PANDA: Adapting Pretrained Features for Anomaly Detection and Segmentation

Tal Reiss *, Niv Cohen *, Liron Bergman & Yedid Hoshen
School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel

Abstract

Anomaly detection methods require high-quality features. In recent years, the anomaly detection community has attempted to obtain better features using advances in deep self-supervised feature learning. Surprisingly, a very promising direction, using pre-trained deep features, has been mostly overlooked. In this paper, we first empirically establish the perhaps expected, but unreported result, that combining pre-trained features with simple anomaly detection and segmentation methods convincingly outperforms, much more complex, state-of-the-art methods.

In order to obtain further performance gains in anomaly detection, we adapt pre-trained features to the target distribution. Although transfer learning methods are well established in multi-class classification problems, the one-class classification (OCC) setting is not as well explored. It turns out that naive adaptation methods, which typically work well in supervised learning, often result in catastrophic collapse (feature deterioration) and reduce performance in OCC settings. A popular OCC method, DeepSVDD, advocates using specialized architectures, but this limits the adaptation performance gain. We propose two methods for combating collapse: i) a variant of early stopping that dynamically learns the stopping iteration ii) elastic regularization inspired by continual learning. Our method, PANDA, outperforms the state-of-the-art in the OCC, outlier exposure and anomaly segmentation settings by large margins[1].

1. Introduction

Detecting anomalous patterns in data is of key importance in science and industry. In the computational anomaly detection task, the learner observes a set of training examples. The learner is then tasked to classify novel test samples as normal or anomalous. There are multiple anomaly detection settings investigated in the literature, corresponding to different training conditions. In this work, we deal with three settings: i) anomaly detection - when only normal images are used for training ii) anomaly segmentation - detecting all the pixels that contain anomalies, given normal images as input. iii) Outlier Exposure (OE) - where an external dataset simulating the anomalies is available.

In recent years, deep learning methods have been introduced for anomaly detection, typically extending classical methods with deep neural networks. Different auxiliary tasks (e.g. autoencoders or rotation classification) are used to learn representations of the data, while a great variety of anomaly criteria are then used to determine if a given sample is normal or anomalous. An important issue for current methods is the reliance on limited normal training data for representation learning, which limits the quality of learned representations. Nearly all state-of-the-art anomaly detection methods rely on self-supervised feature learning - i.e. using the limited normal training data for learning strong features. The motivation for this is twofold: i) the fear that features trained on auxiliary domains will not generalize well to the target domain. ii) the curiosity to investigate the top performance achievable without ever looking at any external dataset (we do not address this question here).

In other parts of computer vision, features pre-trained on external datasets are often used to improve performance on tasks trained on new domains - and our reasonable hypothesis is that this should also be the case for image anomaly detection and segmentation. We present very simple baselines that use pretrained features trained on a large external data and K-nearest neighbor (kNN) retrieval to significantly outperform all previous methods on anomaly detection and segmentation, even on images of distant target domains.

We then tackle the technical challenge of obtaining stronger performance by further adaptation to the normal training data. Although feature adaptation has been extensively researched in the multi-class classification setting, limited work was done in the OCC setting. Unfortunately, it turns out that feature adaptation for anomaly detection often suffers from *catastrophic collapse* - a form of deterioration of the pre-trained features, where all (including anomalous) samples, are mapped to the same point. DeepSVDD [23] proposed to overcome collapse by removing biases from the

*Equal contribution
[1]The code is available at github.com/talreiss/PANDA

model architecture, but this restricts network expressively and limits the pre-trained models that can be borrowed off-the-shelf. Perera and Patel [21] proposed to jointly train OCC with the original task which has several limitations and achieves only limited adaptation success.

Our first finding is that simple training with constant-duration early stopping (with no bells-and-whistles) already achieves top performance. To remove the dependence on the number of epochs for early stopping, we propose two techniques to overcome catastrophic collapse: i) an adaptive early stopping method that selects the stopping iteration per-sample, using a novel generalization criterion - this technique is designed to overcome a special problem of OCC, namely that there are no anomalies in the validation set ii) elastic regularization, motivated by continual learning, that postpones the collapse. Thorough experiments demonstrate that we outperform the state-of-the-art by a wide margin (ROCAUC): e.g. CIFAR10 results: 96.2% vs. 90.1% without outlier exposure and 98.9% vs. 95.6% with outlier exposure. We also achieve 96.0% vs. 89.0% on anomaly segmentation on MVTec.

We present insightful critical analyses: i) We show that pre-trained features strictly dominate current self-supervised RotNet-based feature learning methods. We discuss the relative merits of each paradigm and conclude that for most practical purposes, using pre-trained features is preferable. ii) We analyse the results of the popular DeepSVDD method and discover that its feature adaptation, which is designed to prevent collapse, does not improve over simple data whitening.

Contributions: To summarize our main ocntributions in this paper:

- Demonstrating that a simple baseline outperforms all current methods in image anomaly detection and segmentation - extensive analysis shows the generality of the result.
- Identifying that popular SOTA methods do not outperform linear whitening in OCC feature adaptation.
- Proposing several effective solutions for feature adaptation for OCC.
- Extensive evaluation, obtaining results that significantly improve over the current state-of-the-art.

1.1. Related Work

Classical anomaly detection: The main categories of classical anomaly detection methods are: i) reconstruction-based: compressing the training data using a bottleneck, and using a reconstruction loss as an anomaly criterion (e.g. [4, 17], K nearest neighbors [7] and K-means [12]), ii) probabilistic: modeling the probability density function and labeling unlikely sampled as anomalous (e.g. Ensembles of Gaussian Mixture Models [9], kernel density estimate [19]) iii) one-class classification (OCC): finding a separating manifold between normal data and the rest of input space (e.g. One-class SVM [25]).

Deep learning methods: The introduction of deep learning has affected image anomaly detection in two ways: extension of classical methods with deep representations and novel self-supervised deep methods. Reconstruction-based methods have been enhanced by learning deep autoencoder-based bottlenecks [6] which can provide better models of image data. Deep methods extended classical methods by creating a better representations of the data for parametric assumptions about probabilities, a combination of reconstruction and probabilistic methods (such as DAGMM [28]), or in a combination with OCC methods [23]. Novel deep methods have also been proposed for anomaly detection including GAN-based methods [28]. Another set of novel deep methods use auxiliary self-supervised learning for anomaly detection. The seminal work by [10] was later extended by [15] and [1].

Transferring pretrained representations: Learning deep features requires extensive datasets, preferably with labels. An attractive property of deep neural networks, is that representations learned on very extensive datasets, can be transferred to data-poor tasks. Specifically deep neural representations trained on the ImageNet dataset have been shown by [16] to significantly boost performance on other datasets that are only vaguely related to some of the ImageNet classes. This can be performed with and without finetuning. Although much recent progress has been performed on self-supervised feature learning [8, 5], such methods are typically outperformed by transferred pretrained features. Transferring ImageNet pre-trained features for out-of-distribution detection has been proposed by [13]. Similar pre-training has been proposed for one-class classification has been proposed by [21], however they require joint optimization with the original task. Rippel et. al. [22] follow an early version of this paper and report results with pre-trained features on MVTec using the Mahalanobis distance.

Anomaly segmentation methods: Segmenting the image pixels that contain anomalies has attracted far less research attention than image-level anomaly detection. Several previous anomaly segmentation works used pre-trained features, but they have not convincingly outperformed top self-supervised methods. Napoletano et al. [20] extracted deep features from small overlapping patches, and used a K-means based classifier over dimensionality reduced features. Bergmann et al. [2] evaluated both a ADGAN and autoencoder approaches on MVTec dataset [2] finding complementary strengths. More recently, Venkataramanan et al. [27] used an attention-guided VAE approach combining multiple methods (GAN loss [11], GRADCAM [26]).

Bergmann et al. [3] used a student-teacher based autoencoder approach employing pre-trained ImageNet deep features. Our simple baseline, SPADE, significantly outperforms the previously mentioned approaches.

2. A General Framework and Simple Baselines for Anomaly Detection and Segmentation

2.1. A Three-stage Framework

We present our general framework in which we examine several adaptation-based anomaly detection methods, including our method. Let us assume that we are given a set $\mathcal{D}_{train}$ of normal training samples: $x_1, x_2..x_N$. The framework consists of three steps:

Initial feature extractor: An initial feature extractor ψ_0 can be obtained by pre-training on an auxiliary task with loss function $L_{pretrain}$. The auxiliary task can be either pre-training on an external dataset (e.g. ImageNet) or by self-supervised learning (auto-encoding, rotation or jigsaw prediction). In the former case, the pretrained extractor can be obtained off-the-shelf. The choice of auxiliary tasks is analyzed in Sec. 4.3.

Feature adaptation: Features trained on auxiliary tasks or datasets may require adaptation before being used for anomaly scoring on the target data. This is seen as a finetuning stage of the features on the target training data. We denote the feature extractor after adaptation ψ.

Anomaly scoring: Having adapted the features for anomaly detection, we extract the features $\psi(x_1), \psi(x_2)..\psi(x_N)$ of the training set samples. We then proceed to learn a scoring function, which describes how anomalous a sample is. Typically, the scoring function seeks to measure the density of normal data around the test sample $\psi(x)$ (either by direct estimation or via some auxiliary task) and assign a high anomaly score to low density regions.

2.2. Simple Baselines for Anomaly Detection and Segmentation

We report very simple-to-implement but highly effective baselines for anomaly detection and segmentation, based on our framework. In the anomaly detection baseline "Deep Nearest Neighbours" (**DN2**), the feature extractor is a large ResNet pretrained the ImageNet dataset. We use the distance from the features of the kNN normal images as the anomaly score. In the anomaly segmentation baseline, "Semantic Pyramid Anomaly Detection" (**SPADE**), we use an ImageNet-pretrained ResNet to extract per-pixel features for all images. As both the low-level high-resolution features, and semantic low-resolution context are important for determining if a pixel is anomalous, we join the feature representations of the pixel extracted from multiple layers of the deep neural network. We score the pixel by its kNN distance from the feature descriptors of the pixels of all training images. In both baselines we skip the adaptation stage. Implementation details of both methods, including the K-Means-based speedup, can be found in the Supplementary Material (SM).

Figure 1: (left) An anomalous image (right) The predicted anomalous image pixels.

3. Feature Adaptation for Anomaly Detection

Although our two simple-to-implement baselines, DN2 and SPADE, achieve very strong results, we ask if feature adaptation can improve them further. We first review two existing methods for feature adaption for anomaly detection, and proceed to propose our method, PANDA, which significantly improves over them.

3.1. Background: Existing Feature-Adaptation Methods

DeepSVDD: Ruff et al. [23] suggest to first train an autoencoder on the normal-only train images. The encoder is then used as the initial feature extractor ψ_0. As the features of the encoder are not specifically adapted to anomaly detection, DeepSVDD adapts ψ on the training data. The adaptation takes place by minimizing the compactness loss:

$$L_{compact} = \sum_{x \in \mathcal{D}_{train}} \|\psi(x) - c\|^2 \quad (1)$$

Where c is a constant vector, typically the average of $\psi_0(x)$ on the training set. However, the authors were concerned of the trivial solution $\psi = c$, and suggested architectural restrictions to mitigate it, most importantly removing the biases from all layers. We empirically show that the effect of adaptation of the features in DeepSVDD does not outperform simple feature whitening (see Sec. 4.3.2).

Joint optimization (JO): Perera et al. [21] proposed to use a deep feature extractor trained for object classification on the ImageNet dataset. Due to fear of "learning a trivial solution in the absence of a penalty for miss-classification", the method does not adapt by finetuning on the compactness loss only. Instead, they relaxed the task setting, by assuming that a number ($\sim 50k$) of labelled original ImageNet images, $\mathcal{D}_{pretrain}$, are still available at adaptation time. They

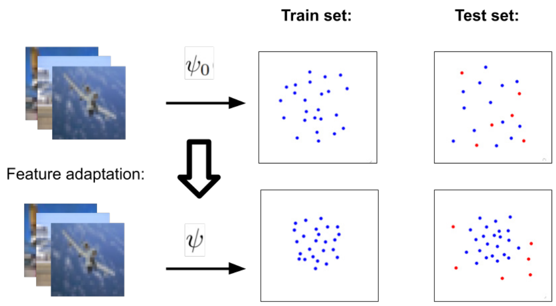

Figure 2: An illustration of our feature adaptation procedure, the pre-trained feature extractor ψ_0 is adapted to make the normal features (blue) more compact resulting in feature extractor ψ. After adaptation, anomalous test features (red) lie in a less dense region of the feature space.

proposed to train the features ψ under the compactness loss jointly with the original ImageNet classification linear layer W and its classification loss, here the CE loss with the true label $\ell_{pretrain}(p, y) = -\log(p_y)$, and SMax indicates Softmax:

$$L_{Joint} = \sum_{(x,y) \in \mathcal{D}_{pretrain}} \ell_{pretrain}(SMax(W\psi(x)), y) + \alpha \cdot \sum_{x \in \mathcal{D}_{train}} \|\psi(x) - c\|^2 \quad (2)$$

Where W is the final linear classification layer and α is a hyper-parameter weighting the two losses. We note that the method has two main weaknesses: i) it requires retaining a significant number of the original training images which can be storage intensive ii) jointly training the two tasks may reduce the anomaly detection task accuracy, which is the only task of interest in this context.

3.2. PANDA: Pre-trained Anomaly Detection Adaptation

We present PANDA, a new method for anomaly detection in images. Similarly to SVDD and Joint Optimization, we also use the compactness loss (Eq. 1) to adapt the general pre-trained features to the task of anomaly detection on the target distribution. Instead of constraining the architecture or introducing external data into the adaptation procedure we tackle catastrophic collapse directly. The main challenge is that the optimal solution of the compactness loss can result in "collapse", where all possible input values are mapped to the same point ($\psi(x) = c, \quad \forall x$). Learning such features will not be useful for anomaly detection, as both normal and anomalous images will be mapped to the same output, preventing separability. The issue is broader than the trivial "collapsed" solution after full convergence, but rather the more general issue of feature deterioration, where the original good properties of the pretrained features are lost. Even a non-trivial solution might lose some of the discrimnative properties of the original features which are none-the-less important for anomaly detection.

To avoid this collapse, we suggest three options: (i) finetuning the pretrained extractor with compactness loss (Eq.1) and stopping after a constrant number of iterations (ii) a novel method for determining early stopping per-sample (iii) when collapse happens prematurely, before any significant adaptation happens, we suggest mitigating it using a Continual Learning-inspired adaptive regularization.

Simple early stopping (PANDA-Early): An embarrassingly simple but effective solution for controlling the collapse of the original features is to stop training after a constant number of iterations (e.g. 15 epochs on CIFAR10). Inversely scaling the number of epochs by dataset size works for most examined datasets (Sec. 4.3).

Sample-wise early stopping (PANDA-SES): A weakness of the simple early-stopping approach, is the reliance on a hyper-parameter that may not generalize to new datasets. Although the optimal stopping epoch can be determined with a validation set containing anomalies, it is not available in our setting. We thus propose "samplewise early stopping" (SES) as an unsupervised way of determining the stopping epoch from a single sample. The intuition for the method can be obtained from Fig. 3. We can see that anomaly detection accuracy is correlated to having a large ratio between the distance of the anomalous samples to the center, and the distance between the normal samples and the center. We thus propose to save checkpoints of our network at fixed intervals (every 5 epochs) during the training process - corresponding to different early stopping iterations ($\psi_1, \psi_2..\psi_T$), for each network ψ_t we compute the average distance on the training set images s_t. During inference, we score a target image x using each model $s_t^{target} = \|\psi_t(x) - c\|^2$, and normalize the score by the training average score by s_t. We set the maximal ratio, as the anomaly score of this sample, as this roughly estimates the model that achieves the best separation between normal and such anomalous samples.

Continual Learning (PANDA-EWC): We propose a new solution for overcoming premature feature collapse that draws inspiration from the field of continual learning. The task of continual learning tackles learning new tasks without forgetting the previously learned ones. We note however that our task is not identical to standard continual learning as: i) we deal with the one-class classification setting whereas continual-learning typically deals with multi-class classification ii) we aim to avoid forgetting the expressivity of the features but do not particularly care if the actual classification performance on the old task is degraded. A simple solution for preventing feature collapse is regularization of the change in value of the weights of the feature extractor.

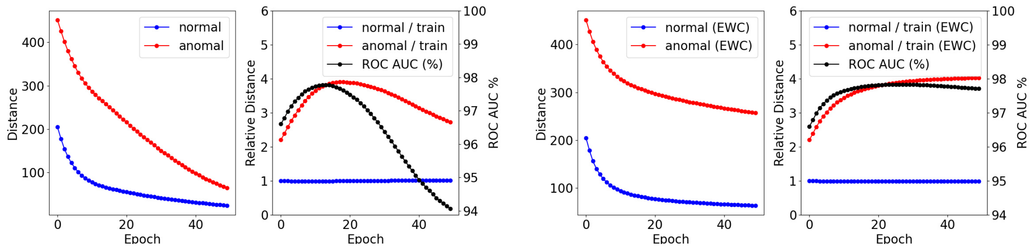

Figure 3: CIFAR100 Class 17 (right to left): (1) - During training all samples approach the center of train set features (2) - When normalized by the train average distance s_t, the normal samples stay dense, while the anomalous ones initially move further away and then "collapse". The ROC AUC performance behaves similarly to the anomalous samples' normalized distance. (3),(4) - when training with EWC the collapse is mitigated.

However, this solution is lacking as some weights influence the features more than others.

Following ideas from continual learning, we use elastic weight consolidation (EWC) [18]. Using a number of mini-batches (we use 100) to pretrain on the auxiliary task. We compute the diagonal of the Fisher information matrix F for all weight parameters of the network. Note that this only needs to happen once at the end of the pretraining stage. The value of the Fisher matrix for diagonal element θ is given by:

$$F_\theta = \mathbb{E}_{(x,y)\in\mathcal{D}_{pretrain}}\left[\left(\frac{\partial}{\partial\theta}L_{pretrain}(x,y)\right)^2\right] \quad (3)$$

We follow [18] in using the diagonal of the Fisher information matrix F_{θ_i}, to weight the squared distance of the change of each network pretrained weight $\theta_i \in \psi_0$ and fine-tuned weight $\theta_i^* \in \psi$. This can be seen as a measure of the loss landscape curvature as function of the weights - larger values imply high curvature, inelastic weights.

We use this regularization in combination with the compactness loss weighted by the factor λ, which is a hyperparameter of the method (we use $\lambda = 10^4$):

$$L_\theta = L_{compact}(\theta^*) + \frac{\lambda}{2}\cdot\sum_i F_{\theta_i}(\theta_i - \theta_i^*)^2 \quad (4)$$

The network ψ is initialized with the parameters of the pretrained extractor ψ_0 and trained with SGD.

Anomaly scoring: As in classical anomaly detection, scoring can be done by density estimation. Unless mentioned otherwise, we use kNN for scoring. We also evaluate faster methods and get similar results (see Sec. 4.3.3).

Outlier Exposure: An extension of the typical image anomaly detection task [14], assumes the existence of an auxiliary dataset of images $\mathcal{D}_{OE}$, which are more similar to the anomalies than normal data. In case such information is available, we simply train a linear classification layer w together with the features ψ under a logistic regression loss (Eq. 5). As before, ψ is initialized with the weights from ψ_0. After training ψ and w, we use $w\cdot\psi(x)$ as the anomaly score. Results and critical analysis of this setting are presented in Sec. 4.3.

$$L_{OE} = \sum_{x\in\mathcal{D}_{train}} log(1-\sigma(w\cdot\psi(x))) + \sum_{x\in\mathcal{D}_{OE}} log(\sigma(w\cdot\psi(x))) \quad (5)$$

4. Image Anomaly Detection

4.1. Experiments

In this section, we present high-level results of the our simple baselines, and our full method PANDA-EWC, (PANDA-SES can be found in Sec.4.3) compared to the state-of-the-art: One-class SVM [25], DeepSVDD [23], Multi-Head RotNet [15]. All the results of others that were available in the original papers were copied exactly. In cases that the result was not available, we run the experiments ourselves (where possible). As Joint Optimization requires extra data, we did not add it to this table, but compare and outperform it in Tab. 6. We compare our PANDA-OE to the OE baseline in [15] on CIFAR10, as the code or results for other classes were unavailable. Note that unless specifically mentioned otherwise, PANDA results were run with kNN. PANDA-OE used the original classifier - which performs a little better than kNN. We compare SPADE and relevant state-of-the-art baselines on anomaly segmentation.

We evaluated our method on a wide range of datasets (Tab. 1 Fig. 4) demonstrating different challenges in image anomaly detection. They are described in the SM.

4.2. High-Level Results

The main results are i) Pre-trained features achieve significantly better results than self-supervised features on all datasets, both in anomaly detection and segmentation. ii) Feature adaptation significantly improves the performance on larger datasets iii) Outlier exposure can further improve

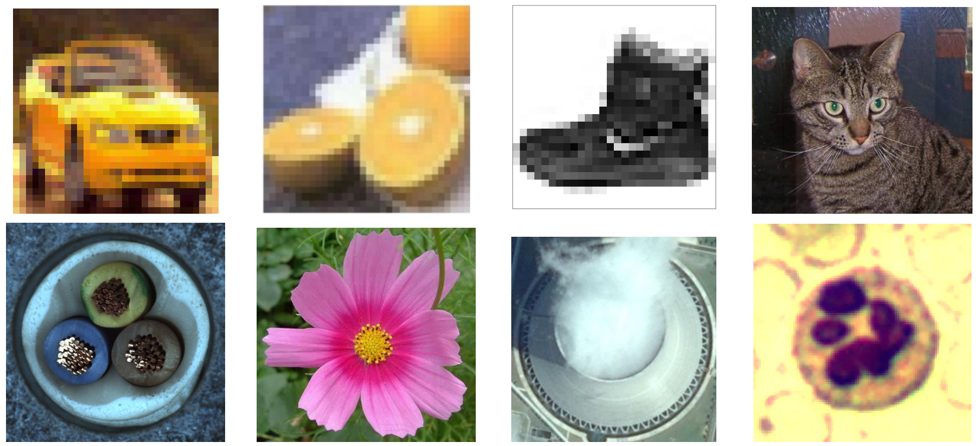

Figure 4: Representative images of the different datasets, from the left clockwise: CIFAR10, CIFAR100, Fashion MNIST, DogsVsCats, WBC, DIOR, Oxford Flowers and MVTec. Following standard propotocol, in all datasets (except MVTec), normal data are one class (e.g. cat in CIFAR10) while anomalies are all other test data from the same dataset (e.g. dog, car in CIFAR10). MVTec contains class-specific anomlies (e.g. for normsl class - Wire, anomalies include bent wires)

Table 1: Details of datasets used for evaluation - number of classes, and average number of normal train and (normal and anomalous) test images per-class

Dataset	$N_{classes}$	N_{train}	N_{test}
CIFAR10	10	5,000	10,000
Fashion MNIST	10	6,000	10,000
CIFAR100	20	2,500	10,000
Flowers	102	10	7,169
Birds	200	30	5,794
CatsVsDogs	2	10,000	5,000
MVTec	15	242	1,725
WBC	4	59	62
DIOR	19	649	9,243

performance in the case where the given outliers are more similar to the anomalies than the normal data.

OE achieves near perfect performance on CIFAR10/100 but hurts performance for Fashion MNIST/CatsVsDogs which are less similar to the 80M Tiny images dataset. A detailed analysis of the reason for better performance for each of these methods and an examination of its appropriateness will be presented in Sec. 4.3.

4.3. Analysis and Further Evaluation

4.3.1 An analysis of feature representations

A comparison of self-supervised and pre-trained features: In Tab. 2 and Tab. 3, we present a comparison between methods that use self-supervised and pre-trained feature representations. We see that the autoencoder used by DeepSVDD is particularly poor. The results of the MHRotNet as a feature extractor are better, but still underperform PANDA methods (see SM for more details). The performance of the raw deep ResNet features without adaptation significantly outperforms all methods, including Fashion MNIST and DIOR which have significant differences from the ImageNet dataset. We can therefore conclude that ImageNet-pretrained features typically have significant advantages over self-supervised features. Tab. 3 shows that self-supervised methods do not perform well on small datasets as such methods require large numbers of normal samples in order to learn strong features. On the other hand ImageNet-pretrained features obtain very strong results.

Does the supriority of pretrained features extend to very different domains? The results in Tab. 3 on FMNIST, DIOR, WBC, MVTec suggest that out-of-domain pretrained features are better at anomaly detection than in-domain self-supervised features. We tested on datasets of various sizes, domains, resolutions and symmetries. On all those datasets pretrained features outperformed the SOTA. These datasets include significantly different objects from those of ImageNet, but also fine-grained intra-object anomalies, and represent a spectrum of data types: aerial images, microscopy, industrial images. This shows that one of the main arguments against using pre-trained features, generalizing to distant domains, is not an issue in practice.

Our simple pretrained feature baseline (SPADE) is extremely effective for anomaly segmentation: In Tab.4 we can see that our simple no-training baseline, SPADE, outperforms previous methods for anomaly segmentation, including those that use trained and pretrained features (see SM for metrics, specifications and detailed results). While we suspect feature adaptation can be used for further performance gain even for anomaly segmentation, we find that the MVTec dataset is too small for significant feature adap-

Table 2: Anomaly detection performance (Average ROC AUC %)

Dataset	Self-Supervised			Pretrained		OE	
	OC-SVM	DeepSVDD	MHRot	DN2	PANDA	MHRot	PANDA-OE
CIFAR10	64.7	64.8	90.1	92.5	**96.2**	95.6	**98.9**
CIFAR100	62.6	67.0	80.1	**94.1**	**94.1**	-	**97.3**
FMNIST	92.8	84.8	93.2	94.5	**95.6**	-	91.8
CatsVsDogs	51.7	50.5	86.0	96.0	**97.3**	-	94.5
DIOR	70.7	70.0	73.3	93.0	**94.3**	-	**95.9**

Table 3: Pretrained feature performance on various small datasets (Average ROC AUC %)

Dataset	Self-Supervised			Pretrained
	OCSVM	DeepSVDD	MHRot	DN2
Birds	62.0	60.8	64.4	**95.3**
Flowers	74.5	78.1	65.9	**94.1**
MvTec	70.8	77.9	65.5	**86.5**
WBC	75.4	71.2	57.7	**87.4**

tation using the compactness loss. We believe that feature adaptation for segmentation calls for new adaptation methods, this is left for future work.

On the different supervision settings for one-class anomaly detection: Anomaly detection methods employ different levels of supervision. Within the one-class classification task, one may use outlier exposure (OE) - an external dataset (e.g. ImageNet), pretrained features, or no external supervision at all. The most extensive supervision is used by OE, which requires a large external dataset at training time, and performs well only when such a dataset is from a similar domain to the anomalies (see Tab. 2). In cases where the dataset used for OE has significantly different properties, the network may not learn to distinguish between normal and anomalous data, as the normal and anomalous data may have more in common than the OE dataset. E.g. both normal and anomalous classes of Fashion MNIST are grayscale, OE using 80M Tiny Images will not be helpful, as the network may learn to classify only according to color. Pretrained features further improve OE, in cases where is suitable e.g. CIFAR10.

Pretraining, like Outlier Exposure, is also achieved through an external labelled dataset, but differently from OE, the external dataset is only required once - at the pretraining stage and is not used again. Additionally, the same features are applicable for very different image domains from that of the pretraining dataset (e.g. Fashion MNIST - grayscale images, DIOR - aerial images, WBC- medical images, MVTec - industrial images). Self supervised feature learning requires no external dataset at all, which can potentially be an advantage. While there might be image anomaly detection tasks where ImageNet-pretrained weights are not applicable, we saw no evidence for such cases after examining a broad spectrum of domains and datasets (Tab. 1). This indicates that the extra supervision of the ImageNet-pretrained weights comes at virtually no cost.

Can pretrained features boost the performance of RotNet-based methods? We did not find evidence that pretrained features improve the performance of RotNet-based AD methods such as [15] (CIFAR10: 90.1% vs. 86.6% without and with pretraining). As can be seen in Tab. 5, pretrained features improve the auxiliary task performance on the normal data, but also on the anomalous samples. As such methods rely on a generalization gap between normal and anomalous samples, deep features actually reduce this gap, as a solution to the auxiliary task becomes feasible for both types of images. For a more detailed analysis see SM.

4.3.2 Feature adaptation methods

Benefits of feature adaptation: Feature adaptation aims to make the distribution of the normal samples more compact, w.r.t. the anomalous samples. Our approach of finetuning pretrained features for compactness under EWC regularization, significantly improves the performance over "raw" pretrained features (see Tab.2). While the distance from the normal train samples center, of both normal and anomalous test samples is reduced (see Fig.3), the average distance from the center of anomalous test samples is typically higher than that of normal samples, in relative terms, which makes anomalies easier to detect.

While PANDA-EWC may train more than $7.8k$ minibatches without catastrophic collapse on CIFAR10, performance of training without regularization usually peaks higher but collapse earlier. We therefore set our constant early stopping epoch such that the net trains with to $2.3k$ minibatches on all datasets for comparison. Our PANDA-SES method usually achieves an anomaly score not far from the unregularized early stopping peak performance, but is most important in cases where unregularized training fails.

A comparison of features adaptation methods: In Tab. 6 we compare PANDA against (i) JO [21] - co-training compactness with ImageNet classification which requires

Table 4: Comparison of anomaly segmentation methods (pixel-level ROCAUC and PRO %)

	AE_{SSIM} [2]	AE_{L2} [2]	AnoGAN [24]	CNN Dict [20]	CAVGA-R_u [27]	Student [3]	SPADE
ROCAUC	87	82	74	78	89	-	**96.2**
PRO	69.4	79	-	51.5	-	85.7	**92.1**

Table 5: Comparison of average transformation prediction accuracy (%), horiz. = horizontal, rot. = rotation.

Method	Normal		Anomalous	
	Horiz.	Rot.	Horiz.	Rot.
Self-supervised	94.0	94.0	67.9	51.6
Pretrained	94.4	92.3	71.4	61.3

Table 6: A comparison of different feature adaptation methods (Avg. ROC AUC %)

Dataset	Baseline	PANDA		
	JO	Early	SES	EWC
CIFAR10	93.2	**96.2**	95.9	**96.2**
CIFAR100	91.1	**94.8**	94.6	94.1
FMNIST	94.9	95.4	95.5	**95.6**
CatsVsDogs	96.1	91.9	95.7	**97.3**
DIOR	93.1	95.4	**95.6**	94.3

ImageNet data at training time. We can see that PANDA - EWC always outperforms JO feature adaptation. (ii) PANDA early stopping, generally has higher performance than PANDA-EWC, but has severe collapse issues on some classes. (iii) PANDA-SES is similar to early stopping, but PANDA-SES does not collapse as badly on CatsVsDogs dataset. We note that replacing the Fisher matrix by equally weighting the changes in all parameters ($\sum_i(\theta_i - \theta_i^*)^2$) achieves similar results to early stopping.

Which are the best layers to finetune? Fine-tuning all layers is prone to feature collapse, even with continual learning. We therefore recommend finetuining only layers 3 & 4 (see ablation in SM).

DeepSVDD architectural changes: DeepSVDD [23] proposes various architectural changes, such as removing the bias parameters from the network, to prevent collapse to trivial features. To understand whether DeepSVDD gains its significant performance from its pretrained features or from its feature adaptation, we tried to replace its feature adaptation by closed-form linear data whitening. For both pretrained features and anomaly scoring, we used the DeepSVDD original code [23]. We found empirically that the results obtained by the constrained architecture were about the same as those achieved with simple whitening of the data (64.8% vs. 64.6%, see SM). We ablated DeepSVDD by running it with the original LeNet (including biases) and found this did not deteriorate its anomaly detection performance. As architectural modifications are not the focus of this work, further investigation into architectures less prone to feature collapse is left for future work.

4.3.3 Anomaly scoring functions

Does kNN improve over distance to the center? kNN achieves an improvement of around 2% on average w.r.t. to distance to the center (CIFAR10: 94.2% vs 96.2%).

Can we improve over the linear complexity of kNN? A naive implementation of kNN has linear runtime complexity in the number of training samples. For anomaly segmentation, 50 means give a speedup from 2.7 frames-per-second to 41 frames-per-second (faster than real-time), with ~0.5% ROCAUC decrease. For anomaly detection, even for very large datasets, or many thousands of means, both kNN and K-means can run faster than real-time.

5. Conclusion and Outlook

We first proposed simple baseline methods for anomaly detection and segmentation, that outperform the state-of-the-art. We further improved over the strong baselines by proposing a method that adapts pretrained features and mitigates catastrophic collapse. We showed that our results significantly outperform current methods while addressing their limitations. We analysed the reasons for the strong performance of our method We note that the question of the optimal performance on image anomaly detection without ever having access to auxiliary data is unaddressed here, however we believe it is of mostly pure academic interest.

The main limitation of this work is the requirement for strong pretrained feature extractors. Much work was done on transferable image and text features and it is likely that current extractors can be effective to obtain features for time series and audio as well. Generic feature extractors are not currently available for tabular data, their development is an exciting direction for future work.

Acknowledgements

This work was partly supported by the Federmann Cyber Security Research Center in conjunction with the Israel National Cyber Directorate.

References

[1] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *ICLR*, 2020. 2

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019. 2, 8

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4183–4192, 2020. 3, 8

[4] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *JACM*, 2011. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2

[6] Pierluca D'Oro, Ennio Nasca, Jonathan Masci, and Matteo Matteucci. Group anomaly detection via graph autoencoders. *NIPS Workshop*, 2019. 2

[7] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pages 77–101. Springer, 2002. 2

[8] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[9] Michael Glodek, Martin Schels, and Friedhelm Schwenker. Ensemble gaussian mixture models for probability density estimation. *Computational Statistics*, 28(1):127–138, 2013. 2

[10] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *NeurIPS*, 2018. 2

[11] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 2

[12] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1979. 2

[13] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *arXiv preprint arXiv:1901.09960*, 2019. 2

[14] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 5

[15] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, 2019. 2, 5, 7

[16] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 2

[17] Ian Jolliffe. *Principal component analysis*. Springer, 2011. 2

[18] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 5

[19] Longin Jan Latecki, Aleksandar Lazarevic, and Dragoljub Pokrajac. Outlier detection with kernel density functions. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 61–75. Springer, 2007. 2

[20] Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Anomaly detection in nanofibrous materials by cnn-based self-similarity. *Sensors*, 18(1):209, 2018. 2, 8

[21] Pramuditha Perera and Vishal M Patel. Learning deep features for one-class classification. *IEEE Transactions on Image Processing*, 28(11):5450–5463, 2019. 2, 3, 7

[22] Oliver Rippel, Patrick Mertens, and Dorit Merhof. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. *arXiv preprint arXiv:2005.14140*, 2020. 2

[23] Lukas Ruff, Nico Gornitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *ICML*, 2018. 1, 2, 3, 5, 8

[24] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg 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. 8

[25] Bernhard Scholkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In *NIPS*, 2000. 2, 5

[26] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[27] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. Attention guided anomaly localization in images. In *European Conference on Computer Vision*, pages 485–503. Springer, 2020. 2, 8

[28] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. *ICLR*, 2018. 2