rich such as images, "unusualness" can be caused by a variety of high-level (or *semantic*) factors, for example the appearance of new objects classes, or unexpected shapes or poses. For these settings, there has been continued interest in developing new deep AD methods (Zhai et al., 2016; Schlegl et al., 2017; Sabokrou et al., 2018; Deecke et al., 2018; Ruff et al., 2018; Golan & El-Yaniv, 2018; Pidhorskyi et al., 2018; Hendrycks et al., 2019b;c; Goyal et al., 2020; Tack et al., 2020) that utilize end-to-end learning, a defining property amongst deep learning approaches (Krizhevsky et al., 2012; He et al., 2016; Goodfellow et al., 2016).

Because of the sheer number of factors that can potentially cause an anomaly, there is no feasible way of a priori describing or anticipating them. As a result, for deep AD there exists no established principal learning objective. Several auxiliary solutions have been proposed: one line of work utilizes self-supervision (Golan & El-Yaniv, 2018; Hendrycks et al., 2019c; Bergman & Hoshen, 2020; Tack et al., 2020; Sohn et al., 2021), for example learning representations from the task of predicting simple geometric transformations (rotations, translations, *etc.*) applied to non-anomalous examples. A second popular approach broadly resembles weak supervision, and uses large unstructured collections of data as auxiliary anomalies (Hendrycks et al., 2019b;c; Ruff et al., 2020a).

Considering the rather ad-hoc nature of many of these approaches, especially given the semantic richness present in natural images, one may wonder whether they learn particularly meaningful features from such auxiliary objectives. This is problematic since anomalies can manifest themselves in ways that require a good semantic understanding — for example when anomalies appear in crowded scenes (Mahadevan et al., 2010).

Here we propose a different perspective and hypothesize that, because there is simply no way of anticipating all potential semantic anomalies for unseen images in advance, the best bet is to follow a transfer-based approach that utilizes the semantically rich features obtained from some semantic task solved on a large, semantically varied dataset. We systematically evaluate different strategies to introduce such an inductive bias in AD, and identify simple strategies that yield surprisingly powerful AD methods.

Our work builds on the recently increased availability and utilization of networks pretrained on semantically rich tasks that incorporate different variations commonly seen in data (edges, color, semantic categories, *etc.*). While He et al. (2019) fundamentally questioned whether actual benefits are achieved from the use of pretrained models, Hendrycks et al. (2019a) painted them in a more positive light, showing they improve robustness and uncertainty calibration.

Rich semantic representations have been shown to boost the performance in many machine learning problems, including image classification (Donahue et al., 2014; Guo et al., 2019), object detection (Girshick et al., 2014; Girshick, 2015), the transfer between large numbers of tasks (Zamir et al., 2018), or from one domain to another (Rebuffi et al., 2017; 2018). In another example, a surge of papers has recently elevated

[1]University of Edinburgh [2]Aignostics (majority of work done while with TU Berlin) [3]ML Group, TU Berlin. Correspondence to: Lucas Deecke <l.deecke@ed.ac.uk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

the role of pretrained models in natural language processing (Mikolov et al., 2018; Devlin et al., 2019; Howard & Ruder, 2018; Adhikari et al., 2019; Hendrycks et al., 2020).

Our central hypothesis is that transferring features from semantic tasks such as ILSVRC image classification (Deng et al., 2009) provides very powerful and generic representations for various AD problems, even when the pretraining task is only loosely related to the task of AD. In doing so, it is important to ensure that the change in representation is not excessive, as this risks *catastrophic forgetting* (Kirkpatrick et al., 2017). For AD in particular, it is crucial to preserve variations incorporated during pretraining that, even though they potentially don't exist in the training data, can nonetheless be meaningful for inferring anomalous semantics at test time (Tax & Müller, 2003; Rippel et al., 2020). Opposed to mere feature extraction (Bergman et al., 2020), our experiments show that it is critical to let the network have *some* flexibility to learn new variations important for AD.

To the best of our knowledge, a rigorous analysis and evaluation of transfer-based approaches for AD is still lacking in the literature. Our experiments show that such strategies provide very powerful methods for AD that outperform previous approaches in the deep AD literature on a set of common benchmarks (Section 5). Moreover they are straightforward to train and deploy, and can be coupled with any modern network architecture.

Besides experiments on the predominant AD benchmarks, we propose the use of disentanglement datasets (Gondal et al., 2019) to evaluate the semantic detection performance of AD models. In doing so, we verify that our proposed method is able to *robustly* detect anomalies under interventions (*e.g*. a change of object color), showing it preserves meaningful semantic variations in its representations.

In addition to verifying the suitability of our method on semantic benchmarks (Section 5.2), models trained on semantic tasks have been shown to learn elements required for non-semantic decisions in early parts of the network (Zeiler & Fergus, 2014). Indeed we find that our methods are suitable for tasks considered non-semantic (Ahmed & Courville, 2020), such as the popular CIFAR-10 one-versus-rest benchmark (Section 5.3).

2. Related Work

Anomaly detection has a long history (Edgeworth, 1887) and has been extensively studied in the machine learning literature, *e.g*. through hidden Markov models for detecting network attacks (Ourston et al., 2003), active learning of anomalies (Pelleg & Moore, 2005), or dynamic Bayesian networks for traffic incident detection (Singliar & Hauskrecht, 2006). An overview over traditional AD methods can be found in Chandola et al. (2009) and Emmott et al. (2013).

Previous deep AD methods utilized autoencoders (Zhou & Paffenroth, 2017; Zong et al., 2018), hybrid methods (Erfani et al., 2016), one-class classification (Ruff et al., 2018; Sabokrou et al., 2018; Ghafoori & Leckie, 2020; Goyal et al., 2020), or GANs (Goodfellow et al., 2014; Schlegl et al., 2017; Akcay et al., 2018; Deecke et al., 2018; Perera et al., 2019; Ngo et al., 2019; Berg et al., 2019). Another line of work explores detecting anomalous videos (Sultani et al., 2018; Ionescu et al., 2019; Ngo et al., 2019; Pang et al., 2020b).

A recent focus has been on developing auxiliary tasks for AD, often following the paradigm of self-supervision, for example predicting geometric transformations of normal data (Golan & El-Yaniv, 2018; Hendrycks et al., 2019c). Different from this, Hendrycks et al. (2019b) propose carrying out AD in a weakly supervised manner in what they call outlier exposure (OE), where one utilizes large unstructured sets of data as auxiliary outliers to improve detection performance. While our approaches also leverage large corpora, they establish inductive biases as a separate crucial element for semantic AD.

Several recent publications have investigated unsupervised mechanisms to learn disentangled representations (Kulkarni et al., 2015; Higgins et al., 2017; Bouchacourt et al., 2018; Burgess et al., 2018; Chen et al., 2018; Kim & Mnih, 2018; Kumar et al., 2018; Locatello et al., 2019; 2020). We propose using image datasets developed for disentanglement (Gondal et al., 2019) for gaining better insights into AD methods, letting us show that an inductive bias seems necessary to detect anomalies on a semantic level.

3. Motivation

To motivate our approach we investigate the semantic viability of features learned under different auxiliary AD objectives. We employ a strategy inspired by linear probing (Zhang et al., 2017) in a two-stage setup very commonly used in AD applications (Erfani et al., 2016; Sohn et al., 2021): after a feature extraction phase f over some data $x \sim p(x)$, a subsequent one-class model g is learned to encapsulate the normal class w.r.t. the push-forward $f_*(p(x))$.

For the data we make use of a standard AD benchmark (Ruff et al., 2018): single classes from CIFAR-10 (*e.g*. dogs) constitute the normal class, and the one-class model g is learned over all embedded training examples of this class. At test time, we measure whether the two-stage model can successfully identify the appearance of the remaining object classes (cats, deers, *etc*.) as anomalous. The metric reported in Table 1 results from repeating this procedure for all ten classes, and recording the area under the ROC curve (AUC) relative to that of a random baseline (AUC of 0.5).

The initial extraction phase occurs at one of three layers

Layer	(i) Self-sup.	(ii) Weakly-sup.	(iii) Transfer-b.
$f_{\text{conv},1}$	1.44 *(0.19)*	1.58 *(0.20)*	2.02 *(0.21)*
$f_{\text{conv},2}$	4.60 *(0.17)*	3.83 *(0.16)*	5.48 *(0.19)*
$f_{\text{conv},3}$	4.63 *(0.15)*	5.12 *(0.12)*	6.72 *(0.15)*

Table 1. Percent improvement in AUC relative to a random baseline on CIFAR-10 AD for one-class SVMs on top of features extracted from LeNet layers (`conv1-3`) trained through different learning paradigms (transfer-based *etc*.). Standard deviations in parentheses were computed over five random seeds.

(`conv1-3`) of a LeNet architecture (LeCun et al., 1989) trained via three different paradigms (see next paragraph). The subsequent one-class stage always uses the exact same OC-SVM (Schölkopf et al., 1999). Fixing a simple model g on top of $f_{\text{conv},i}(x)$ allows direct insights into the viability of each layer's features for semantic AD.

We compare $f_{\text{conv},i}$ after training the extraction model with three different learning paradigms for AD:

(i) self-supervision through geometric transformation of the input data as in Hendrycks et al. (2019c);

(ii) weakly supervised classification via outlier exposure (CIFAR-100 as OE dataset) (Hendrycks et al., 2019b);

(iii) transferring from another task (CIFAR-100 classification) and subsequent finetuning through OE.

For (i), (ii), and (iii), deeper features always result in performance improvements (see Table 1). When extracting at deeper layers – which are typically associated with higher semantic function (Yosinski et al., 2014; Zeiler & Fergus, 2014; Mahendran & Vedaldi, 2016; Asano et al., 2020) – there is however a performance gap between paradigms: (i) self-supervised features do not improve from `conv2` to `conv3`, indicating they learn predominantly low-level features. For (ii) OE-based extraction performance increases a little at every layer, but overall AUCs are most improved by the (iii) transfer-based approach, which raises mean AUC by 6.72% in `conv3`. From this, we can already observe that transfer-based features can have a favorable impact for robust downstream AD detection performance.

In our experimental section (Section 5) we expand on this finding and show that, when transferring semantic representations to complex AD tasks, it is crucial to ensure models do not suffer from *catastrophic forgetting*. The next section introduces ways in which this can be achieved, *e.g*. through adequate regularization.

4. Methods

We review components of our proposed approach in Sections 4.1 and 4.2, and subsequently introduce two methods for semantic AD with an inductive bias: ADIB (Section 4.3) and ADRA (Section 4.4).

4.1. AD using an Unstructured Corpus

In AD, a (semantic) understanding of normal examples is extracted from a set $S_n = \{x_j\}_{j=1}^n$ assumed to have been sampled i.i.d. from the *normal* data distribution $\mathbb{P}^+$ over some space $\mathcal{X}$. The goal is to learn a one-class model $f_\theta\colon \mathcal{X} \to [0,1]$ with parameters $\theta \in \Theta$ that decides whether a previously unseen $x \in \mathcal{X}$ is *normal* (s.t. $f_\theta(x) \approx 0$) or *anomalous* ($f_\theta(x) \approx 1$).

The way S_n is used to learn f_θ defines how different approaches in the AD literature can be categorized, *e.g*. in an unsupervised way (Ruff et al., 2018), or through self-supervision (Golan & El-Yaniv, 2018) (*c.f*. Section 2). The concept of outlier exposure (OE) (Hendrycks et al., 2019b) utilizes a large number of unlabelled images from some unstructured corpus of data Q_m (where commonly $m \gg n$), for example 80 Million Tiny Images (Torralba et al., 2008), on which models are trained to identify whether samples belong to the corpus or the normal data $\mathbb{P}^+$. Importantly, this is a form of weak supervision via existing resources (Zhou, 2018), and not equivalent to supervised classification: images from the auxiliary corpus are not necessarily true anomalies (and may even contain samples from $\mathbb{P}^+$). For an N-sized batch of samples, the associated learning objective can be formulated as

$$\arg\min_\theta \Big\{ \mathcal{L}[f_\theta] = \mathcal{L}_{S_n}[f_\theta] + \mathcal{L}_{Q_m}[f_\theta] = \frac{1}{N}\Big[\sum_{x\in S_n} \log f_\theta(x) + \sum_{x \in Q_m} \log(1 - f_\theta(x))\Big]\Big\}. \quad (1)$$

Recent state-of-the-art AD methods have proposed to modify this objective by using radial functions Ruff et al. (2020a), which is in line with the so-called concentration assumption common in AD (Schölkopf & Smola, 2002; Steinwart et al., 2005). We include such radial functions in our ablation (Table 4), however empirically observed that – when paired with an explicit inductive bias – standard classifiers typically performed better.

4.2. Transfer-Based AD

Following work that investigated the prospects of large pretrained networks (Zamir et al., 2018; Adhikari et al., 2019; Hendrycks et al., 2020), a recent study proposed carrying out AD through a nearest neighbor search on top of features extracted from a large pretrained residual network (Bergman et al., 2020). However as can be seen from the experimental results in Table 5, simply transferring over fixed representations to an unrelated task seems subpar for semantic AD. Next, we outline how parameters obtained from some pretraining task can be more effectively transferred to AD.

Pretraining itself tends to follow a simple standard protocol: a model's parameters are randomly initialized with some distribution, for example Xavier initialization (Glorot & Bengio, 2010). Optimization of a suitable transfer task $\mathcal{T}$ (*e.g.* ImageNet object classification) yields a set of general-purpose parameters $\theta_0 \in \Theta$. The so-obtained model f_{θ_0} is then ready to be transferred to some downstream task $\mathcal{B}$.

The traditional methodology for leveraging pretrained models is to continue to optimize the model parameters (or a subset thereof) on $\mathcal{B}$. One crucial limitation of this learning protocol is that when learning on $\mathcal{B}$ isn't carried out carefully through the introduction of some explicit inductive bias (Li et al., 2018), this risks *catastrophic forgetting* of information previously extracted from $\mathcal{T}$. To alleviate this issue, a common approach is to use regularization, *e.g.* as in continual learning (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017). In the following two sections, we introduce two new AD-specific learning methods that prevent catastrophic forgetting.

4.3. Anomaly Detection with an Inductive Bias

Due to the complex ways in which semantic anomalies may manifest themselves in images, we hypothesize that our best bet for robust semantic AD is to introduce a semantic inductive bias into models.

To achieve this, we augment the learning criterion introduced in eq. (1) with an additional regularizer $\Omega\colon \Theta \to \mathbb{R}_+$ that constraints models, resulting in the following objective:

$$\arg\min_{\theta\in\Theta}\big\{\mathcal{L}_{S_n}[f_\theta] + \mathcal{L}_{Q_m}[f_\theta] + \Omega(\theta)\big\}. \tag{2}$$

As our ablations (Table 4) show, an inductive bias such as L_2 regularization toward initial pretrained parameters $\theta_0 \in \Theta$ is crucial for robust semantic AD performance. Motivated by this finding, in *Anomaly Detection with an Inductive Bias (ADIB)* we set $\Omega(\theta) = \alpha||\theta - \theta_0||^2$ scaled by $\alpha \in \mathbb{R}_+$.

We find that ADIB outperforms previous state-of-the-art AD methods on semantic anomaly benchmarks. For the CIFAR-10 semantic AD benchmark, for example, it raises the state of the art to 74.6 versus 41.6 mean AP reported previously by Ahmed & Courville (2020). Moreover, ADIB sets a new state of the art on the widespread one-versus-rest AD benchmark, raising the bar from 96.1 (Ruff et al., 2020a) to 99.1 mean AUC.

4.4. Anomaly Detection with Residual Adaptation

Regularization can also be formulated to bolster parameter efficiency. For this, we propose constraining the underlying generating function of residual networks (He et al., 2016) $\Phi(x) = x + f(x)$ to allow at most a linear change from the pretrained mapping Φ_0 with f_0 in every layer, whereby $\Phi(x) - \Phi_0(x) = Vx$. This is then rearranged to:

$$\Phi(x) = x + f_0(x) + Vx, \tag{3}$$

where V linearly corrects from adjacent layers (and can be implemented via 1x1 convolutions), and f_0 is the residual 3x3 convolution obtained from some transfer task $\mathcal{T}$. Assuming that pretrained models will have obtained strong general-purpose representations that should require only minimal changes to adapt to new tasks, only the V are learned, while f_0 is left unchanged. Similar strategies have been used in multi-task (Rebuffi et al., 2017; 2018; Deecke et al., 2020) and NLP (Stickland & Murray, 2019) settings to restrict the number of learnable parameters there.

We apply this strategy in *Anomaly Detection with Residual Adaptation (ADRA)*, and our experiments in Section 5 demonstrate that its performance is often comparable to that of regularizing all parameters via $\Omega(\theta)$. At the same time, as only V gets learned at each layer, ADRA is highly efficient. Such savings are crucial for applications in which multiple normal datasets exist but memory footprints are restrictive (*e.g.* federated learning scenarios (Yang et al., 2019; Bhagoji et al., 2019)).

5. Experiments

For evaluation, we first propose a novel experiment that is based on disentanglement datasets that have been introduced recently (Section 5.1). We then evaluate ADIB and ADRA on two benchmark settings: semantic AD (Section 5.2), and the widely adopted one-versus-rest AD (Section 5.3). All experiments have been implemented with PyTorch (Paszke et al., 2019).[1]

5.1. Examining Models through Interventions

As previous authors have emphasized, curating datasets with semantic anomalies is challenging (Ahmed & Courville, 2020). We here propose to achieve this via datasets originally developed for disentangled representations (Kulkarni et al., 2015; Higgins et al., 2017; Bouchacourt et al., 2018; Burgess et al., 2018; Chen et al., 2018; Kim & Mnih, 2018; Kumar et al., 2018; Locatello et al., 2019; 2020) that contain underlying ground-truth factors of images, in particular high-resolution, realistic datasets such as the recently released MPI3D (Gondal et al., 2019). In contrast to previous evaluations for semantic AD, for example those that modify CIFAR-10 to such a task (Ahmed & Courville, 2020), interventions on ground-truth factors allow for principled measurements of semantic capabilities of a given model, as for example the color of an object can be changed in a systematic fashion.

[1] Code available at `https://github.com/VICO-UoE/TransferAD`.

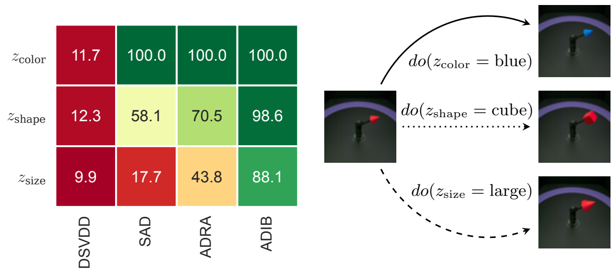

Figure 1. Detection performance in AUC under interventions on color, shape, and size of images in MPI3D.

MPI3D contains joint pairs of latent ground-truth factors z (color, shape, angle, *etc.*), and corresponding images x_z of a robot arm mounted with an object. The original dataset comes in three styles (photo-realistic, simple, or detailed animation); because the models we evaluate use rich deep architectures, we skip evaluation on simple and animated images (which are useful for simpler models) and focus on the photo-realistic images here.

All models use the same number of parameters, and differ only in which AD loss is optimized — DSVDD (Ruff et al., 2018) uses eq. (1) without any weak supervision (no $\mathcal{L}_{Q_m}[f_\theta]$ term). SAD (Ruff et al., 2020a;b) differs from DSVDD only in that it uses OE. Our models combine both OE and an inductive bias, see eq. (2). To ensure fair comparison, we use the exact same ResNet26 for all methods, and initialize all of them in exactly the same way, *i.e.* with the same pretrained weights. Note however our proposed ADRA has less modeling power than DSVDD and SAD, due to having fewer learnable parameters.

For semantic AD experiments on MPI3D, we propose fixing a red cone as the normal object (chosen arbitrarily), and train models on all available views. Anomalies are obtained by interventions on three underlying factors: (i) changing color to blue, (ii) transforming shape to cube, and (iii) increasing size. Two additional degrees of freedom exist in the dataset: background color and camera height. Interventions on these have an outsized impact on images however, and do not provide any real challenge to a residual network (or any other modern vision architecture, for that matter), which is why we do not consider them here.

For weak supervision through OE we use all remaining images that do not belong to neither the normal nor the anomaly class. For example white, green, brown, and olive all appear in the corpus Q_m.

Optimization The underlying model for DSVDD, SAD, ADRA, ADIB is the exact same ResNet26, optimized via stochastic gradient descent (momentum parameter of 0.9, weight decay of 10^{-4}) for a total of 100 epochs, with learning rate reductions by 1/10 after 60 and 80 epochs. The batch size is fixed to 128, and we only use standard augmentations. For all models, we initialize parameters via θ_0 obtained from pretraining on ImageNet and then train them further on the downstream AD task.

In ADRA only linear corrections V are learned, while θ_0 is fixed. For an explicit inductive bias in ADIB, we scale the regularization term $\Omega(\theta)$ with $\alpha = 10^{-2}$, as recommended by Li et al. (2018). Results are averaged over 5 seeds.

Results AUCs for different interventions are displayed in Fig. 1. Detecting even the most simple semantic anomaly, such as a change in object color z_{color} = red $\rightarrow$ blue is impossible when learning without any weak supervision, as is the case for DSVDD (11.7 AUC).

Our proposed intervention protocol confirms that it is beneficial to introduce a concept of differentness via OE. In other words, exposing models to the concept of red being normal, while also showing it examples of other colors (brown, green, *etc.*) *prepares* the model for *potential* anomalous shifts — although SAD has never seen a blue example, OE enables it to identify it as "*not* red", and hence an anomaly.

To obtain more robust models that can pick up on less obvious interventions such as changing the shape z_{shape} = cone $\rightarrow$ cube or z_{size} = small $\rightarrow$ large, adequate forms of regularization appear to be critical. While it has fewer learnable parameters, ADRA improves performance over SAD under all interventions. Some performance gap remains, however, which is likely a consequence of the parameter-efficiency of ADRA, letting it rely more on the weights of the base network which potentially aren't *particularly* well suited for the task.

ADIB has a higher degree of flexibility, thus allowing for sample-efficient utilization of those features which *are* useful from the pretrained network. While ADIB might be a simple strategy for the transfer of rich semantic features to AD, the performance under all three interventions shows that it can *robustly* detect semantic anomalies.

We finally note that weak supervision through OE consistently increased disentanglement in the learned representations. DCI disentanglement (Eastwood & Williams, 2018) almost doubles from 0.068 for DSVDD to 0.103 for SAD, their only distinction being the absence and presence of weak supervision via Q_m, respectively. Locatello et al. (2020) made a similar observation in the context of unsupervised learning, finding that *some* weak supervision is required for disentanglement.

Non-Semantic Shift Recent work examined model robustness toward non-semantic shift, such as the appearance of color not contained in the training data, which can confuse models from their primary objective of detecting se-

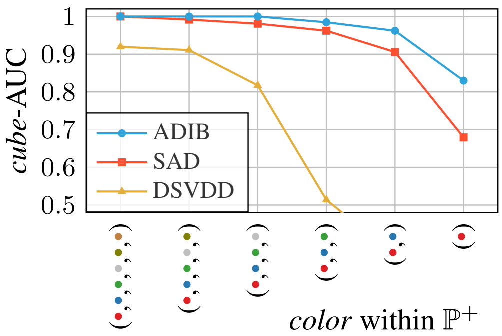

Figure 2. Robustness of detecting *cube* under non-semantic color shifts for DSVDD (no OE), SAD (uses OE), and our ADIB. The x-axis indicates colors that are included in the normal distribution.

mantic categories (Ahmed et al., 2021). In order to examine this setting on MPI3D we (w.l.o.g.) set *cube* = anomalous and *cone* = normal and consider color a non-semantic factor.

The experiment consists of a controlled sequence of trials: a single color (red) is included in the normal data at first, and the detection performance of models for *cone vs. cube* (of any color) is evaluated. Then another color is picked and added to the normal data (which now contains red & blue), and models are evaluated again. Repeating this for green, white, *etc*. yields $\mathbb{P}^+_{\text{red}}, \ldots, \mathbb{P}^+_{\text{all}}$ and gives precise control over the degree in which semantic context may be established.

Fig. 2 shows the extent to which our transfer-based approach improves robustness to non-semantic shifts and underlines the importance of preventing drift from the transfer task. SAD makes use of OE (which in this experiment includes shapes other than *cone* and *cube*, but never additional colors) and enhances performance relative to DSVDD (which does not use OE). A gap remains however when context is established only through OE (SAD *vs*. ADIB). Especially for few colors in $\mathbb{P}^+$ transfer-based AD appears very useful to manifesting the right semantic context.

5.2. Semantic AD

In this section, we evaluate recently proposed benchmarks for semantic AD (Ahmed & Courville, 2020). This setup is equivalent to that presented in our motivation (Section 3), but here we evaluate a broad range of recent state-of-the-art AD models.

In the CIFAR-10 and STL-10 semantic AD benchmarks 9 out of 10 object classes form the normal data S_n (*e.g*. all classes except dogs), so that images from multiple classes form a multi-modal normal distribution $\mathbb{P}^+$. The single class that is left out (*i.e*. dogs) is declared anomalous and never seen during training. At test time, the AD model has to identify the held-out class, *i.e*. we measure whether $f_\theta(x) \approx 1$ when x contains a dog. This requires that the AD model has a good semantic understanding of the objects in the normal distribution $\mathbb{P}^+$, and this benchmark has been shown to be more difficult than the popular one-versus-rest AD benchmarks (Ahmed & Courville, 2020; Bergman et al., 2020) (which we also evaluate in Section 5.3).

Ahmed & Courville (2020) determine semantic anomalies via MSP (Hendrycks & Gimpel, 2017) and ODIN (Liang et al., 2018) using an auxiliary self-supervised criterion akin to RotNet (Gidaris et al., 2018), while Bergman et al. (2020) use a nearest neighbor search over fixed pretrained features. We include all existing results in Table 2.

Optimization As before, for ADIB we set $\alpha = 10^{-2}$ following the suggestion of Li et al. (2018); in elastic weight consolidation (EWC) we set the Fisher multiplier to 400, as recommended by Kirkpatrick et al. (2017). For experiments on CIFAR-10 (Krizhevsky & Hinton, 2009), we introduce an inductive bias by regularizing the network weights towards those of ResNet26 trained on ImageNet at 32x32 resolution. We use the same architecture for STL-10 (Coates et al., 2011), but since images have a higher resolution the initial model weights were obtained by training on ImageNet at a resolution of 96x96.

For eqs. (1) and (2) we contrast S_n against images from an unstructured corpus Q_m. Following previous work that uses OE (Hendrycks et al., 2019b; Ruff et al., 2020a), for CIFAR-10 we fix this to contain all samples from the CIFAR-100 training split. As already emphasized, Q_m equals weak supervision: CIFAR-100 gives a viable surrogate learning signal, however does not contain examples of the anomalous CIFAR-10 categories. STL-10 contains a large unlabeled split, which we use for OE.

Results There are discrepancies in how performance is reported in the semantic AD literature: some authors recommend average precision (AP) (Ahmed & Courville, 2020), while others report AUC (Bergman & Hoshen, 2020). We include both metrics in Table 2, and report AP for STL-10 (Table 3) as this benchmark was so far only evaluated by Ahmed & Courville (2020) who report AP as AUC is overly optimistic for the STL-10 semantic AD benchmark (Davis & Goadrich, 2006).

On the CIFAR-10 semantic AD benchmark, ADIB outperforms the previously reported methods by a substantial margin: 74.6 *vs*. 41.6 mAP, and 95.1 *vs*. 71.7 mAUC. Even though it requires a smaller number of learnable parameters ADRA comes very close: 95.0 mAUC, and 72.9 mAP.

As our results confirm, inferring anomalies on STL-10 is significantly harder. In particular, even when using a state-of-the-art HSC classifier (Ruff et al., 2020a) initialized with pretrained θ_0 but without regularization $\Omega(\theta)$, this does not successfully address the semantic AD task (mAP of 27.7,

	mAUC	mAP
ODIN (Ahmed & Courville, 2020)	—	41.6
GT (Golan & El-Yaniv, 2018)	61.7	—
kNN-AD (Bergman & Hoshen, 2020)	71.7	—
ADRA *(ours)*	95.0 *(0.1)*	72.9 *(0.4)*
ADIB *(ours)*	**95.1** *(0.1)*	**74.6** *(0.3)*

Table 2. Results on the CIFAR-10 semantic AD benchmark; GT reported in Bergman & Hoshen (2020).

Class	ODIN	HSC	ADRA	ADIB
Airplane	23.4	23.1	**49.3** *(8.9)*	41.4 *(7.4)*
Bird	40.1	13.8	18.9 *(8.1)*	**44.0** *(2.9)*
Car	16.9	39.9	**74.6** *(6.5)*	72.2 *(10.5)*
Cat	31.4	18.9	29.6 *(3.4)*	**51.0** *(2.1)*
Deer	29.7	25.3	20.7 *(1.9)*	**43.0** *(5.7)*
Dog	26.1	17.3	26.6 *(3.5)*	**32.2** *(3.1)*
Horse	23.6	30.1	52.5 *(5.9)*	**53.7** *(2.5)*
Monkey	28.3	18.4	23.0 *(2.7)*	**46.6** *(1.9)*
Ship	15.4	49.2	**69.2** *(2.6)*	51.7 *(8.7)*
Truck	16.6	40.7	**64.3** *(2.2)*	58.7 *(3.6)*
mAP	25.1	27.7	42.9 *(1.4)*	**49.5** *(1.2)*

Table 3. APs on the STL-10 semantic AD benchmark.

Table 3). When adding a regularization term performance improves to 35.0 mAP ($\mathtt{a}_3$ in Table 4), supporting our assumption that variations that are important to determining anomalies at test time are *forgotten* during training, yielding poorer performance across classes.

ADIB improves performance to 49.5 mean AP. While having much fewer effective parameters, ADRA almost matches this performance (42.9 mAP) — interestingly, ADRA outperforms all other AD models on STL-10 man-made objects (cars, trucks, *etc*.), potentially due to there being a minority of examples of human-made objects in CIFAR-10 and a reported increase in robustness of linear bypasses on smaller modes (Deecke et al., 2020).

Ablation The transfer of features from rich semantic tasks to AD has to be carried out *carefully*. We examine this in an ablation shown in Table 4, for which we use the exact same model in each experiment $\mathtt{a}_1$–$\mathtt{a}_7$, and only switch on and off individual components: starting from random ($\mathtt{a}_1$) or pretrained models without regularization ($\mathtt{a}_2$) is not sufficient, as also highlighted in our intervention experiments in Section 5.1. Using an HSC loss (Ruff et al., 2020a) with the exact same explicit inductive bias through $\Omega(\theta)$ that we use in ADIB reduces performance ($\mathtt{a}_3$ *vs*. $\mathtt{a}_7$). DOC (Perera & Patel, 2019) is conceptually very similar to HSC, combining a radial compactness loss with a descriptiveness loss that requires ImageNet data. Our result ($\mathtt{a}_3$ *vs*. $\mathtt{a}_4$) confirms they also behave very similarly performance-wise.

	$\mathcal{L}$	Tr.	Reg.	CIFAR-10	STL-10
$\mathtt{a}_1$	eq. (1)	✗	✗	60.3	32.9
$\mathtt{a}_2$	eq. (1)	✓	✗	64.9	38.6
$\mathtt{a}_3$	HSC	✓	L_2	68.5	35.0
$\mathtt{a}_4$	DOC	✓	✗	65.8	35.2
$\mathtt{a}_5$	eq. (2)	✓	EWC	66.4	39.7
$\mathtt{a}_6$	ADRA *(ours)*			72.9 *(0.3)*	42.9 *(1.4)*
$\mathtt{a}_7$	ADIB *(ours)*			**74.6** *(0.3)*	**49.5** *(1.2)*

Table 4. Ablations in terms of mAP. Tr. indicates absence or presence of transfer learning; Reg. that of regularization. Included are comparisons against hyperspherical classifiers (Ruff et al., 2020a) in $\mathtt{a}_3$, and EWC (Kirkpatrick et al., 2017) in $\mathtt{a}_5$.

In $\mathtt{a}_5$ we find that EWC, a popular strategy for continual learning that regularizes weights via the Fisher information (Kirkpatrick et al., 2017), performs poorly compared to ADIB. This makes perfect sense, as EWC was designed to slow down learning on model weights relevant for the pretraining task: when pretraining on a demanding task like ImageNet, this can restrain capacity. Crucially for our scenario we need model capacity to free up and focus on semantic AD instead. In other words, as we never return the model to ImageNet classification, there simply is *no good reason* why we would want to preserve it. This ablation shows that, while they may be simple, our proposed strategies are surprisingly effective AD strategies.

Qualitative Analysis In Figure 3 we visualize the high semantic association of the representation learned by ADIB, and compute feature embeddings for samples from CIFAR-10, mapped to two dimensions using t-SNE (van der Maaten & Hinton, 2008). The representation was learned on the CIFAR-10 semantic AD setup, *i.e.* trained on a multi-category $\mathbb{P}^+$ that contains 9 out of 10 classes (• cat, • dog, *etc*.). At test time the singular anomalous category (• bird) gets revealed to the model.

While ADIB has no access to semantic categories or labels, it organizes the feature space in a *highly semantic* manner: • deer and • horses are semantically similar and cluster together, so do • cats and • dogs, while • frogs are separated from the other animals. Moreover, man-made objects such as • cars, • trucks, *etc*. are clearly separated from animal categories. This matches human intuition.

Even though the anomalous • bird category is never seen during training, it is located near other animals. We display one bird that has a similar feature representation to man-made objects: for this image, it is indeed difficult to identify it as a bird, and it should be no surprise that it is located far away from the •-cluster in feature space.

In Figure 4 we display examples from STL-10 that have been assigned a high anomaly score by ADIB. The anoma-

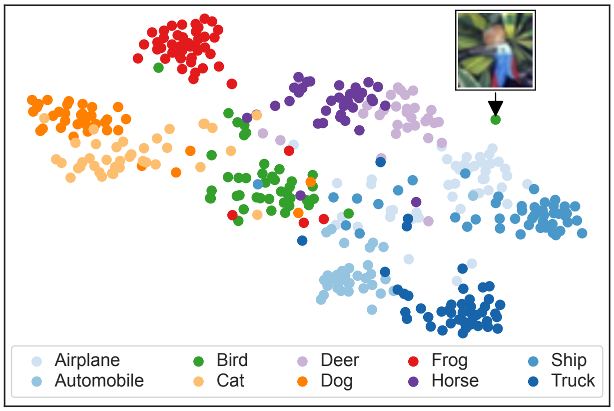

Figure 3. ADIB learns a feature space that is *semantically* meaningful. It separates objects that occur naturally (• cats, • dogs, *etc.*) from man-made ones (•, •, •, •). It even locates examples from the *unseen* bird category (•) nearby other animals. The arrow highlights a bird that gets mapped close to man-made objects, and identifying it as one indeed requires a fair bit of imagination.

(a) Cats. (b) Dogs.

(c) Horses. (d) Monkeys.

Figure 4. Anomalous examples from STL-10.

lous images are indeed unusual: either because animals appear in an unexpected pose (*e.g.* cat reaching for camera), because of the presence of captions, or in some cases – such as dogs – because the underlying object class is almost impossible to discern from the image.

5.3. Non-Semantic AD

We evaluate the performance of ADIB and ADRA on the standard CIFAR-10 one-versus-rest AD benchmark, which has recently been deemed a *non-semantic* problem by Ahmed & Courville (2020). While this is a less complex benchmark that can be solved using shallower feature representations, it is reported across large parts of the AD literature (Ruff et al., 2018; Deecke et al., 2018; Golan & El-Yaniv, 2018; Hendrycks et al., 2019b; Abati et al., 2019; Hendrycks et al., 2019c; Perera et al., 2019; Bergman & Hoshen, 2020; Ruff et al., 2020b;a) and therefore is still meaningful for comparison of our proposed methods to previous AD models.

In some sense, this benchmark can be viewed as opposite of semantic AD: only a single object class is fixed as the normal class — say, dogs. All dogs in the CIFAR-10 training split are collected into S_n (so 5000 out of 50 000 total samples), from which models are trained. Models are then evaluated against the entire CIFAR-10 test split, and performance is measured by checking whether anomaly scores assigned to dogs are lower than scores assigned to all nine remaining non-dog classes.

For this benchmark previous works almost exclusively report AUC, and we follow this custom here. Note that the optimization settings remain unchanged from those in Section 5.2.

Class	GT	kNN	GT+	HSC	ADRA	ADIB
Airplane	74.7	93.9	90.4	96.7	99.0 *(0.1)*	**99.2** *(0.3)*
Automobile	95.7	97.7	99.3	98.9	99.7 *(0.1)*	**99.8** *(0.1)*
Bird	78.1	85.5	93.7	93.2	97.5 *(0.4)*	**98.6** *(0.2)*
Cat	72.4	85.5	88.1	90.6	96.3 *(0.4)*	**97.0** *(0.7)*
Deer	87.8	93.6	97.4	97.1	98.9 *(0.1)*	**99.3** *(0.1)*
Dog	87.8	91.3	94.3	94.7	97.7 *(0.2)*	**98.2** *(0.3)*
Frog	83.4	94.3	97.1	98.0	**99.6** *(0.1)*	**99.6** *(0.2)*
Horse	95.5	93.6	98.8	97.9	99.6 *(0.1)*	**99.8** *(0.1)*
Ship	93.3	95.1	98.7	98.2	99.5 *(0.1)*	**99.6** *(0.1)*
Truck	91.3	95.3	98.5	97.7	99.4 *(0.1)*	**99.5** *(0.2)*
mAUC	86.0	92.5	95.6	96.3	98.7 *(0.1)*	**99.1** *(0.1)*

Table 5. AUCs for different methods on the CIFAR-10 one-versus-rest AD benchmark. Included are geometric transformations (GT) (Golan & El-Yaniv, 2018), kNN-AD (Bergman et al., 2020), self-supervised transformations (GT+) (Hendrycks et al., 2019c), and hyperspherical classifiers (HSC) (Ruff et al., 2020a).

OE Dataset	HSC	ADRA	ADIB
SVHN	70.2	75.3 (+5.1)	79.8 (+9.6)
CIFAR-100	96.3	98.7 (+2.4)	99.1 (+2.8)

Table 6. Ablations on the CIFAR-10 one-versus-rest AD benchmark for different choices of OE. Relative gain (*vs.* HSC) displayed in parentheses.

Results As shown in Table 5, ADIB raises the current state of the art to 99.1 mAUC, a marked gap to the previous best method with 96.1 mAUC. As demonstrated by the performance of kNN-AD (Bergman et al., 2020), simply using features from a large pretrained network is inferior when looking to detecting anomalies.

These results suggest that favorable inductive biases are critical for utilizing AD models to their full potential. ADRA once again comes very close in terms of performance, while requiring a much smaller number of learnable parameters.

Ablation Recent work examined the hierarchical relationship between distributions for out-of-distribution detection (Schirrmeister et al., 2020). We take inspiration from their work and critically examine the role of CIFAR-100 as OE in an ablation that compares it to the use of SVHN as OE.[2]

Results in Table 6 make it evident that SVHN is less well suited for CIFAR-10, as performance drops for all methods. For HSC (the current state-of-the-art AD method using OE) we obtain 70.2 mAUC here. A sizeable drop, but still improving from 64.8 mAUC for DSVDD (the mathematical equivalent to using no OE).

Our method obtains 79.8 mAUC when coupled with SVHN, a gain of +9.6 over HSC. This is a considerably larger difference than the gain for CIFAR-10 coupled with CIFAR-100 of +2.8 reported in Table 5 (ours: 99.1 *vs*. HSC: 96.3 mAUC), indicating that transfer-based AD benefits performance *more* when not using CIFAR-100 as OE (albeit it is better suited overall). In other words, the importance of the transfer task *increases* as the suitability of OE decreases.

5.4. Robustness to Small Modes

An ideal AD model has the ability to incorporate information from normal examples even if they form only a minor mode of $\mathbb{P}^+$, in the sense that only few samples from this class are contained in S_n — for example a rare dog breed. Since AD is concerned with low-probability events, the ability to robustly incorporate such small modes from few examples is of special importance.

To measure AD robustness, we let the normal class be constituted by samples associated with two classes (y_a, y_b), such that $S_n \sim \frac{1}{r+1}\mathbb{P}_{y_a} + \frac{r}{r+1}\mathbb{P}_{y_b}$, where the minor mode amplitude $r \in [0,1]$ controls the number of examples from y_b in the normal data. For a robust AD model, even as S_n is relaxed to contain only examples from y_a, its ability to identify the smaller category y_b as non-anomalous would remain intact.

We use CIFAR-10 here, and report primary and secondary AUCs as a function of r for $y_a =$ "ship" and $y_b =$ "truck" in Figure 5. We compare our methods to SAD (Ruff et al., 2020b) with pretrained weights, which corresponds to ADIB with $\alpha = 0$, *i.e.* without a regularization term $\Omega(\theta)$.

For SAD performance for the secondary class decreases much faster than for our methods. This trend is consistent across class pairings (more pairs are displayed in the supplements), and indicates that adequate transfer-based regularization as in ADRA and ADIB is crucial to robustly incorporating small modes of data in AD.

[2] While some previous work such as Hendrycks et al. (2019b) or Hendrycks et al. (2019c) used 80 Million Tiny Images instead of CIFAR-100 for OE, the dataset was withdrawn and further use has been discouraged by the authors.

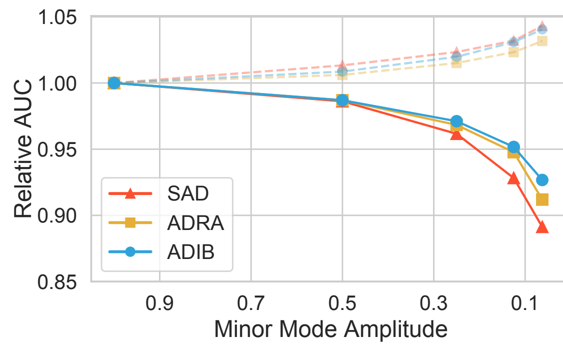

Figure 5. Relative AUCs for the secondary object category $y_b =$ "truck". Dashed curves display primary class performance.

6. Conclusion

Detecting semantic anomalies is a difficult task, due to the infinite and complex ways these can manifest themselves. We have proposed two methods to account for such complexities: ADIB sets a new state of the art in semantic AD tasks and ADRA provides a highly efficient, yet surprisingly effective learning protocol.

We used interventions to show that our methods can detect subtle semantic anomalies, and verified that ADIB and ADRA offer high AD robustness. An interesting question for future research is whether detecting anomalies requires disentanglement, and if it can benefit from the ongoing development of disentangled representations.

7. Acknowledgments

We thank the anonymous reviewers for their help in improving this paper. 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). HB is funded by the EPSRC programme grant Visual AI EP/T028572/1.

References

Abati, D., Porrello, A., Calderara, S., and Cucchiara, R. Latent space autoregression for novelty detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 481–490, 2019.

Adhikari, A., Ram, A., Tang, R., and Lin, J. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*, 2019.

Ahmed, F. and Courville, A. Detecting semantic anomalies. In *AAAI Conference on Artificial Intelligence*, pp. 3154–3162, 2020.

Ahmed, F., Bengio, Y., van Seijen, H., and Courville, A. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021.

Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P. GANomaly: Semi-supervised anomaly detection via adversarial training. In *Asian Conference on Computer Vision*, pp. 622–637. Springer, 2018.

Asano, Y. M., Rupprecht, C., and Vedaldi, A. A critical analysis of self-supervision, or what we can learn from a single image. In *International Conference on Learning Representations*, 2020.

Berg, A., Felsberg, M., and Ahlberg, J. Unsupervised adversarial learning of anomaly detection in the wild. In *European Conference on Artificial Intelligence*, 2019.

Bergman, L. and Hoshen, Y. Classification-based anomaly detection for general data. In *International Conference on Learning Representations*, 2020.

Bergman, L., Cohen, N., and Hoshen, Y. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*, 2020.

Bhagoji, A. N., Chakraborty, S., Mittal, P., and Calo, S. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, pp. 634–643, 2019.

Bouchacourt, D., Tomioka, R., and Nowozin, S. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI Conference on Artificial Intelligence*, pp. 2095–2102, 2018.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3): 15, 2009.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2610–2620, 2018.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 215–223, 2011.

Davis, J. and Goadrich, M. The relationship between precision-recall and ROC curves. In *International Conference on Machine Learning*, pp. 233–240, 2006.

Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., and Kloft, M. Image anomaly detection with generative adversarial networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 3–17. Springer, 2018.

Deecke, L., Hospedales, T., and Bilen, H. Latent domain learning with dynamic residual adapters. *arXiv preprint arXiv:2006.00996*, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. DeCAF: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pp. 647–655, 2014.

Eastwood, C. and Williams, C. K. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Edgeworth, F. On discordant observations. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 23(5):364–375, 1887.

Emmott, A. F., Das, S., Dietterich, T., Fern, A., and Wong, W.-K. Systematic construction of anomaly detection benchmarks from real data. In *ACM SIGKDD Workshop on Outlier Detection and Description*, pp. 16–21. ACM, 2013.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., and Leckie, C. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition*, 58:121–134, 2016.

Ghafoori, Z. and Leckie, C. Deep multi-sphere support vector data description. In *SIAM International Conference on Data Mining*, pp. 109–117, 2020.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Girshick, R. Fast R-CNN. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1440–1448, 2015.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

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

Gondal, M. W., Wuthrich, M., Miladinovic, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., and Bauer, S. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In *Advances in Neural Information Processing Systems*, pp. 15714–15725, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Goyal, S., Raghunathan, A., Jain, M., Simhadri, H. V., and Jain, P. DROCC: Deep robust one-class classification. In *International Conference on Machine Learning*, pp. 11335–11345, 2020.

Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., and Feris, R. SpotTune: transfer learning through adaptive fine-tuning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4805–4814, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

He, K., Girshick, R., and Dollár, P. Rethinking ImageNet pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4918–4927, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pp. 2712–2721, 2019a.

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

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

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Ionescu, R. T., Khan, F. S., Georgescu, M.-I., and Shao, L. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7842–7851, 2019.

Kim, H. and Mnih, A. Disentangling by factorising. In *International Conference on Learning Representations*, 2018.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pp. 2539–2547, 2015.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten ZIP code recognition. *Neural Computation*, 1(4):541–551, 1989.

Li, X., Grandvalet, Y., and Davoine, F. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, pp. 2825–2834, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pp. 4114–4124, 2019.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pp. 6348–6359, 2020.

Lopez-Paz, D. and Ranzato, M. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pp. 6467–6476, 2017.

Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. Anomaly detection in crowded scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1975–1981, 2010.

Mahendran, A. and Vedaldi, A. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, pp. 233–255, 2016.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. Advances in pre-training distributed word representations. In *International Conference on Language Resources and Evaluation*, 2018.

Ngo, P. C., Winarto, A. A., Kou, C. K. L., Park, S., Akram, F., and Lee, H. K. Fence GAN: towards better anomaly detection. In *IEEE International Conference on Tools with Artificial Intelligence*, pp. 141–148, 2019.

Ourston, D., Matzner, S., Stump, W., and Hopkins, B. Applications of hidden Markov models to detecting multi-stage network attacks. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*. IEEE, 2003.

Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. Deep learning for anomaly detection: A review. *ACM Comput. Surv.*, 54(2), 2020a.

Pang, G., Yan, C., Shen, C., Hengel, A. v. d., and Bai, X. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12173–12182, 2020b.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Pelleg, D. and Moore, A. W. Active learning for anomaly and rare-category detection. In *Advances in Neural Information Processing Systems*, pp. 1073–1080, 2005.

Perera, P. and Patel, V. M. Learning deep features for one-class classification. *IEEE Transactions on Image Processing*, 28(11):5450–5463, 2019.

Perera, P., Nallapati, R., and Xiang, B. OCGAN: One-class novelty detection using GANs with constrained latent representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2898–2906, 2019.

Pidhorskyi, S., Almohsen, R., and Doretto, G. Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in Neural Information Processing Systems*, pp. 6822–6833, 2018.

Rebuffi, S.-A., Bilen, H., and Vedaldi, A. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, pp. 506–516, 2017.

Rebuffi, S.-A., Bilen, H., and Vedaldi, A. Efficient parametrization of multi-domain deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8119–8127, 2018.

Rippel, O., Mertens, P., and Merhof, D. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In *International Conference on Pattern Recognition*, 2020.

Ruff, L., Vandermeulen, R., Görnitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep one-class classification. In *International Conference on Machine Learning*, pp. 4393–4402, 2018.

Ruff, L., Vandermeulen, R. A., Franks, B. J., Müller, K.-R., and Kloft, M. Rethinking assumptions in deep anomaly detection. *arXiv preprint arXiv:2006.00339*, 2020a.

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

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

Sabokrou, M., Khalooei, M., Fathy, M., and Adeli, E. Adversarially learned one-class classifier for novelty detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3379–3388, 2018.

Schirrmeister, R. T., Zhou, Y., Ball, T., and Zhang, D. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. In *Advances in Neural Information Processing Systems*, pp. 21038–21049, 2020.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pp. 146–157. Springer, 2017.

Schölkopf, B. and Smola, A. J. *Learning with Kernels*. MIT Press, 2002.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. Technical Report MSR-TR-99-87, Microsoft Research, 1999.

Singliar, T. and Hauskrecht, M. Towards a learning traffic incident detection system. In *Workshop on Machine Learning Algorithms for Surveillance and Event Detection, International Conference on Machine Learning*, 2006.

Sohn, K., Li, C.-L., Yoon, J., Jin, M., and Pfister, T. Learning and evaluating representations for deep one-class classification. In *International Conference on Learning Representations*, 2021.

Steinwart, I., Hush, D., and Scovel, C. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(Feb):211–232, 2005.

Stickland, A. C. and Murray, I. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pp. 5986–5995, 2019.

Sultani, W., Chen, C., and Shah, M. Real-world anomaly detection in surveillance videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018.

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

Tax, D. M. and Müller, K.-R. Feature extraction for one-class classification. In *Artificial Neural Networks and Neural Information Processing*, pp. 342–349. Springer, 2003.

Torralba, A., Fergus, R., and Freeman, W. T. 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.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605, 2008.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):1–19, 2019.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pp. 3320–3328, 2014.

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., and Savarese, S. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833, 2014.

Zhai, S., Cheng, Y., Lu, W., and Zhang, Z. Deep structured energy based models for anomaly detection. In *International Conference on Machine Learning*, pp. 1100–1109, 2016.

Zhang, R., Isola, P., and Efros, A. A. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Zhou, C. and Paffenroth, R. C. Anomaly detection with robust deep autoencoders. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 665–674, 2017.

Zhou, Z.-H. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018.