Sub-Image Anomaly Detection with Deep Pyramid Correspondences

Niv Cohen and Yedid Hoshen

School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel.
{niv.cohen2,yedid.hoshen}@mail.huji.ac.il

预训练+KNN，缺失语义特征定位

Abstract. Nearest neighbor (kNN) methods utilizing deep pre-trained features exhibit very strong anomaly detection performance when applied to entire images. A limitation of kNN methods is the lack of segmentation map describing where the anomaly lies inside the image. In this work we present a novel anomaly segmentation approach based on alignment between an anomalous image and a constant number of the similar normal images. Our method, Semantic Pyramid Anomaly Detection (SPADE) uses correspondences based on a multi-resolution feature pyramid. SPADE is shown to achieve state-of-the-art performance on unsupervised anomaly detection and localization while requiring virtually no training time.

对齐

多分辨率特征金字塔

Keywords: anomaly detection, nearest-neighbors, feature pyramid

相对于图片，特征语义信息，细粒度和全文信息具有对齐作用

1 Introduction

Humans observe many images throughout their lifetimes, most of which are of little interest. Occasionally, an image indicating an opportunity or danger appears. A key human ability is to detect the novel images that deviate from previous patterns triggering particular vigilance on the part of the human agent. Due to the importance of this function, allowing computers to detect anomalies is a key task for artificial intelligence.

As a motivational example, let us consider the setting of an assembly-line fault detection. Assembly lines manufacture many instances of a particular product. Most products are normal and fault-free. Unfortunately, on isolated occasions, the manufactured products contain some faults e.g. dents, wrong labels or part duplication. As reputable manufacturers strive to keep a consistent quality of products, prompt detection of the faulty products is very valuable. As mentioned earlier, humans are quite adept at anomaly detection, however having a human operator oversee every product manufactured by the assembly line has several key limitations: i) high wages earned by skilled human operators ii) limited human attention span ([14] states this can be as low as 20 minutes!) iii) a human operator cannot be replicated between different assembly lines. iv) different operators typically do not maintain a consistent quality level. Anomaly detection therefore calls for computer vision solutions.

人工检测的缺陷

Although visual anomaly detection is very valuable, it is also quite challenging. One challenge common to all anomaly detection methods is the unexpectedness of anomalies. Typically in supervised classification, test classes come from the a similar distribution to the train data. In most anomaly detection settings, the distribution of anomalies is not observed during training time. Different anomaly detection methods differ by the way the anomalies are observed at training time. In this paper, we deal with the setting where at training time only normal data (but no anomalies) are observed. This is a practically useful setting, as obtaining normal data (e.g. products that contain no faults) is usually easy. This setting is sometimes called semi-supervised ([7]). As this notation is ambiguous, we shall refer to this setting as the normal-only training setting. An easier scenario is fully supervised i.e. both normal and anomalous examples are presented with labels during training. As this training setting is similar to standard supervised classification, a mature task with effective solutions, it will not be dealt with in this work.

Another challenge particular to visual anomaly detection (rather than non-image anomaly detection methods) is the localization of anomalies i.e. segmenting the parts of the image which the algorithm deems anomalous. This is very important for the explainability of the decision made by the algorithm as well as for building trust between operators and novel AI systems. It is particularly important for anomaly detection, as the objective is to detect novel changes not seen before which humans might not be familiar with. In this case, the computer may teach the human operator of the existence of new anomalies or alternatively the human may decide that an anomaly is not of interest thus not rejecting the product, resulting in cost-saving

We present a new method for solving the task of sub-image anomaly detection and segmentation. Our method does not require an extended training stage, it is fast, robust and achieves state of the art performance. Out methods consists of several stages: i) image feature extraction using a pre-trained deep neural network (e.g. an ImageNet trained ResNet) ii) nearest neighbor retrieval of the nearest K normal images to the target iii) finding dense pixel-level correspondence between the target and the normal images, target image regions that do not have near matches in the retrieved normal images are labeled as anomalous. Our method is extensively evaluated on an industrial product dataset (MVTech) as well as a surveillance dataset in a campus setting (Shanghai Tech Campus). Our method achieves state-of-the-art performance both on image-level and pixel-level anomaly detection.

1. 预训练网络提取特征
2. KNN找到K个最近正样本图片
3. 目标和正常样本像素级对应关系，不匹配地方认为异常。

2 Previous Work

Anomaly detection has attracted a large body of work over the last several decade[illegible] of image-level and sub-image anomaly detection metho[illegible]

图片级方法：这些方法不是专门为分割图像中的异常而设计的

Im[illegible]iew methods that detect if an image is anomalous that are not particularly designed for segmenting the anomaly within the image.

这些方法都不是专门作用于图像级别，也可以用于像素

Note that many of these methods are not specific to images. There are three main classes of methods for image-level anomaly detection: reconstruction-based, distribution-based and classification-based.

基于重构的方法是在正常样本数据上学习一系列的基础函数，然后去重构测试图片。如果测试图片不能被很好地重构，表明它是异常。

Reconstruction-based methods learn a se[illegible] ing data, and attempt to reconstruct the test [illegible] basis functions. If the test image cannot be faithfully reconstructed using the basis functions, it is denoted as anomalous, as it is likely that it came from a different basis from that of the normal training data. Different methods vary in terms of the set of basis functions and loss functions they use. Popular choices of basis functions include: K-means [15], K nearest neighbors (kNN) [9], principal component analysis (PCA) [20]. The loss functio[illegible] vector metrics such as Euclidean or L_1 losses an[illegible] ceptual losses such as structural similarity (SSIM) [illegible] methods have broadened the toolbox of reconstruct[illegible] components ha[illegible]ctions learned by autoencoders [27], including [illegible]al autoencoders (VAEs). Deep perceptual loss[illegible]ove over traditional perceptual loss [illegible]struction-based loss functions are: [illegible]used for evaluating the quality of re[illegible]n non-obvious and hurting performance ii) determining the correct functional basis.

不同的方法基础函数(K-means、KNN、PCA、AE、VAEs)和损失函数（Euclidean数、L1 损失、SSIM、深度感知损失函数)不同。

两个主要的[illegible]失函数评价[illegible]失函数不降[illegible]基础函数。

基于分布的方法：对正常数据分布的概率密度函数（PDF）进行建模，使用PDF对测试样本进行评估，并将具有低概率密度值的测试样本指定为异常。

The second class of methods is distribution-based. The main principle is to model the probability density function (PDF) of the distribution of the normal data. Test samples are evaluated using the PDF, and test samples with low probability density values are designated as anomalous. Different distribution-based methods differ by the distributional assumptions that they make, the approximations used to estimate the true PDF, and by the training procedure. Parametric methods include Gaussian or mixture of Gaussians (GMM). Kernel density estimation [21] is a notable non-parametric method. Nearest neighbors [9] can also be seen as a distributional (as it performs density estimation), but note that we also designated it a[illegible]ethod. Recently deep learning methods have improved [illegible] by mapping high-dimensional data distributions into a[illegible] PD[illegible] easier in lower dimensional spaces. Learning the deep pro[illegible]l modeling can be done jointly as done by [36]. Another recent development, adversarial training, was also applied to anomaly detection e.g. ADGAN [8]. Although in principle distributional-methods are very promising, they suffer from some critical drawbacks: i) real image data rarely follows simple parametric distributional assumptions ii) non-paramet[illegible] require large training set tha[illegible]

深度学习映射高维数据分布到低纬密度空间，在低维空间上估计PDF.

对抗训练的原因？训练分布吗？

两个缺点：1.真实样本的分布复杂；2.非参数方法需要很高的复杂度，因为需要很大的数据集，而这在实践中通常是不可用的

Recently, classification-based methods have achieved dominance for image-level anomaly detection. One such paradigm is one-class support vector machines (OC-SVM) [28]. One of its most successful variants is support vector data description (SVDD) [30] which can be seen as a finding the minimal sphere which contains at least a given fraction of the data. These methods are very sensitive to

基于分类的方法

传统方法依赖核函数、深度学习依赖特征提取

自监督方法

the feature space used giving rise to both kernel methods as well as deep methods [26] for learning features. Another set of methods is based on self-supervised learning. Golan and El-Yaniv [11] proposed a RotNet-based [10] approach, which performs geometric transformations on the input data and trains a network that attempts to recognize the transformation used. They use the idea that the trained classifier will generalize well to new normal images but not to anomalous images allowing it to be used as an anomaly detection criterion. Hendrycks et al. [18] improved the architecture and training procedure achieving strong performance. Bergman and Hoshen [4] combined this work with an SVDD type criterion and extended it to non-image data. Very recently Bergman et al. [3] showed that the features learned using such self-supervised methods are not competitive with generic ImageNet-based feature extractors. A simple method based on kNN (or efficient approximations) significantly outperfo

自监督学习弱于迁移学习：简单的KNN优于自监督。

ods.

Sub-image methods: The methods previously described tackled the task of classifying a whole image as normal or anomalous, and most of the techniques were not specific to images. The task of segmenting the particular pixels containing anomalies is special to images and has achieved far less attention from the deep learning community. Napoletano et al. [24] extracted deep features from small overlapping patches, and used a K-means based classifier over dimensionality reduced features. Bergmann et al. [5] evaluated both a ADGAN and autoencoder approaches on MVTech finding complementary strengths. More recently, Venkataramanan et al. [31] used an attention-guided VAE approach combining multiple methods (GAN loss [13], GRADCAM [29]). Bergmann et al. [6] used a student-teacher based autoencoder approach employing pre-trained ImageNet deep features (still requiring an expensive training stage). In this work, we present a novel sub-image alignment approach which is more accurate, faster, more stable than previous methods and does not require a dedicated training stage. To support research on sub-image anomaly detection, high quality datasets for evaluating this task have been introduced, such as: MVTech [5] - a dataset simulating an industrial fault detection where the objective is to detect parts of images of products that contain faults such as dents or missing parts. The ShanghaiTech Campus dataset [23] - a dataset simulating a surveillance setting where cameras observe a busy campus and the objective is to detect anomalous objects and activities such as fights. Hendrycks et al. [17] also proposed a new dataset containing anomalies such as road hazards. We evaluate our work on the two most used datasets, MVTech and ShanghaiTech Campus (STC).

3 Correspondence-based Sub-Image Anomaly Detection

We present our method for sub-image anomaly detection. Our method consists of several parts: i) image feature extraction ii) K nearest neighbor normal image retrieval iii) pixel alignment with deep feature pyramid correspondences.

3.1 Feature Extraction

自监督提取特征，但数据量少，不利于高质量的相似度测量。Bergman证明预训练ImageNet网络优于自监督网络。

The first stage of our method is the extraction of strong image level features. The same features are later used for pixel-level image alignment. There are multiple options for extracting features. The most commonly used option is self-supervised feature learning, that is, learning features from scratch directly on the input normal images. Although it is an attractive option, it is not obvious that the features learned on small training datasets will indeed be sufficient for serving as high-quality similarity measures. The analysis performed in Bergman et al. [4] illustrates that self-supervised features underperform ImageNet-trained ResNet features for the purposes of anomaly detection. We therefore used a ResNet feature extractor pre-trained on the ImageNet dataset. As image-level features we used the feature vector obtained after global-pooling the last convolutional layer. Let us denote the global feature extractor F, for a given image x_i, we denote the extracted features f_i:

最后一个卷积层后的全局池化特征

$$f_i = F(x_i) \tag{1}$$

At initialization, the features for all training images (which are all normal) are computed and stored. At inference, only the features of the target image are extracted.

计算所有正常样本的距离，排序，前K个距离求和。

K邻近如何检索正常图片？

3.2 K Nearest Neighbor Normal Image Retrieval

The first stage in our method is determining which images contain anomalies using DN2 [4]. For a given test image y, we retrieve its K nearest normal images from the training set, $N_k(f_y)$. The distance is measured using the Euclidean metric between the image-level feature representations.

$$d(y) = \frac{1}{K} \sum_{f \in N_K(f_y)} \|f - f_y\|^2 \tag{2}$$

Images are labelled at this stage as normal or anomalous. Positive classification is determined by verifying if the kNN distance is larger than a threshold τ. It is expected that most images are normal, and only few images are designated as anomalous.

KNN距离与阈值比较

如何图片对齐？

3.3 Sub-image Anomaly Detection via Image Alignment

After being labelled as anomalous at the image-level stage, the objective is to locate and segment the pixels of one or multiple anomalies. In the case that the image was falsely classified as anomalous, our objective would be to mark no pixels as anomalous.

As a motivational idea, let us consider aligning the test image to a retrieved normal image. By finding the differences between the test and normal image, we would hope to detect the anomalous pixels. This naive method has several flaws i) assume that there are multiple normal parts the object may possibly

1. 假设物体可能由多个正常部分组成，对齐整个图片可能失败

2.小数据集或者物体复杂变化，可能不能找到一个相似的训练图片去检测预测图片，造成很高的假阳率

consist of, alignment to particular normal images may fail ii) for small datasets or objects that experience complex variation, we may never in fact find a normal training image which is similar to the test image in every respect triggering false positive detections iii) computing the image difference would be very sensitive to the loss function being used.

3.计算图像差异将对所使用的损失函数非常敏感

To overcome the above issues, we present a multi-image correspondence method. We extract deep features at every pixel location $p \in P$ using feature extractor $F(x_i, p)$ of the relevant test and normal training images. The details of the feature extractor will be described in Sec. 3.4. We construct a gallery of features at all pixel locations of the K nearest neighbors $G = \{F(x_1, p) | p \in P\} \cup \{F(x_2, p) | p \in P\} .. \cup \{F(x_K, p) | p \in P\}$. The anomaly score at pixel p, is given by the average distance between the features $F(y, p)$ and its κ nearest features from the gallery G. The anomaly score of pixel p in target image y is therefore given by:

异常得分：特征与集合中k邻近特征的平均距离

一个集合：K个邻近特征的所有像素位置特征

$$d(y, p) = \frac{1}{\kappa} \sum_{f \in N_\kappa(F(y,p))} \|f - F(y, p)\|^2 \tag{3}$$

异常得分大于0表示异常

For a given threshold θ, a pixel is determined as anomalous if $d(y, p) > \theta$, that is, if we cannot find a closely corresponding pixel in the K nearest neighbor normal images.

3.4 Feature Pyramid Matching

如何通过密集的对应关系进行对齐？

Alignment by dense correspondences is an effective way of determining the parts of the image that are normal vs. those that are anomalous. In order to perform the alignment effectively, it is necessary to determine the features for matching. As in the previous stage, our method uses features from a pre-trained deep ResNet CNN. The ResNet results in a pyramid of features. Similarly to image pyramids, earlier layers (levels) result in higher resolution features encoding less context. Later layers encode lower resolution features which encode more context but at lower spatial resolution. To perform effective alignment, we describe each location using features from the different levels of the feature pyramid. Specifically, we concatenate features from the output of the last M blocks, the results for different numbers of M is shown in the experimental section. Our features encode both fine-grained local features and global context. This allows us to find correspondences between the target image and $K \geq 1$ normal images, rather than having to explicitly align the images, which is more technically challenging and brittle. Our method is scalable and easy to deploy in practice. We will show in Sec. 4 that our method achieves the state-of-the-art sub-image anomaly segmentation accuracy.

细粒度的局部特征和全局上下文

M块特征拼接

3.5 Implementation Details

In all experiments, we use a Wide-ResNet50 $\times$ 2 feature extractor, which was pre-trained on ImageNet. MVTec images were resized to 256×256 and cropped

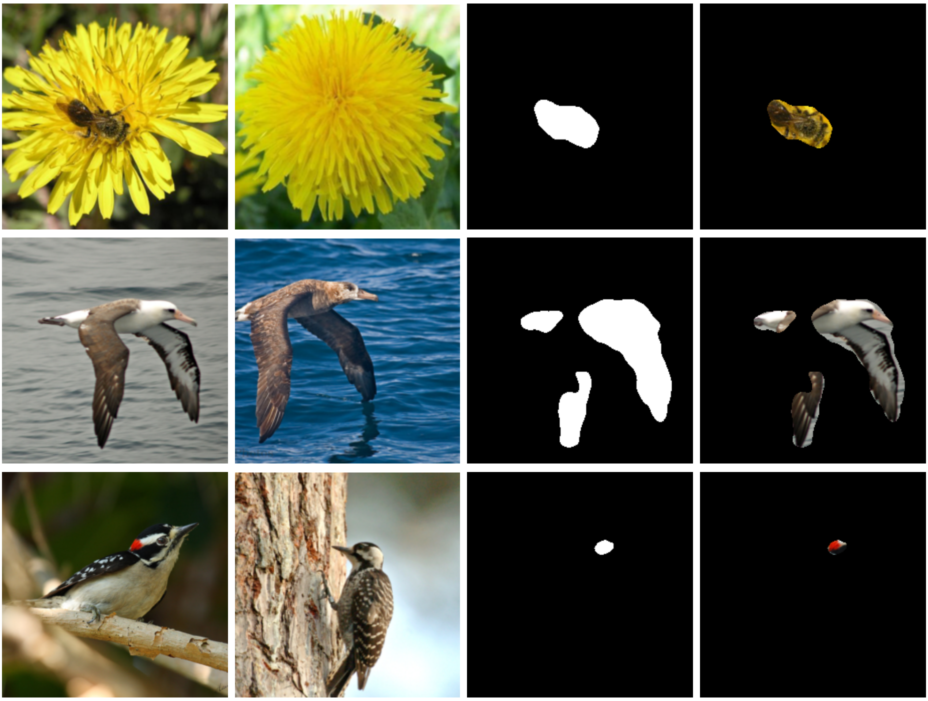

Fig. 1. An evaluation of SPADE on detecting anomalies between flowers with or without insects (taken from one category of 102 Category Flower Dataset [25]) and bird varieties (taken from Caltech-UCSD Birds 200) [33]. (left to right) i) An anomalous image ii) The retrieved top normal neighbor image iii) The mask detected by SPADE iv) The predicted anomalous image pixels. SPADE was able to detect the insect on the anomalous flower (top), the white colors of the anomalous albatross (center) and the red spot on the anomalous bird (bottom).

to 224×224. STC images were resized to 256×256. Due to the large size of STC dataset, we subsampled its training data to roughly 5000 images. To be comparable with [31], we subsampled the STC test set by a factor of 5. All metrics were calculated at 256×256 image resolution, and we used cv2.INTERAREA for resizing when needed. Unless otherwise specified, we used features from the ResNet at the end of the first block (56×56), second block (28×28) and third block (14×14), all with equal weights. We used $K = 50$ nearest neighbours for the MVTtec experiments and $K = 1$ nearest neighbours for the STC experiments (due to the runtime considerations). In all experiments we used $\kappa = 1$.

After achieving the pixel-wise anomaly score for each image, we used smoothed the results with a Gaussian filter ($\sigma = 4$).

4 Experiments

We perform an extensive evaluation of our method against the state-of-the-art in sub-image anomaly detection.

4.1 MVTec

To simulate anomaly detection in industrial settings, [5] Bregmann et al. presented a dataset consisting of images from 15 different classes. 5 classes consist of textures such as wood or leather. The other 10 classes contain objects (mostly rigid). For each class, the training set is composed of normal images. The test set is composed of normal images as well as images containing different types of anomalies. This dataset therefore follows the standard protocol where no anomalous images are used in training. The anomalies in this dataset are more finegrained than those typically used in the literature e.g. in CIFAR10 evaluation, where anomalous images come from a completely different image category. Instead, anomalies take the form of a slightly scratched hazelnut or a lightly bent cable. As the anomalies are at the sub-image level, i.e. only affect a part of the image, the dataset provides segmentation maps indicating the precise pixel positions of the anomalous regions.

An example of the operation of our method on the MVTec dataset can be observed in Fig. 2. The anomalous object (a hazelnut) contain a scratch. The retrieved nearest neighbor normal image, contains a complete nut without scratches. By search for correspondences between the two images, our method is able to find correspondences for the normal image regions but not for the anomalous region. This results in an accurate detection of the anomalous region of the image. The anomalous images pixels are presented on the right-most image.

We compared our method against several methods that were introduced over the last several months, as well as longer standing baseline such as OCSVM and nearest neighbors. For each setting, we compared against the methods that reported the suitable metric.

We first compare the quality of deep nearest neighbor matching as a means for finding anomalous images. This is computed by the distance between the test image and the nearest neighbor normal images. Larger distances indicate more anomalous images. We compared the ROC area under the curve (ROCAUC) of the first step of our method and other state-of-the-art methods for image level anomaly detection. We report the average ROCAUC across the 15 classes. Please note that the first stage of our method is identical with DN2 [3]. This comparison is important as it verifies if deep nearest neighbors are effective on these dataset. The comparison is presented in Tab. 1. Our method is shown to outperform a range of state-of-the-art methods utilizing a range of self-supervised anomaly detection learning techniques. This gives evidence that deep features trained on ImageNet (which is very different from ImageNet dataset) are very effective even on datasets that are quite different from ImageNet.

We proceed to evaluate our method on the task of pixel-level anomaly detection. The objective here is to segment the particular pixels that contain anoma-

Table 1. Image-level anomaly detection accuracy on MVTec (Average ROCAUC %)

	Geom [11]	GANomaly [1]	AE_{L2}	ITAE [19]	SPADE
Average	67.2	76.2	75.4	83.9	**85.5**

Table 2. Sub-Image anomaly detection accuracy on MVTec (ROCAUC %)

	AE_{SSIM}	AE_{L2}	AnoGAN	CNN Dict	TI	VM	CAVGA-R_u	SPADE
Carpet	87	59	54	72	88	-	-	97.5
Grid	94	90	58	59	72	-	-	93.7
Leather	78	75	64	87	97	-	-	97.6
Tile	59	51	50	93	41	-	-	87.4
Wood	73	73	62	91	78	-	-	88.5
Bottle	93	86	86	78	-	82	-	98.4
Cable	82	86	78	79	-	-	-	97.2
Capsule	94	88	84	84	-	76	-	99.0
Hazelnut	97	95	87	72	-	-	-	99.1
Metal nut	89	86	76	82	-	60	-	98.1
Pill	91	85	87	68	-	83	-	96.5
Screw	96	96	80	87	-	94	-	98.9
Toothbrush	92	93	90	77		68	-	97.9
Transistor	90	86	80	66	-	-	-	94.1
Zipper	88	77	78	76	-	-	-	96.5
Average	87	82	74	78	75	77	89	**96.0**

lies. We evaluate our method using two established metrics. The first is per-pixel ROCAUC. This metric is calculated by scoring each pixel by the distance to its K nearest correspondences. By scanning over the range of thresholds, we can compute the pixel-level ROCAUC curve. The anomalous category is designated as positive. It was noted by several previous works that ROCAUC is biased in favor of large anomalies. In order to reduce this bias, Bergmann et al [6] propose the PRO (per-region overlap) curve metric. They first separate anomaly masks into their connected components, therefore dividing them into individual anomaly regions. By changing the detection threshold, they scan over false positive rates (FPR), for each FPR they compute PRO i.e. the proportion of the pixels of each region that are detected as anomalous. The PRO score at this FPR is the average coverage across all regions. The PRO curve metric computes the integral across FPR rates from 0 to 0.3. The PRO score is the normalized value of this integral.

In Tab. 2, we compare our methods on the per-pixel ROCAUC metric against state-of-the-art results reported by Bergmann et al. [5] as well as newer results by Venkataramanan et al. [31]. Most of the methods use different varieties of autoencoders, including the top-performer CAVGA-R_u. Our method significantly outperforms all methods. This attest to the strength of our pyramid based correspondence approach.

Table 3. Sub-Image anomaly detection accuracy on MVTec (PRO %)

	Student	1-NN	OC-SVM	ℓ_2-AE	VAE	SSIM-AE	CNN-Dict	SPADE
Carpet	69.5	51.2	35.5	45.6	50.1	64.7	46.9	94.7
Grid	81.9	22.8	12.5	58.2	22.4	84.9	18.3	86.7
Leather	81.9	44.6	30.6	81.9	63.5	56.1	64.1	97.2
Tile	91.2	82.2	72.2	89.7	87.0	17.5	79.7	75.9
Wood	72.5	50.2	33.6	72.7	62.8	60.5	62.1	87.4
Bottle	91.8	89.8	85.0	91.0	89.7	83.4	74.2	95.5
Cable	86.5	80.6	43.1	82.5	65.4	47.8	55.8	90.9
Capsule	91.6	63.1	55.4	86.2	52.6	86.0	30.6	93.7
Hazelnut	93.7	86.1	61.6	91.7	87.8	91.6	84.4	95.4
Metal nut	89.5	70.5	31.9	83.0	57.6	60.3	35.8	94.4
Pill	93.5	72.5	54.4	89.3	76.9	83.0	46.0	94.6
Screw	92.8	60.4	64.4	75.4	55.9	88.7	27.7	96.0
Toothbrush	86.3	67.5	53.8	82.2	69.3	78.4	15.1	93.5
Transistor	70.1	68.0	49.6	72.8	62.6	72.5	62.8	87.4
Zipper	93.3	51.2	35.5	83.9	54.9	66.5	70.3	92.6
Average	85.7	64	47.9	79	63.9	69.4	51.5	**91.7**

Table 4. Image-level anomaly detection accuracy on STC (Average ROCAUC %)

TSC [23]	StackRNN [23]	AE-Conv3D [35]	MemAE [12]	AE(2D) [16]	ITAE [19]	SPADE
67.9	68.0	69.7	71.2	60.9	**72.5**	71.9

In Tab. 3, we compare our method in terms of PRO. As explained above, this is another per-pixel accuracy measure which gives larger weight to anomalies which cover few pixels. Our method is compared with the auto-encoder with pre-trained features based approach of Bregmann et al. [6] and the baselines presented in their paper. Our approach achieves significantly better results than all previous methods. We note than Bregmann et al also presented an ensemble approach with better results. While our method does not use ensembles (which will probably improve our method too), we outperform the ensemble approach as well. We present more qualitative results of our method in Fig. 1 that show that our method is able to recover accurate masks of the anomalous regions.

4.2 Shanghai Tech Campus Dataset

We evaluate our method on the Shanghai Tech Campus dataset. It simulates a surveillance setting, where the input consists of videos captured by surveillance cameras observing a busy campus. The dataset contains 12 scenes, each scene consists of training videos and a smaller number of test images. The training videos do not contain anomalies while the test videos contain normal and anomalous images. Anomalies are defined as pedestrians performing non-standard ac-

Fig. 2. (left to right) i) An anomalous image ii) The retrieved top normal neighbor image iii) The mask detected by SPADE iv) The predicted anomalous image pixels. We can see how in this example, SPADE detects the anomalous image region by finding the correspondence with the nearest-neighbor image. The anomalous parts did not have correspondences in the normal image and were therefore detected.

Table 5. Pixel-level anomaly detection accuracy on STC (Average ROCAUC %)

AE_{L2}	AE_{SSIM}	CAVGA-R_u [31]	SPADE
74	76	85	**89.9**

tivities (e.g. fighting) as well as any moving object which is not a pedestrian (e.g. motorbikes).

We began by evaluating our first stage for detecting image-level anomalies against other state-of-the-art methods. We show in Tab. 4 that our first stage has comparable performance to the top performing method [19]. More interestingly, we compare in Tab. 5 the pixel-level ROCAUC performance with the best reported method, CAVGA-R_u [31]. Our method significantly outperforms the best reported method by a significant margin. Note that we compared to the best method that did not use anomaly supervision, as we do not use it and as anomaly supervision is often not available in practice.

4.3 Ablation Study

We conduct an ablation study on our method in order to understand the relative performance of its different parts. In Tab. 6, we compare using different level of the feature pyramid. We experienced that using activations of too high resolution (56×56) significantly hurts performance (due to limited context) while using

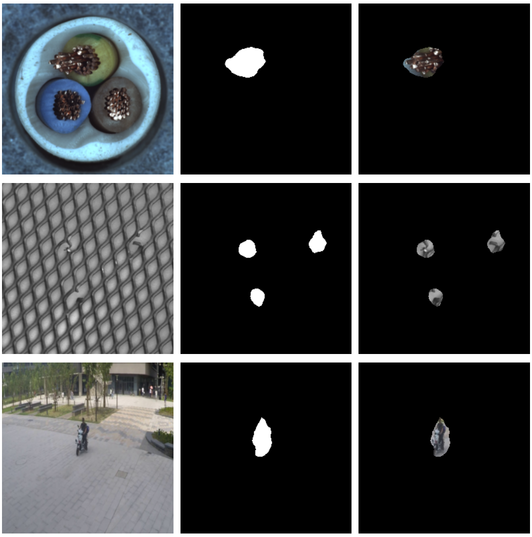

Fig. 3. (top rows) Anomaly detection on images from the Cable and Grid categories of the MVTec dataset (bottom) Detecting bike anomaly on the STC dataset.

the higher levels on their own results in diminished performance (due to lower resolution). Using a combination of all features in the pyramid results in the best performance. In Tab. 7, we compared using the top K neighboring normal images as performed by our first stage vs. choosing them randomly from the dataset. We observe that choosing the kNN images improves performance. This does not affect all classes equally. As an example, we report the numbers for the class "Grid" which has much variation between images. For this category, using the kNN images results in much better performance than randomly choosing K images.

5 Discussion

Anomaly detection via alignment: Most current sub-image anomaly detection methods take the approach of learning a large parametric function for auto-encoding images, making the assumption that anomalous regions will not be reconstructed well. Although this approach does achieve some success, we take a much simpler approach. Similarly to image alignment methods and differently from other sub-image anomaly detection methods, our method does not require feature training and can work on very small datasets. A difference between our method and standard image alignment is that we find correspondences between the target image and parts of K normal images, as opposed to an entire single normal image in simple alignment approaches. The connection with alignment methods, can help in speeding up our method e.g. by combining it with the

Table 6. Pyramid level ablation for sub-image anomaly detection accuracy on MVTec (PRO %)

Used layers size:	(14)	(28)	(56)	SPADE
Carpet	93.5	93.4	91.0	94.7
Grid	80.9	88.0	89.1	86.7
Leather	96.6	97.5	97.3	97.2
Tile	74.5	65.9	73.8	75.9
Wood	84.7	87.7	87.5	87.4
Bottle	93.7	94.7	88.3	95.5
Cable	89.3	87.3	73.5	90.9
Capsule	90.5	92.8	91.4	93.7
Hazelnut	92.7	95.8	96.2	95.4
Metal nut	91.3	93.1	86.1	94.4
Pill	89.2	94.4	96.3	94.6
Screw	90.7	95.9	96.1	96.0
Toothbrush	90.9	93.5	94.5	93.5
Transistor	91.3	72.1	62.5	87.4
Zipper	90.9	92.4	92.5	92.6
Average	89.38	89.6	87.74	**91.7**

Table 7. Evaluating the effectiveness of our kNN retrieval state. We use here 10 nearest neighbours, chosen according to stage 1, or randomly selected. We also show the "Grid" class to indicate that stage 1 is more important to some classes then others

Stage 1:	SPADE (10 Random)	SPADE (10NN)
Grid	73.2	86.3
Average	89.2	**91.4**

PatchMatch [2] method which used locality for significant speedup of the kNN search.

The role of context for anomaly detection: The quality of the alignment between the anomalous image and retrieved normal images is strongly affected by the quality of extracted features. Similarly to other works dealing with detection and segmentation, the context is very important. Local context is needed for achieving segmentation maps with high-pixel resolutions. Such features are generally found in the shallow layers of a deep neural networks. Local context is typically insufficient for alignment without understanding the global context i.e. location of the part within the object. Global context is generally found in the deepest layers of a neural network, however global context features are of low resolution. The combination of features from different levels allows both global context and local resolution giving high quality correspondences. The idea is quite similar to that in Feature Pyramid Networks [22].

Optimizing runtime performance: Our method is significantly reliant on the K nearest neighbors algorithm. The complexity of kNN scales linearly with the size of the dataset used for search which can be an issue when the dataset is very large or of high dimensionality. Our approach is designed to mitigate the complexity issues. First, we compute the initial image-level anomaly classification on global-pooled features which are 2048 dimensional vectors. Such kNN computation can be achieved very quickly for moderate sized datasets and different speedup techniques (e.g. KDTrees) can be used for large scale datasets. The anomaly segmentation stage requires pixel-level kNN computation which is significantly slower than image-level kNN. However, our method limits the sub-image kNN search to only the K nearest neighbors of the anomalous image significantly limiting computation time. We assume that the vast majority of images are normal, therefore only a small fraction of images require the next stage of anomaly segmentation. Additionally, the anomaly segmentation stage is required for explainability and trust building with the human operators, but in many cases it is not time-critical therefore putting a laxer requirement on computation time. Our method is therefore quite suitable for practical deployment from a complexity and runtime perspective.

Pre-trained vs. learned features: Previous sub-image anomaly detection methods have either used self-learned features or a combination of self-learned and pre-trained images features. Self-learned approaches in this context, typically train an autoencoder and use its reconstruction error for anomaly detection. Other approaches have used a combination of pre-trained and self-learned methods e.g. methods that use perceptual losses and [6] which uses a pre-trained encoder. Our numerical results have shown that our method significantly outperforms such approaches. We believe that given the limited supervision and small dataset size in normal-only training set as tackled in this work, it is rather hard to beat very deep pre-trained networks. We therefore use pre-trained features and do not attempt to modify them. The strong results achieved by our method attest to the effectiveness of this approach. We believe that future work should focus on methods for finetuning the deep pre-trained features for this particular task and expect it it improve over our method. That not-withstanding the ease of deployment and generality of our approach should make it a good choice in many practical settings.

6 Conclusion

We presented a novel alignment-based method for detecting and segmenting anomalies inside images. Our method relies on K nearest neighbors of pixel-level feature pyramids extracted by pre-trained deep features. Our method consists of two stages, which are designed to achieve high accuracy and reasonable computational complexity. Our method was shown to outperform the strongest current methods on two realistic sub-image anomaly detection datasets, while being much simpler. The ease of deployment enjoyed by our method should make it a good candidate for practitioners.

7 Acknowledgements

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

References

1. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: Ganomaly: Semi-supervised anomaly detection via adversarial training. In: ACCV (2018)
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. In: ACM ToG (2009)
3. Bergman, L., Cohen, N., Hoshen, Y.: Deep nearest neighbor anomaly detection. arXiv preprint arXiv:2002.10445 (2020)
4. Bergman, l., Hoshen, Y.: Classification-based anomaly detection for general data. In: ICLR (2020)
5. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In: CVPR (2019)
6. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. arXiv preprint arXiv:1911.02357 (2019)
7. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys (CSUR) (2009)
8. Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., Kloft, M.: Anomaly detection with generative adversarial networks. In: ICLR (2018)
9. Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S.: A geometric framework for unsupervised anomaly detection. In: Applications of data mining in computer security (2002)
10. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. ICLR (2018)
11. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. In: NeurIPS (2018)
12. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: ICCV (2019)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
14. Green, M.W.: The appropriate and effective use of security technologies in US schools: A guide for schools and law enforcement agencies. US Department of Justice, Office of Justice Programs, National Institute of ... (1999)
15. Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) (1979)
16. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: CVPR (2016)
17. Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., Song, D.: A benchmark for anomaly segmentation. arXiv preprint arXiv:1911.11132 (2019)
18. Hendrycks, D., Mazeika, M., Dietterich, T.G.: Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606 (2018)
19. Huang, C., Cao, J., Ye, F., Li, M., Zhang, Y., Lu, C.: Inverse-transform autoencoder for anomaly detection. arXiv preprint arXiv:1911.10676 (2019)
20. Jolliffe, I.: Principal component analysis. Springer (2011)
21. Latecki, L.J., Lazarevic, A., Pokrajac, D.: Outlier detection with kernel density functions. In: International Workshop on Machine Learning and Data Mining in Pattern Recognition (2007)

22. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
23. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked rnn framework. In: ICCV (2017)
24. Napoletano, P., Piccoli, F., Schettini, R.: Anomaly detection in nanofibrous materials by cnn-based self-similarity. Sensors (2018)
25. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing (2008)
26. Ruff, L., Gornitz, N., Deecke, L., Siddiqui, S.A., Vandermeulen, R., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: ICML (2018)
27. Sakurada, M., Yairi, T.: Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: MLSD (2014)
28. Scholkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C.: Support vector method for novelty detection. In: NIPS (2000)
29. 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: ICCV (2017)
30. Tax, D.M., Duin, R.P.: Support vector data description. Machine learning (2004)
31. Venkataramanan, S., Peng, K.C., Singh, R.V., Mahalanobis, A.: Attention guided anomaly detection and localization in images. arXiv preprint arXiv:1911.08616 (2019)
32. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004)
33. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200 (2010)
34. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
35. Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.S.: Spatio-temporal autoencoder for video anomaly detection. In: Proceedings of the 25th ACM international conference on Multimedia (2017)
36. Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. ICLR (2018)