SPot-the-Difference Self-Supervised Pre-training for Anomaly Detection and Segmentation

Yang Zou[1], Jongheon Jeong[2⋆], Latha Pemula[1]
Dongqing Zhang[1], Onkar Dabeer[1]

[1]AWS AI Labs [2]KAIST
{yanzo,lppemula,zdongqin,onkardab}@amazon.com, jongheonj@kaist.ac.kr

Abstract. Visual anomaly detection is commonly used in industrial quality inspection. In this paper, we present a new dataset as well as a new self-supervised learning method for ImageNet pre-training to improve anomaly detection and segmentation in 1-class and 2-class 5/10/high-shot training setups. We release the Visual Anomaly (VisA) Dataset consisting of 10,821 high-resolution color images (9,621 normal and 1,200 anomalous samples) covering 12 objects in 3 domains, making it the largest industrial anomaly detection dataset to date. Both image and pixel-level labels are provided. We also propose a new self-supervised framework - SPot-the-difference (SPD) - which can regularize contrastive self-supervised pre-training, such as SimSiam, MoCo and SimCLR, to be more suitable for anomaly detection tasks. Our experiments on VisA and MVTec-AD dataset show that SPD consistently improves these contrastive pre-training baselines and even the supervised pre-training. For example, SPD improves Area Under the Precision-Recall curve (AU-PR) for anomaly segmentation by 5.9% and 6.8% over SimSiam and supervised pre-training respectively in the 2-class high-shot regime. We open-source the project at http://github.com/amazon-research/spot-diff.

Keywords: Representation learning, pre-training, anomaly detection, anomaly segmentation, industrial anomaly dataset

1 Introduction

Visual surface anomaly detection and segmentation identify and localize defects in industrial manufacturing [3]. While anomaly detection and segmentation are instances of image classification and semantic segmentation problems, respectively, they have unique challenges. First, defects are rare, and it is hard to obtain a large number of anomalous images. Second, common types of anomalies, such as surface scratches and damages, are often small. Fig. 1 (a) gives an example. Third, manufacturing is a performance sensitive domain and usually requires highly accurate models. Fourth, inspection in manufacturing spans a wide range of domains and tasks, from detecting leakages in capsules to finding damaged millimeter-sized components on a complex circuit board.

⋆ work done during an Amazon internship

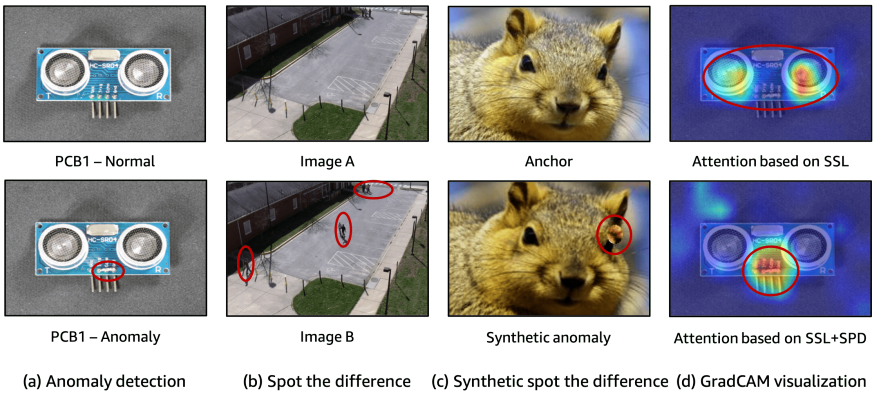

Fig. 1. (a) Normal and anomalous samples of VisA - PCB1 with real defect (molten metal), anomaly highlighted by red ellipse; (b) A pair of images for the spot-the-difference (SPD) puzzle [25]; (c) An anchor image and its variant augmented by SmoothBlend for synthetic spot-the-difference; (d) GradCAM attention visualization for PCB1 - Anomaly image based on self-supervised ImageNet pre-training w/wo proposed SPD. With SPD, attention is more focused on the local defects.

Upon the aforementioned challenges, previous surface anomaly detection models have been typically trained for a particular object and require re-training for different ones. For each object, there are only slight global differences in lighting and object pose/positions across images while the diversity in the defects on objects is large. Moreover, due to the rarity of anomalous data, there has been a predominant focus on 1-class anomaly detection, which only requires normal images for model training [6,10,14,26,33,44]. In mature manufacturing domains, anomalous samples are also available and sometimes sufficient. In such cases, one can improve over 1-class methods with a standard 2-class model [12,18,21,27] by incorporating the anomalous data in training, which is in fact a well-established practice in commercial visual inspection AI services [1,2]. For both setups, existing state-of-the-art methods for surface anomaly detection commonly leverage supervised representations pre-trained on ImageNet [16], either as feature extractors [14,33] or as initialization for fine-tuning on the target dataset [26,44].

Meanwhile, recent advances in self-supervised learning (SSL) have shown that pre-trained representations learned without categorical labels might be a better choice for transfer learning compared to those from supervised in object detection and segmentation [8,9,23]. However, their application to anomaly detection and segmentation is underdeveloped. SSL for surface anomaly detection was explored in CutPaste [26] to learn representation from downstream images for each specific object. However, such representations hardly generalize to different objects and can lead to overfitting in a practical setting where only 1-20 normal samples are available. Also, there are previous works focusing on SSL for high-level semantic anomaly detection such as cat among a distribution of dogs [11,13,37]. However, as [35] pointed out, surface anomaly detection aims

to spot the low-level textual anomalies such as scratch and crack which has challenges different from semantic anomaly detection. Until now, the universal self-supervised pre-trained representation with good generalization ability have not yet been attempted for surface anomaly detection and segmentation.

Regarding the evaluation protocol, the community has been experiencing the lack of challenging benchmarks. The popular MVTec Anomaly Detection (AD) benchmark [3] is saturating with the Area Under the Receiver Operating Characteristic (AU-ROC) approaching ∼95% [14,26], and the benchmark is limited to the 1-class setup. But the anomaly detection problems in practice is still far from solved, demanding new datasets and metrics that better represent the real-world. In this paper, we introduce a new challenging Visual Anomaly (VisA) dataset. VisA is collected to present several new characteristics: objects with complex structures such as printed circuit board (PCB), multiple instances with different locations in a single view, 12 different objects spanning 3 domains, and multiple anomaly classes (up to 9) for each object. VisA contains 10,821 high-resolution color images - 9,621 normal and 1,200 anomalous - with both image and pixel-level labels. To our best knowledge, VisA is currently the largest and most challenging public dataset for anomaly classification and segmentation. Moreover, to cover different use cases in practice, we establish benchmarks not only in standard 1-class training setup but also 2-class training setups with 5/10/high-shot. For evaluation, we propose to use Area Under the Precision-Recall curve (AU-PR) in combination with standard AU-ROC. In the imbalanced defect dataset, AU-ROC might present inflated view of performances and AU-PR is more informative to measure anomaly detection performance [11,13,37].

In addition to an improved dataset, we also explore self-supervision to improve anomaly detection. As we argue below, our hypothesis is that previous contrastive SSL methods [8,9,23] are sub-optimal to transfer learning for anomaly detection. Specifically, SimCLR, MoCo and other methods regard globally augmented images of a given image as one class and other images in the same batch as negative classes. Transformations, such as cropping and color jittering, are applied globally to the anchor for positives generation. The InfoNCE or cosine similarity losses [8,9,23] encourage invariance to these global deformations, and capturing semantic information instead of local details [19]. However, anomaly detection relies on local textual details to spot defects. Thus the subtle and local intra-object (or intra-class) differences are important but not well modeled by previous methods. Figure 1 (d) illustrates the sub-optimality in one of the previous SSL methods using the GradCAM attention map [38]. As far as we know, improving representations by self-supervision for better downstream anomaly detection/segmentation has not been studied before and we explore this angle.

Inspired by the spot-the-difference puzzle shown in Fig. 1 (b), we propose a contrastive SPot-the-Difference (SPD) training to promote the local sensitivity of previous SSL methods. In the puzzle, players need to be sensitive to the subtle differences between the two globally alike images, which is similar to anomaly detection. In the contrastive SPD training, as shown in Fig. 1 (c), a novel augmentation called SmoothBlend is proposed to produce the local perturbations on

SPD negatives for synthetic spot-the-difference. The (locally) augmented images are regarded as negatives, which is different from regarding (globally) augmented images as positives in SimCLR/MoCo. Moreover, weak global augmentations, such as weak cropping and color jittering, are also applied to the SPD negatives as anomaly detection should spot defects under slight global changes in lighting and object pose/position. Additionally, to prevent models from using the slight global changes as shortcuts to differentiate negatives, SPD positives are generated by applying weak global augmentations on the anchor. Lastly, SPD training minimizes the feature similarities between SPD negative pairs while maximizing the similarities between SPD positives, which encourages models to be locally sensitive to anomalous patterns and invariant to slight global variations.

Our main contributions are as follows:

1. We propose a new VisA dataset, 2× larger than MVTec-AD, with both image and pixel-level annotations. It spans 12 objects across 3 domains, with challenging scenarios including complex structures in objects, multiple instances and object pose/location variations. Moreover, we establish both 1-class and 5/10/high-shot 2-class benchmarks to cover different use cases.
2. To promote the local sensitivity to anomalous patterns, a SPot-the-Difference (SPD) training is proposed to regularize self-supervised ImageNet pre-training, which benefits their transfer-learning ability for anomaly detection and localization. As far as we know, we are the first one to explore self-supervised pre-training on large-scale datasets for surface defect detection tasks.
3. Compared to strong self-supervised pre-training baselines such as SimSiam, MoCo and SimCLR, extensive experiments show our proposed SPD learning improves them for better anomaly detection and segmentation. We also show the SPD improves over supervised ImageNet pre-training for both tasks.

2 Related Works

Unsupervised Anomaly Detection and Segmentation use only normal samples to train models, which have drawn extensive attention. Many recent methods are proposed to detect low-level texture anomalies [35], such as scratches and cracks, which are common cases in industrial visual inspection [15,31,34,44]. SPADE [14] and PatchCore [33] extract features at patch level and use nearest neighbor methods to classify patches and images as anomalies. PaDiM [14] learns a parametric distribution over patches for anomaly detection. CutPaste [26] learns a representation based on images augmented by cut-and-pasted patches. The supervised ImageNet models are used in these methods either as feature extractors or initialization for fine-tuning. However, self-supervised pre-training on large-scale datasets is an unexplored area for quality inspection applications. In addition, several works [30,36,39,40] focus on high-level semantic anomaly detection. As mentioned in [35], semantic anomaly detection approaches can be less effective for texture anomaly detection as their challenges are different.

Self-Supervised Learning (SSL) have gathered momentum in the last 5 years. Several surrogate tasks have been proposed for self-supervision, such as

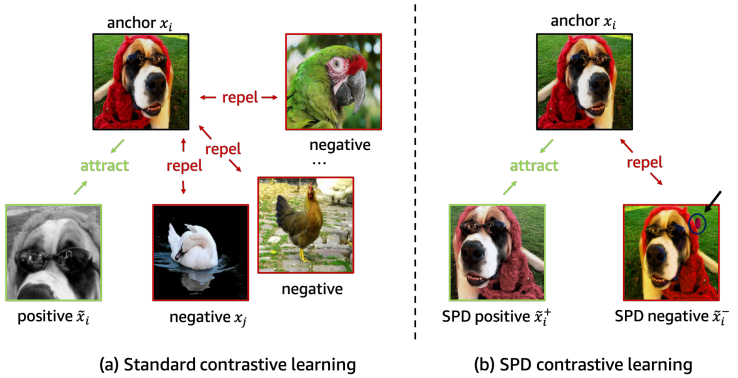

Fig. 2. (a) Contrastive learning in SimCLR, MoCo and SimSiam; (b) Contrastive learning in SPD training. Local deformation in SPD negative is highlighted by circle.

image colorization [46], rotation prediction [20], jigsaw puzzles [29]. Recently, multi-view based methods such as MoCo [23], SimCLR [8], SimSiam [9] and BYOL [22] present better or comparable performances than supervised pre-training in transfer learning tasks including image classification, object detection [43] and semantic segmentation [41]. Moreover, to promote spatial details of representations for localization tasks, several approaches proposed to encourage the invariance of patch features to global augmentations [7,28,41,42], although they may not lead to local sensitivity to tiny defects. As far as we know, none of these works explored their generalization ability to surface defect detection tasks.

3 SPot-the-Difference (SPD) Regularization

To promote local sensitivity of standard self-supervised contrastive learning, we propose a contrastive SPot-the-Difference (SPD) regularization. As mentioned earlier, SPD aims to increase model invariance to slight global changes by maximizing the feature similarity between an image and its weak global augmentation, while forcing dissimilarity for local perturbations, as shown in Fig. 2 (b). In the following, we first present background in contrastive learning, and then the augmentations used in SPD followed by the learning with SPD.

3.1 Background on Self-supervised Contrastive Learning

Many self-supervised learning methods, such as SimCLR [8] and MoCo [23], are based on contrastive learning. As shown in Fig. 2 (a), given an image, these methods maximize the feature similarity between two strongly augmented samples x_i and $\hat{x}_i$ while minimizing the similarities between the anchor x_i and other images x_j's in the same batch of size N. Strong global augmentations, such as grayscaling, large cropping and strong color jittering, are used to get positives.

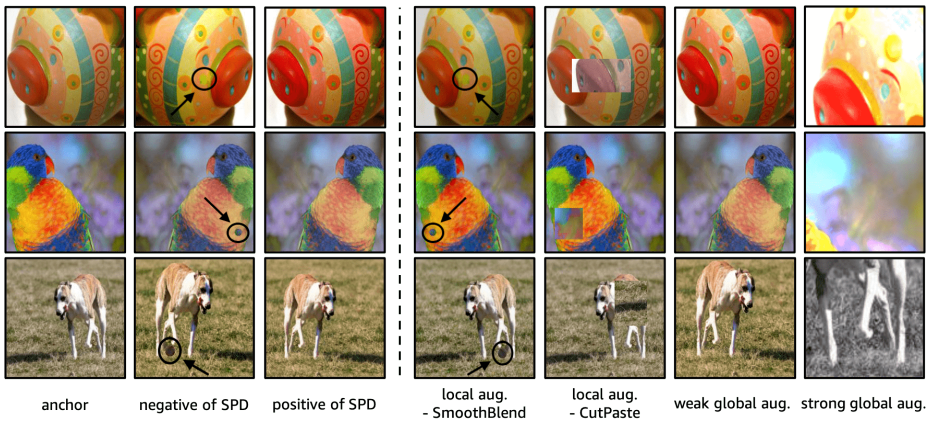

Fig. 3. (a) Samples for synthetic spot-the-difference; (b) Augmentation comparison

Typically, an encoder extracts features $h_i, \hat{h}_i$ and h_j's which are inputs to a multilayer perceptron (MLP) head. The MLP head extracts the L2 normalized embeddings $z_i, \hat{z}_i$ and z_j's to compute the InfoNCE loss defined as follows.

$$\mathcal{L}_{\text{NCE}}(x_i, \hat{x}_i) = -\log \frac{\exp\left(z_i \cdot \hat{z}_i/\tau\right)}{\exp\left(z_i \cdot \hat{z}_i/\tau\right) + \sum_{j=1}^{N} \mathbb{1}_{j \neq i} \exp\left(z_i, z_j/\tau\right)} \tag{1}$$

τ is a temperature scaling hyperparameter. In addition, SimSiam [9] shows that self-supervised models can be trained even without negatives where only similarity modeling is implemented for positives.

Remark: Images augmented by most strong global transformations in SSL, such as grayscaling and large cropping, share semantics with anchor but with different local details (a dog v.s. a dog head). Thus to maximize their similarity, the features are forced to be invariant about local details and capture the global semantics. This is even enforced by minimizing similarities between anchor and different images in a batch as they have different global structures [8,17]. This further motivates us to promote local sensitivity in SSL for anomaly detection.

3.2 Augmentations for SPD

Local augmentation: In SPD, the locally deformed images, rather than other images of a batch in standard contrastive training, are used as negatives. SmoothBlend is proposed to produce local deformations. The first column in Fig. 3 (b) presents the samples augmented by SmoothBlend. It is implemented by a smoothed alpha blending between an image and a small randomly cut patch of the same image. Specifically, color jittering is applied to a cut patch. Then an all-zero foreground layer u is created with the patch pasted to a random location. An alpha mask α is created where the pixels corresponding to the pasted patch are set to 1 otherwise 0, followed by a Gaussian blur. Finally, the augmented sample is obtained by $\bar{x} = (1 - \alpha) \odot x + \alpha \odot u$. $\odot$ is the element-wise product.

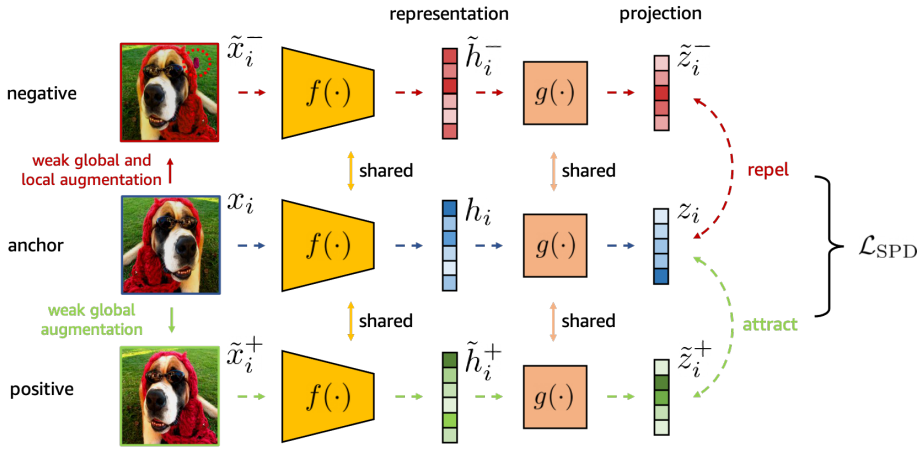

Fig. 4. The contrastive spot-the-difference learning

Global augmentation: To generate global variations for both SPD positives and negatives, we use weak global augmentation. Adding global variations to SPD is motivated by the potentially small global variations in realistic manufacturing environment, such as lighting, object positions, etc. To simulate such slight changes, we choose weak random cropping, Gaussian blurring, horizontal flipping and color jittering. Such weak global augmentations are different from strong transformations used in SimSiam, SimCLR and MoCo which is illustrated by last two columns in Fig. 3 (b). As we can see, there might be just 20% overlap between the anchor and strongly augmented positive. If the network is designed to maximize the distance between negatives with only subtle changes while minimizing the distance between positives with largely global transformations, it is a confusing task which might harm representation learning for anomaly detection.
Remark: SmoothBlend is a smoothed version of CutPaste augmentation proposed in [26]. Both of them can be used to generate structural local deformations, illustrated by the first two columns in Fig. 3 (b). Unlike the sharp edges of the CutPaste patches, the local and subtle perturbations with smooth edges from SmoothBlend provides a challenging puzzle for models.

3.3 Training with SPD

Based on the above augmentations, we propose the SPD learning illustrated by Fig. 4 with Fig. 3 (a) presents more SPD training samples. For an anchor image x_i, a negative $\tilde{x}_i^-$ is generated by applying weak global augmentations followed by SmoothBlend. The positive $\tilde{x}_i^+$ is produced by weak global transformations only. Then a shared feature extractor $f(\cdot)$ extracts the representations $h_i, \tilde{h}_i^-, \tilde{h}_i^+$ (h_i's are used for downstream anomaly detection tasks). They are further inputted into a shared multilayer perceptron (MLP) $g(\cdot)$ to get the projections $z_i, \tilde{z}_i^-, \tilde{z}_i^+$. The cosine similarity between $z_i, \tilde{z}_i^-$ is minimized while similarity between $z_i, \tilde{z}_i^+$

is maximized. In summary, the SPD learning minimizes the following SPD loss.

$$\mathcal{L}_{\text{SPD}}(x_i, \tilde{x}_i^-, \tilde{x}_i^+) = \cos(z_i, \tilde{z}_i^-) - \cos(z_i, \tilde{z}_i^+). \tag{2}$$

Standard contrastive SSL with SPD: Regularizing SSL with SPD is simple. Taking SimCLR as an example baseline, for a given image, SimCLR generates the anchor x_i and positive $\hat{x}_i$ via strong global augmentations with other images x_j's in the same batch as negatives. Then SPD positives $\hat{x}_i^+$ and negatives $\hat{x}_i^-$ are generated by SmoothBlend and weak global augmentations. The shared encoder and MLP head in SimCLR are used to extract the image feature projections for loss computation. Finally the network is trained by the following combined loss.

$$\mathcal{L}(x_i, \hat{x}_i, \tilde{x}_i^-, \tilde{x}_i^+) = \mathcal{L}_{\text{NCE}}(x_i, \hat{x}_i) + \eta \cdot \mathcal{L}_{\text{SPD}}(x_i, \tilde{x}_i^-, \tilde{x}_i^+) \tag{3}$$

Similary, we can apply SPD to MoCo. For SimSiam, $\mathcal{L}_{\text{NCE}}(x_i, \hat{x}_i)$ loss is replaced by a cosine distance loss for positive pairs without considering negatives [9].
Standard supervised pre-training with SPD: With the class labels, standard supervised pre-trained features also capture global semantics to distinguish categories with less attention to local details, similar to SSL. Thus SPD could improve its local sensitivity. Specifically, on top of the last feature layer of the standard supervised model (ResNet-50 [24]), an auxiliary classifier is added to classify if an augmented SPD image has a local perturbation or not, which is trained by cross-entropy loss. The backbone is shared to extract features.

4 Visual Anomaly (VisA) Dataset

4.1 Dataset Description

The VisA dataset contains 12 subsets corresponding to 12 different objects. Fig. 5 gives images in VisA. There are 10,821 images with 9,621 normal and 1,200 anomalous samples. Four subsets are different types of printed circuit boards (PCB) with relatively complex structures containing transistors, capacitors, chips, etc. For the case of multiple instances in a view, we collect four subsets: Capsules, Candles, Macaroni1 and Macaroni2. Instances in Capsules and Macaroni2 largely differ in locations and poses. Moreover, we collect four subsets including Cashew, Chewing gum, Fryum and Pipe fryum, where objects are roughly aligned. The anomalous images contain various flaws, including surface defects such as scratches, dents, color spots or crack, and structural defects like misplacement or missing parts. There are 5-20 images per defect type and an image may contain multiple defects. The

Table 1. Overview of VisA dataset

	Object	# normal samples	# anomaly samples	# anomaly classes
Complex structure	PCB1	1,004	100	4
	PCB2	1,001	100	4
	PCB3	1,006	100	4
	PCB4	1,005	100	7
Multiple instances	Capsules	602	100	5
	Candle	1,000	100	8
	Macaroni1	1,000	100	7
	Macaroni2	1,000	100	7
Single instance	Cashew	500	100	9
	Chewing gum	503	100	6
	Fryum	500	100	8
	Pipe fryum	500	100	6

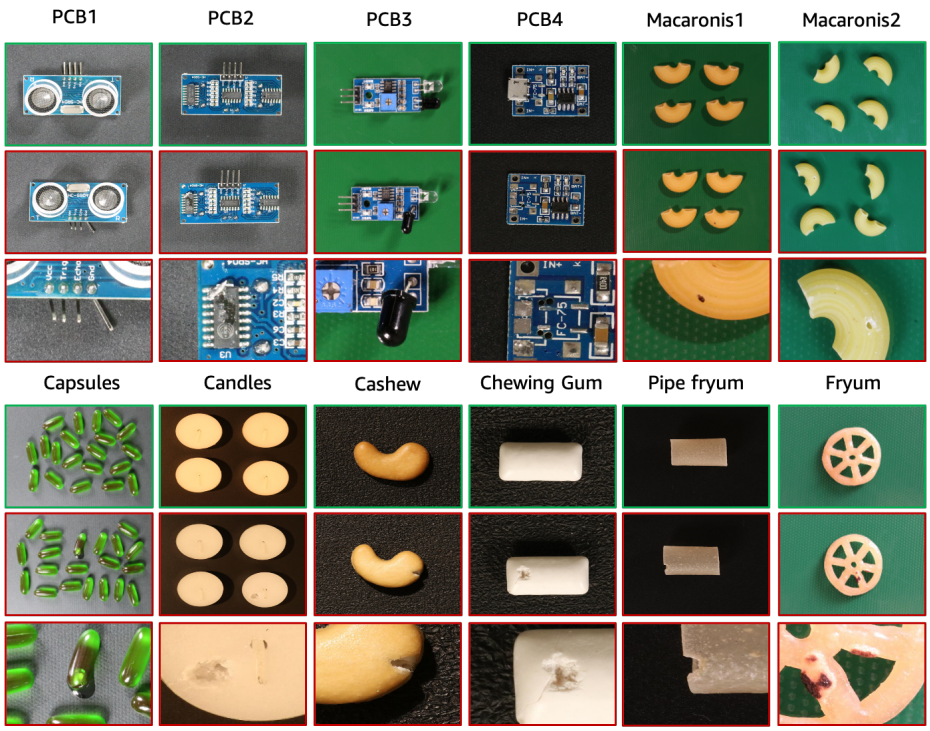

Fig. 5. Samples of VisA datasets. First row: normal images; Second row: anomalous images; Third row: anomalies viewed by zooming in.

defects were manually generated to produce realistic anomalies. All images were acquired using a $4,000 \times 6,000$ high-resolution RGB sensor. Both image and pixel-level annotations are provided. Table 1 gives the statistics of VisA dataset.

Fig. 6 illustrates the differences between VisA and MVTec-AD. First, VisA considers more complex structures, comparing the VisA - PCB3 with multiple electronic components to a single one of MVTec - transistor as an example. Second, multiple objects can appear in VisA (Capsules) as opposed to a single object in MVTec-AD. Third, large variation in object locations is covered by VisA (Capsules) while almost all objects in MVTec-AD are roughly aligned. Lastly, MVTec-AD has $5,354$ images and VisA is $2\times$ larger with $10,821$ images.

4.2 Evaluation Protocol and Metrics

We establish three evaluation protocols for each of 12 objects in VisA dataset. First, following MVTec-AD 1-class protocol, we establish VisA 1-class protocol by assigning 90% normal images to train set while 10% normal images and all anomalous samples are grouped as test set. Second, we establish 2-class high/low-

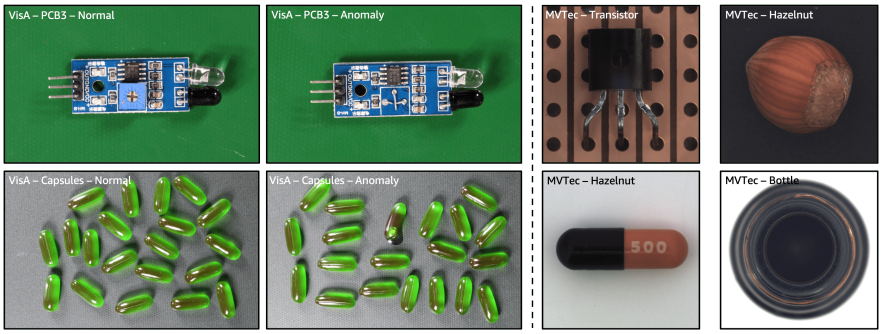

Fig. 6. Comparing VisA and MVTec-AD. VisA is more challenging due to the complex object structures, multiple instances, large variations of objects and scale.

shot evaluation protocols as proxies for realistic 2-class setups in commercial products [1,2]. In high-shot setup, for each object, 60%/40% normal and anomalous images are assigned to train/test set respectively. For low-shot benchmark, firstly, 20%/80% normal and anomalous images are grouped to train/test set respectively. Then the k-shot (k=5,10) setup randomly samples k images from both classes in train set for training. The averaged performances over 5 random runs will be reported. Note that for both 1-class and 2-class training setups, test sets have samples from both classes. In addition, we report model performances averaged over all subsets of VisA and MVTec-AD in Sec. 5. The model performances for each subset are reported in Sec. D of supplementary.

For metrics, we report Area Under Precision-Recall curve (AU-PR) in combination with the Area Under Receiver Operator Characteristic curve (AU-ROC). AU-ROC is the most widely used metric for anomaly detection tasks [14,33,44]. But as pointed out in [11,13,37], in imbalanced dataset where performance of minor class is more important, AU-ROC might provide an inflated view of performance which may cause challenges in measuring models' true capabilities. This is true for anomaly detection where anomalies are often rare. In [3], the best method is Student-Teacher [5] with 92.2% AU-ROC which seems to be close to perfection. However, it only gets 59.9% AU-PR which is far-from satisfactory. The imbalance issue is more extreme in anomaly segmentation where normal pixels (negatives) can be tens/hundreds times more than anomalous pixels (positives). Even for a bad model, the false positive rate can be small due to numerous negatives, leading to a high AU-ROC. Thus we argue AU-PR is a better performance measurement. Our experiments also demonstrate this point.

5 Experiments

Datasets: For self-supervised as well as supervised pre-training, we use ImageNet 2012 classification dataset [16]. ImageNet consists 1,000 classes with 1.28 million training images. For downstream tasks, in addition to our VisA dataset,

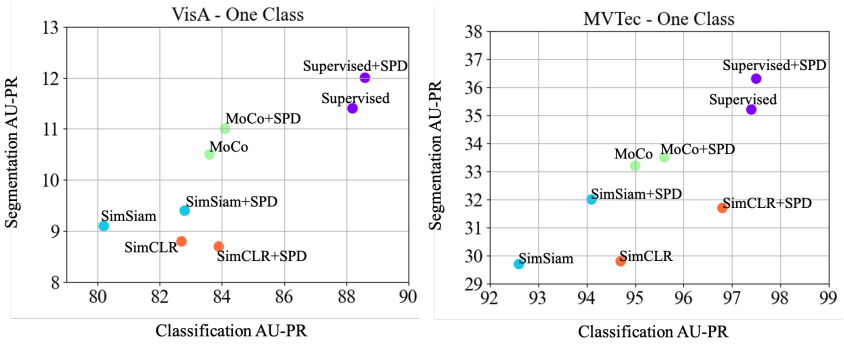

Fig. 7. Scatter plots for various ImageNet pre-training models in 1-class setup.

Table 2. 1-class performance evaluation of various ImageNet pre-training options on VisA and MVTec-AD with PaDiM. Bold numbers refers to the highest score. In the brackets are the gaps to the ImageNet supervised/self-supervised pre-training counterpart. In green are the gaps of at least +0.5 point.

	ImageNet labels	VisA (1-class)				MVTec-AD (1-class)			
		Classification		Segmentation		Classification		Segmentation	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Sup. pre-train	✓	88.2	87.8	11.4	93.1	97.4	94.5	35.2	94.4
SimSiam	✗	80.2	78.1	9.1	93.1	92.6	83.9	29.7	92.1
+SPD	✗	82.8 (+2.6)	81.2 (+3.1)	9.4 (+0.3)	92.7 (-0.4)	94.1 (+1.5)	88.0 (+4.1)	32.0 (+2.3)	92.2 (+0.1)
MoCo	✗	83.6	83.4	10.5	93.4	95.0	90.4	33.2	93.4
+SPD	✗	84.1 (+0.5)	83.0 (-0.4)	11.0 (+0.5)	93.5 (+0.1)	95.6 (+0.6)	90.5 (+0.1)	33.5 (+0.3)	93.5 (+0.1)
SimCLR	✗	82.7	81.6	8.8	89.7	94.7	90.7	29.8	92.1
+SPD	✗	83.9 (+0.8)	82.6 (+1.0)	8.7 (-0.1)	89.9 (+0.2)	96.8 (+2.1)	93.8 (+3.1)	31.7 (+1.9)	92.9 (+0.8)
Sup. pre-train+SPD	✓	**88.6** (+0.4)	**87.8** (+0.0)	**12.0** (+0.6)	**93.8** (+0.7)	**97.5** (+0.1)	**94.6** (+0.1)	**36.3** (+1.1)	**94.6** (+0.2)

we use MVTec-AD dataset [4] as a 1-class training benchmark. MVTec-AD contains 15 sub-datasets with a total of 5,354 images.

Anomaly detection and segmentation algorithms: To evaluate the transfer learning performances of different pre-training, we adopt the following algorithms for anomaly detection and segmentation.

1-class anomaly classification/segmentation: We leverage PaDiM [14] which is one of the top performing 1-class anomaly detection/localization methods.

2-class anomaly classification/segmentation: We train a standard binary ResNet [24] as the supervised model for classification. A U-Net [32] is used as segmentation model. The focal loss [27] is used to overcome the data imbalance.

Implementation details: Unless otherwise noted, we choose ResNet-50 as the major backbone. We adopt exactly the same hyperparameters in SimSiam, MoCo, SimCLR and supervised learning for pre-training. More implementation details are in the supplementary.

5.1 SPD in high-shot 1-class/2-class Regimes

For the 1-class setting, the results of PaDiM with various pre-training options w/wo SPD are shown in Table 2. The results are also visualized as scatter plots in Fig. 7. We have several key observations. First, SPD improves performances of

Table 3. 2-class fine-tuning with different pre-training on VisA high-shot setup.

	ImageNet labels	VisA (2-class, high-shot)			
		Classification		Segmentation	
		AU-PR	AU-ROC	AU-PR	AU-ROC
Sup. pre-train	✓	97.5	99.5	65.1	97.3
SimSiam	✗	88.7	97.9	53.8	97.3
+SPD	✗	93.2 (+4.5)	98.7 (+0.8)	59.7 (+5.9)	98.1 (+0.8)
MoCo	✗	93.9	98.8	62.4	98.0
+SPD	✗	94.2 (+0.3)	98.8 (+0.0)	64.4 (+2.0)	97.9 (-0.1)
SimCLR	✗	93.4	98.5	67.7	95.3
+SPD	✗	92.7 (-0.7)	98.6 (+0.1)	68.2 (+0.5)	95.7 (+0.4)
Sup. pre-train+SPD	✓	**98.3** (+0.8)	**99.7** (+0.2)	**71.9** (+6.8)	**98.5** (+1.2)

both anomaly detection and segmentation across almost all pre-training baselines on both VisA and MVTec-AD. While we report both AU-PR and AU-ROC, the former metric is more relevant to the application and we see that self-supervised methods are improved up to AU-PR of 2.6%. Note both metrics are averaged over the 12 objects in VisA. For different objects, the gains differ and are given in Sec. D of the supplementary. Second, the gap between self-supervised pre-training with SimSiam, SimCLR, MoCo, and supervised pre-training is large. SPD reduces this gap, but no combination of SSL and SPD beats supervised pre-training. This is in contrast to the low-shot regime in Section 5.2, where self-supervision has advantages in some cases. Third, PaDiM is one of the SOTA methods with $>$ 97% AU-ROC in MVTec. But it just achieves $<$ 90% AU-PR and AU-ROC in VisA - classification. For VisA - segmentation, PaDiM only achieves about 10% AU-PR. This shows the difficulty of the VisA 1-class benchmark. Moreover, the gap between low AU-PR and high AU-ROC for both VisA/MVTec segmentation justifies the inflated performance view of AU-ROC, in favor of AU-PR as a more suitable metric in imbalanced datasets. In addition, even in terms of AU-ROC, the SPD consistently improves almost all baselines.

In Table 3, we show the results for the 2-class high-shot regime on the VisA and observe similar trends as above. However, the AU-PR gains from SPD on top of SimSiam and supervised pre-training are higher at 5.9% and 6.8% respectively for segmentation. Another key point to note here is that the AU-ROC metrics are saturating even though AU-PR metrics show room for improvement, particularly for segmentation. This another data point for preferring AU-PR metric. Comparing Tables 2 and 3, there is a significant gap between 1-class and 2-class performance on VisA. As anomalies are harder to obtain compared to normal images, bridging the gap is an open challenge to the research community.

5.2 SPD in Low-shot 2-class Regime

Low-shot anomaly segmentation: With different ImageNet pre-training as initialization, a 2-class U-Net with ResNet-50 encoder is trained for each 5/10-shot segmentation setup. From Table 4, SPD again improves all baselines in both 5-shot and 10-shot evaluation, with AU-PR gain up to 2.3%. One departure from the high-shot regime is that for few-shot anomaly segmentation, MoCo+SPD is the best method, even outperforming supervised pre-training.

Table 4. Low-shot anomaly detection and segmentation on VisA.

	ImageNet labels	Classification (2-class, low-shot)				Segmentation (2-class, low-shot)			
		5-shot		10-shot		5-shot		10-shot	
		AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Sup. pre-train	✓	59.2	85.5	70.4	91.7	17.8	74.6	28.3	81.8
SimSiam	✗	51.9	82.3	65.0	89.4	17.3	75.2	28.5	81.6
+SPD	✗	56.1 (+4.2)	84.0 (+1.7)	67.6 (+2.6)	90.8 (+1.4)	18.2 (+0.9)	**76.0** (+0.8)	29.7 (+1.2)	83.2 (+1.6)
MoCo	✗	56.1	83.8	68.7	90.6	21.5	80.5	32.3	**85.7**
+SPD	✗	56.4 (+0.3)	83.9 (+0.1)	68.0 (-0.7)	90.1 (-0.5)	**22.1** (+0.6)	78.5 (-2.0)	**32.8** (+0.5)	84.9 (-0.8)
SimCLR	✗	48.4	79.6	58.2	86.0	18.4	71.2	23.0	75.1
+SPD	✗	47.4 (-1.0)	79.9 (+0.3)	59.0 (+0.8)	86.1 (+0.1)	18.9 (+0.5)	74.5 (+3.3)	25.1 (+2.1)	78.2 (+3.1)
Sup. pre-train+SPD	✓	**59.8** (+0.6)	**85.9** (+0.4)	**71.2** (+0.8)	**92.1** (+0.4)	18.7 (+0.9)	75.9 (+1.3)	30.6 (+2.3)	81.8 (+0.0)

Table 5. Ablation study

	VisA (1-class)				MVTec-AD (1-class)			
	Classification		Segmentation		Classification		Segmentation	
	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
SimSiam w/ Res50	80.2	78.1	9.1	93.1	92.6	83.9	29.7	92.1
+SPD ($\eta = 0.1$)	82.8	81.2	9.4	92.7	94.1	88.0	32.0	92.2
+SPD ($\eta = 0.5$)	80.5	79.3	8.7	93.0	93.3	84.9	30.1	91.9
+SPD ($\eta = 1.0$)	81.5	79.8	9.4	92.8	93.4	85.8	30.0	92.0
+SPD w/ CutPaste	78.8	77.0	9.7	93.1	93.5	85.2	28.2	91.3
+SPD w/ Xent	71.4	66.6	2.7	84.8	86.3	71.0	15.2	82.6
SimSiam w/ WideRes50	80.3	77.7	9.9	93.6	93.0	84.7	31.3	92.2
+SPD	81.9	80.4	10.5	93.7	93.4	85.4	32.5	92.8

Table 6. 1-class performance evaluation on VisA and MVTec-AD with PatchCore.

Backbone: Wide ResNet50	VisA (1-class)				MVTec-AD (1-class)			
	Classification		Segmentation		Classification		Segmentation	
	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC	AU-PR	AU-ROC
Sup. pre-train	93.3	92.4	38.4	98.4	99.2	99.8	48.8	97.6
Sup. pre-train+SPD	93.8 (+0.5)	92.5 (+0.1)	39.3 (+0.9)	98.1 (-0.3)	99.0 (-0.2)	99.7 (-0.1)	49.3 (+0.5)	97.5 (-0.1)

Low-shot anomaly detection: Initialized with different ImageNet pre-training, a 2-class ResNet-50 is trained in 5/10-shot setups for anomaly detection. From Table 4, overall the supervised pre-training with SPD outperforms both supervised pre-training only and other SSL's. Moreover, SPD significantly improves SimSiam with 4.2% AU-PR in 5-shot and 2.6% AU-PR in 10-shot, although it's still inferior to supervised pre-training.

5.3 Ablation Study

We conduct extensive ablation studies based on ImageNet SimSiam pre-training and PaDiM as the anomaly detection and segmentation algorithms trained in the 1-class setups of VisA and MVTec-AD. Results are shown in Table 5.

Sensitivity analysis on SPD loss weight η: From Table 5, we see consistent improvement for $\eta = 0.1, 0.5, 1.0$ in at least one task for both datasets. SPD loss with $\eta = 0.1$ gives us the best performances in both datasets, which is chosen as the default SPD loss weight for all pre-training with SPD. So the SimSiam+SPD ($\eta = 0.1$) is regarded as SimSiam+SPD for better clarity.

Comparison between SPD and CutPaste [26]: CutPaste and cross-entropy loss used in [26] for anomaly detection training can also be used in ImageNet pre-training. An ablation study is done to demonstrate the superiority of the proposed SmoothBlend and SPD loss. With $\mathcal{L}_{\text{SPD}}$, SmoothBlend is arguably

better than CutPaste by 4.0% and 3.8% AU-PR improvement in VisA - classification and MVTec - segmentation (+SPD v.s. +SPD w/ CutPaste). With the SmoothBlend, the SPD loss significantly outperforms cross-entropy loss (+SPD v.s. +SPD w/ Xent). Such results demonstrate the validity of proposed methods.

SPD with different backbones: ResNet-50 is adopted as the backbone for all major experiments in this paper. We demonstrate the SPD can generalize to different network architectures by experiments of SimSiam w/wo SPD on wide ResNet-50 [45]. As in Table 5, SPD still improves the baseline.

Results with PatchCore: In addition to PaDiM, we also evaluate supervised pre-trained models based on another state-of-the-art 1-class method PatchCore [33]. Wide ResNet-50 is chosen as the backbone network. As in Table 6, on VisA, SPD improves supervised pre-trained model by 0.5% and 0.9% AU-PR for both classification and segmentation. On MVTec-AD, SPD improves by 0.5% AU-PR for segmentation with slightly performance decreased in classification.

Extending SPD to other tasks: Besides improvement on defect detection and segmentation, SPD also improves ImageNet supervised classification accuracy: 69.8% $\rightarrow$ 70.2% for ResNet-18 and 76.1% $\rightarrow$ 76.4% for ResNet-50. Pre-trained models with better ImageNet accuracy are expected to benefit downstream tasks more. Thus we speculate that SPD will work well for object recognition and detection, especially on fine-grained classification and small object detection as SPD promotes local sensitivity. In addition, we will leverage the proposed SPD training as a 1-class anomaly detection model to be trained by downstream data.

Qualitative results: To qualitatively demonstrate the effectiveness of SPD regularization, we present attention maps of anomalous samples and anomaly segmentation results in Sec. E of the supplementary due to page limits.

6 Conclusions

In this work, we present a spot-the-difference (SPD) training to regularize pre-trained models' local sensitivity to anomalous patterns. We also present a novel Visual Anomaly (VisA) dataset which is the largest industrial anomaly detection dataset. Extensive experiments demonstrate the benefits of SPD for various contrastive self-supervised and supervised pre-training for anomaly detection and segmentation. Compared to standard supervised pre-training, SimSiam with SPD obtains superior or competitive performances in low-shot regime while supervised learning with SPD presents better performances in various setups.

Acknowledgments

The authors would like to thank Fanyi Xiao, Erhan Bas, Aditya Deshpande and Joachim Stahl for idea brainstorming and providing insightful comments on the manuscript.

References

1. AWS Lookout for Vision. https://aws.amazon.com/lookout-for-vision/
2. Google Visual Inspection AI. https://cloud.google.com/solutions/visual-inspection-ai
3. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C.: The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. International Journal of Computer Vision **129**(4), 1038–1059 (2021)
4. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: MVTec AD–a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9592–9600 (2019)
5. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4183–4192 (2020)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems **33**, 9912–9924 (2020)
7. Chen, K., Hong, L., Xu, H., Li, Z., Yeung, D.Y.: MultiSiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7546–7554 (2021)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
9. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
10. Cohen, N., Hoshen, Y.: Sub-image anomaly detection with deep pyramid correspondences. arXiv preprint arXiv:2005.02357 (2020)
11. Cook, J., Ramadas, V.: When to consult precision-recall curves. The Stata Journal **20**(1), 131–148 (2020)
12. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9268–9277 (2019)
13. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240 (2006)
14. Defard, T., Setkov, A., Loesch, A., Audigier, R.: PaDim: a patch distribution modeling framework for anomaly detection and localization. In: International Conference on Pattern Recognition. pp. 475–489. Springer (2021)
15. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9737–9746 (June 2022)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

17. Ericsson, L., Gouk, H., Loy, C.C., Hospedales, T.M.: Self-supervised representation learning: Introduction, advances and challenges. arXiv preprint arXiv:2110.09327 (2021)
18. Feng, T., Qi, Q., Wang, J., Liao, J.: Few-shot class-adaptive anomaly detection with model-agnostic meta-learning. In: 2021 IFIP Networking Conference (IFIP Networking). pp. 1–9. IEEE (2021)
19. Geirhos, R., Narayanappa, K., Mitzkus, B., Bethge, M., Wichmann, F.A., Brendel, W.: On the surprising similarities between supervised and self-supervised models. In: NeurIPS 2020 Workshop SVRHM (2020), https://openreview.net/forum?id=q2ml4CJMHAx
20. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (2018)
21. Görnitz, N., Kloft, M., Rieck, K., Brefeld, U.: Toward supervised anomaly detection. Journal of Artificial Intelligence Research **46**, 235–262 (2013)
22. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
23. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
25. Jhamtani, H., Berg-Kirkpatrick, T.: Learning to describe differences between pairs of similar images. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2018)
26. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: CutPaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9664–9674 (2021)
27. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
28. Liu, S., Li, Z., Sun, J.: Self-EMD: Self-supervised object detection without ImageNet. arXiv preprint arXiv:2011.13677 (2020)
29. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016)
30. Reiss, T., Hoshen, Y.: Mean-shifted contrastive loss for anomaly detection. arXiv preprint arXiv:2106.03844 (2021)
31. Ristea, N.C., Madan, N., Ionescu, R.T., Nasrollahi, K., Khan, F.S., Moeslund, T.B., Shah, M.: Self-supervised predictive convolutional attentive block for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
32. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
33. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2022)

34. Rudolph, M., Wehrbein, T., Rosenhahn, B., Wandt, B.: Fully convolutional cross-scale-flows for image-based defect detection. In: Winter Conference on Applications of Computer Vision (WACV) (2022)
35. Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., Müller, K.R.: A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE **109**(5), 756–795 (2021)
36. Ruff, L., Vandermeulen, R.A., Görnitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: Proceedings of the 35th International Conference on Machine Learning. vol. 80, pp. 4393–4402 (2018)
37. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PloS one **10**(3), e0118432 (2015)
38. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
39. Sohn, K., Li, C.L., Yoon, J., Jin, M., Pfister, T.: Learning and evaluating representations for deep one-class classification. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=HCSgyPUfeDj
40. Tack, J., Mo, S., Jeong, J., Shin, J.: CSI: Novelty detection via contrastive learning on distributionally shifted instances. Advances in neural information processing systems **33**, 11839–11852 (2020)
41. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3024–3033 (June 2021)
42. Xie, E., Ding, J., Wang, W., Zhan, X., Xu, H., Sun, P., Li, Z., Luo, P.: DetCo: Unsupervised contrastive learning for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8392–8401 (2021)
43. Yang, C., Wu, Z., Zhou, B., Lin, S.: Instance localization for self-supervised detection pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3987–3996 (June 2021)
44. Yi, J., Yoon, S.: Patch SVDD: Patch-level svdd for anomaly detection and segmentation. In: Proceedings of the Asian Conference on Computer Vision (2020)
45. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: British Machine Vision Conference 2016. British Machine Vision Association (2016)
46. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision. pp. 649–666. Springer (2016)