OmniAL: A unified CNN framework for unsupervised anomaly localization

Ying Zhao
Ricoh Software Research Center (Beijing) Co., Ltd.
zy_deepwhite_zy@hotmail.com

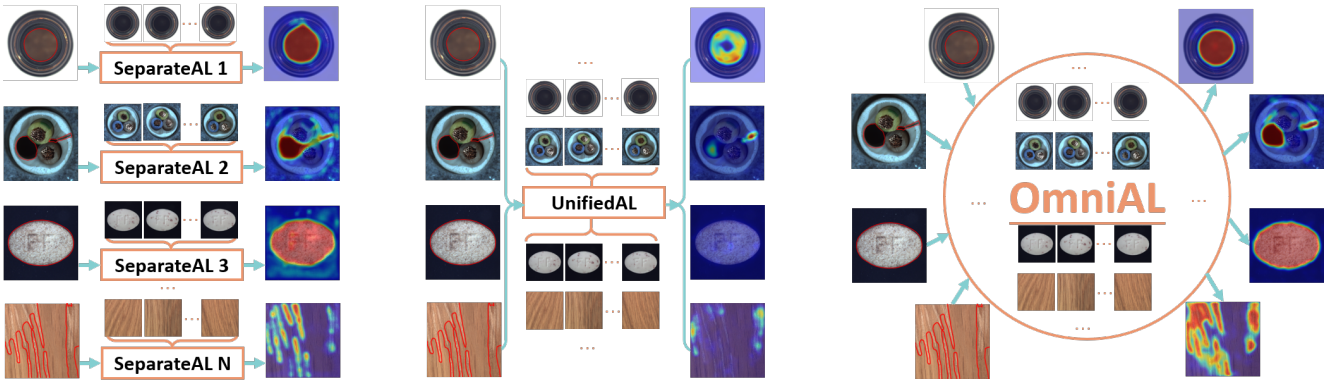


Figure 1. **OmniAL is a unified framework for unsupervised anomaly localization.** As shown in (c), OmniAL not only simplifies existing learning paradigm from (a) Separate: N models for N classes to (b) Unified: one model for N classes but also produces high quality anomaly localization results. Results in (a) and (b) are from separate and unified models of JNLD [38] on MVTecAD [1] dataset.

Abstract

Unsupervised anomaly localization and detection is crucial for industrial manufacturing processes due to the lack of anomalous samples. Recent unsupervised advances on industrial anomaly detection achieve high performance by training separate models for many different categories. The model storage and training time cost of this paradigm is high. Moreover, the setting of one-model-N-classes leads to fearful degradation of existing methods. In this paper, we propose a unified CNN framework for unsupervised anomaly localization, named OmniAL. This method conquers aforementioned problems by improving anomaly synthesis, reconstruction and localization. To prevent the model learning identical reconstruction, it trains the model with proposed panel-guided synthetic anomaly data rather than directly using normal data. It increases anomaly reconstruction error for multi-class distribution by using a network that is equipped with proposed Dilated Channel and Spatial Attention (DCSA) blocks. To better localize the anomaly regions, it employs proposed DiffNeck between reconstruction and localization sub-networks to explore multi-level differences. Experiments on 15-class MVTecAD and 12-class VisA datasets verify the advantage of proposed OmniAL that surpasses the state-of-the-art of unified models. On 15-class-MVTecAD/12-class-VisA, its single unified model achieves 97.2/87.8 image-AUROC, 98.3/96.6 pixel-AUROC and 73.4/41.7 pixel-AP for anomaly detection and localization respectively. Besides that, we make the first attempt to conduct a comprehensive study on the robustness of unsupervised anomaly localization and detection methods against different level adversarial attacks. Experiential results show OmniAL has good application prospects for its superior performance.

1. Introduction

In real industrial scenarios, the location of anomaly [22, 26] reveals important information, such as defective types and degrees. It is essential not only to inspect whether a sample is defective but also to know where the specific anomaly regions are. Since anomaly appearance is inexhaustible, it is almost impossible and infeasible to collect and manually annotate all kinds of abnormal data. Thus, only normal samples are available for training a detector that is robust enough to find out unseen anomalies during inference phase. Considering the diversity of classes and various types of one class, the conventional training paradigm of N models for N classes, as shown in Fig.1a, may not be the best solution. The model storage and training time cost increase with the number of classes. As shown in Fig.1b, existing method severely degrades anomaly local-

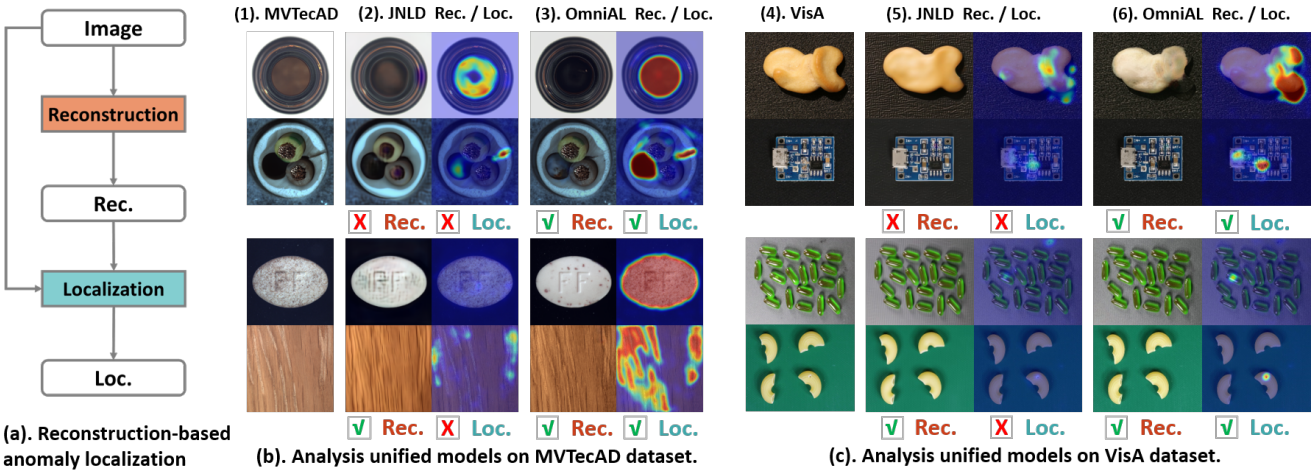


Figure 2. Problem analysis. The final failure may caused by reconstruction and localization.

ization performance if the training paradigm changes to one model for N classes. Therefore, a robust unified framework for unsupervised anomaly localization is highly demanded for intelligent industrial.

With the limitation of available training data, many appealing unsupervised approaches [14, 18, 35, 38] using synthesized anomaly data are proposed. These approaches generate anomalous instances to inspire the anomaly detector to learn discriminative features. Their experiments show that the realisticness of generated anomalous instances had a strong impact on the quality of anomaly localization. However, none of these methods consider the training paradigm of one model for N classes. When switching to the unified training paradigm, they are more prone to learn an identical short-cut and fail to discriminate the anomaly.

With the normal and synthesized anomalous samples, recent unsupervised learning methods train a deep anomaly detector by either a distance-based [6, 14, 19–21, 27] or reconstruction-based [2, 9, 35, 36, 38] way. The reconstruction-based architectures [35, 38] are supposed to reconstruct normal images more accurately than the unseen anomalous. The anomaly localization is then calculated from the reconstruction error between the original and reconstructed versions of the input image, as shown in Fig.2a. The prediction of anomaly location is not only based on the reconstruction quality but also the ability of spotting the reconstruction error. The typical reconstruction-based method JNLD [38] learns a joint representation of an anomalous image and its anomaly-free reconstruction, while simultaneously learning a decision boundary between normal and simulated anomalous examples. As shown in Fig.2b and Fig.2c, under the unified setting, JNLD [38] fails to produce correct results either because of the reconstruction failure or the localization failure.

To conquer aforementioned problems, we propose a novel unified framework OmniAL for effectively localizing anomaly pixels of different classes only by using a single model. OmniAL uses a panel-guided anomaly synthesis method that controls the portion of normal and anomaly regions for each training sample. By doing this, OmniAL blocks the chance of learning identical shortcut from the source. To increase the anomaly reconstruction error for multi-class distribution, OmniAL constructs a reconstruction and a localization sub-networks that are equipped with proposed Dilated Channel and Spatial Attention (DCSA) blocks. To better localize the anomaly regions, OmniAL employs a DiffNeck module between the reconstruction and localization sub-networks to explore multi-level reconstruction errors. As shown in Fig.1 and Fig.2, OmniAL learns a single unified model for multiple classes that produces high quality reconstruction and precise anomaly localization. Furthermore, we conduct an exhaustive evaluation of reconstruction and localization performance against to multi-level adversarial attacks.

In summary, we make following main contributions:

- We construct a unified CNN framework OmniAL for unsupervised anomaly localization that is equipped with proposed panel-guided anomaly synthesis, DCSA block, and DiffNeck module. OmniAL achieves superior performance for anomaly localization on challenging MVTecAD [1] and VisA [39] datasets compared to the state-of-the-art.

- By preventing model from learning identical reconstruction, our proposed panel-guided anomaly synthesis method also brings substantial improvement for existing methods under the unified setting. It boosts the image-AUROC/pixel-AUROC/pixel-AP from 88.7/87.1/49.4 to 92.5/94.5/57.4 for Draem [35].

- We make a comprehensive study on the robustness of separate/unified anomaly localization methods against different level adversarial attacks. Our synthesized

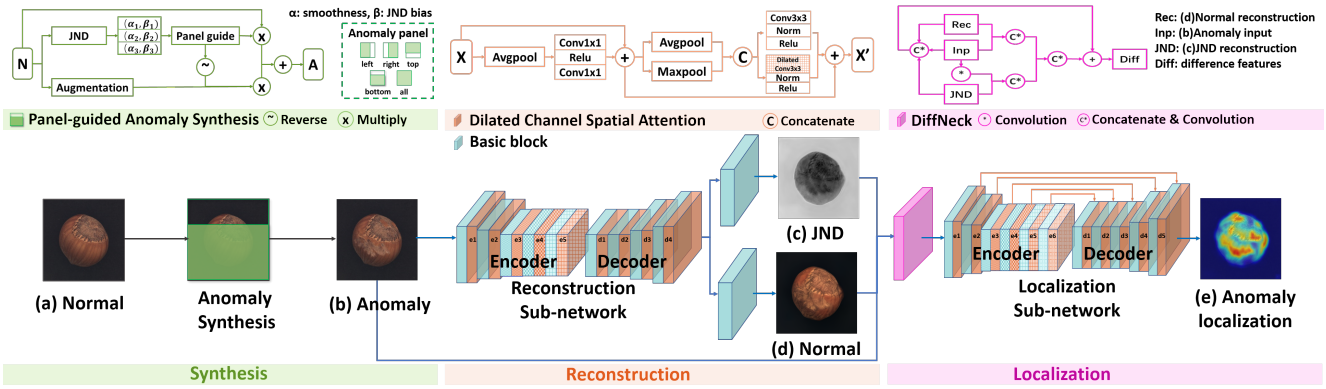


Figure 3. **Framework of OmniAL.** It consists of panel-guided anomaly synthesis, reconstruction and localization. Anomaly synthesis is based on anomaly panel and three variants of the Just Noticeable Distortion (JND) map. The synthetic anomaly is reconstructed into normal image and corresponding JND map by the Dilated Channel Spatial Attention (DCSA) modules equipped reconstruction sub-network. The localization sub-network with a DiffNeck module localizes the anomaly regions by exploring the difference between reconstructed and original data.

adversarial datasets exhibit strong attack capability against anomaly detection, reconstruction and localization, also helping to analyse the risks of existing methods.

2. Related Work

Anomaly synthesis. Due to the lack of anomaly samples, unsupervised learning methods are commonly used in industrial quality inspection. These methods are typically trained with normal data and synthesized anomalous that highly effect the performance. Therefore, how to synthesize the anomalous [7, 8, 14, 35, 37, 38] also draws extensive attention. CutPaste [14] learns representations by classifying normal data from the Cut-and-Paste augmentation. SPD [39] uses a smoothed version of CutPaste [14] augmentation. Instead of using simple regular shaped anomalous, Draem [35] simulates just-out-of-distribution anomalies having random shape and texture. To simulate photo-realistic anomaly samples, JNLD [38] proposes a multi-scale noticeable anomalous generation method based on just noticeable distortion [31]. The open-set supervised anomaly detection method DRA [7] adapts the popular CutMix [34] and the outlier exposure method [10] to generate pseudo anomalies from normal images for training. DSR [37] generates the anomalies at the feature level by sampling the learned quantized feature space, which allows a controlled generation of near-in-distribution anomalies. For the unified paradigm, we propose a panel-guided anomaly synthesis method that controls the portion of normal and anomaly regions for each training sample.

Anomaly localization. Anomaly detection [22, 26], also known as outlier detection or one-class classification, refers to the task of distinguishing defective image at the image-level from the majority of anomaly-free images. Anomaly localization (segmentation), on the other hand, aims to segment out the pixel-level anomaly regions. For industrial visual inspection, many recent methods [5, 6, 19] achieve high performance in anomaly detection but anomaly localization. SPADE [5] detects anomaly based on alignment between an anomalous image and a constant number of the similar normal images. It relies on K nearest neighbors of pixel-level feature pyramids extracted by pre-trained deep features. PSVDD [32] extends deep Support Vector Data Description (SVDD) to a patch-wise detection method. PaDim [6] also relies on ImageNet pretrained feature extractor with multi-scale pyramid pooling. Instead of using time-costly clustering, it uses a well-known Mahalanobis distance metric [17] as an anomaly score. CutPaste [14] uses GradCAM [24] to get the defect localization. MKD [23] proposes to use multi-level features alignment to increase the discriminating capability of the Teacher-Student model on various types of abnormalities. PatchCore [19] combines patch-level embeddings from ImageNet models with an outlier detection model. However, PatchCore is inherently not suitable for unified setting (larger normal dataset) since it needs to build the coreset with all normal data. Draem [35] and JNLD [38] use segmentation sub-network to predict the defective regions. This paper differs from these previous works by focusing on the paradigm of only using a single unified model for N classes.

Unified anomaly localization. Recently, how to use a unified model to localize anomaly for different objects has already become researchers' concern. RegAD [11] trains a single generalizable model for few-shot anomaly detection, where a limited number of normal images are provided for each category at training. It employs SimSiam [3] with three spatial transformer network [13] blocks to solve the category-agnostic feature registration proxy task. It identifies anomalies by comparing the registered features of the test image and the corresponding normal images. UniAD [33] constructs a transformer with a layer-wise query de-

coder and a neighbor masked attention module to model the multi-class distribution. It uses the layer-wise query decoder to intensify the use of query embedding. To avoid the information leak, it employs a neighbor masked attention module, where a feature point relates to neither itself nor its neighbors. In this paper, we propose a unified CNN framework for unsupervised anomaly localization. It manifests the problem-solving ability of fully convolutional networks.

3. Method

In this work, we construct a unified CNN framework OmniAL for unsupervised anomaly localization under the practical setting of only using one model for N classes. Fig.3 shows the overview of proposed OmniAL. To train only with normal data, we firstly synthesize photo-realistic anomalous in the regions that are selected based 5 predefined panels. The synthesized anomaly samples are then used to train the OmniAL network that consists of an anomaly-free reconstruction and an anomaly localization sub-networks. The reconstruction and localization sub-networks are composed of alternative basic blocks and DCSA blocks. The reconstruction sub-network learns to recover the synthesized anomalous back to normal. The localization sub-network with a DiffNeck module localizes the anomaly regions by exploring the difference between reconstructed and original data.

3.1. Panel-guided anomaly synthesis

Reconstruction-based anomaly localization methods rely on the hypothesis that the reconstruction errors in unseen anomalous regions are larger than the normal regions. The reconstruction models are trained with alternative normal and synthetic anomaly samples. For normal sample, the supervision ground truth is the direct copy of the input. Thus, directly using normal samples for training increases the chance of information leak and leads to identical reconstruction for any input. To prevent the learning of identical short-cut, the intuitive idea is that only using synthetic anomaly rather than normal data. However, without using normal samples for training leads to more false alarm. Then, the problem is transformed into how to better synthesis anomaly for unified model training.

To solve the problem, we propose a panel-guided anomaly synthesis method that takes into account both normal and anomaly for training. As shown in Fig.4, we build 5 types of panels ('left','right','top','bottom','all') to control the portion of normal and anomaly regions. That is, we only generate the anomaly in one of the panel region for each sample in each iteration. Except the type of 'all', the area of panel region is also randomly adjusted in range of [0.5, 0.8]*ImageWidth and [0.5, 0.8]*ImageHeight. Given a panel, we synthesize anomaly in the selected region according to following equation.

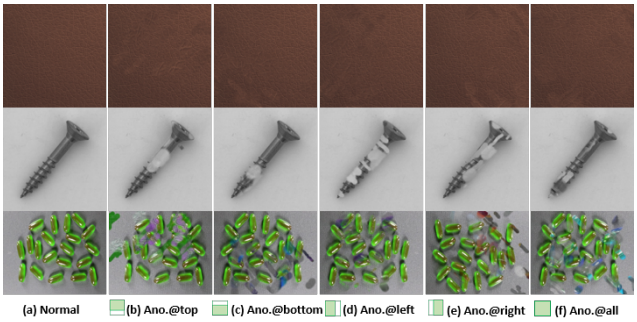


Figure 4. Visualization of our panel-guided anomaly synthetic data. The anomaly source is selected either from augmented normal images (row1,2) or DTD [4] dataset (row3).

$$A = \sum_{i=1}^{3}(W_i T + (1 - W_i)I) \tag{1}$$

Where A is the synthesized anomaly image, W_i is the pixel-wise fusion weight map, i indicates one of the 3 defect levels (easy, medium and hard), I is the anomaly-free image, T is the anomaly source image. To synthesize natural and diversity anomalies, T is generated by applying a set of random augmentation (resize, crop, flip, color-jitter) to I or image from the Describable Textures Dataset(DTD) [4].

To get the 3-level defects, we first need to find out the baseline anomaly fusion weight map. Following JNLD [38], we use the JND [31] to guide the anomaly synthesis. JND [31] refers to the minimum visibility threshold of the Human Visual System, is useful in perceptual image/video processing systems. It reveals a perceptual threshold of intensity change in an image that can be noticed by the human vision system. We use a nonlinear additivity JND model to define the anomaly weight map J as [25] that proposes a novel hybrid exposure weight measurement using the JND. Each level anomaly fusion weight map W_i is defined as:

$$W_i = \alpha_i * (J + \beta_i) \tag{2}$$

Where J indicates the JND map, β_i is a bias term for different level of noticeable, α_i is a smooth kernel applied to the mask and produces soft boundary between anomaly and normal regions. For the easier level anomaly, we use a smaller smooth kernel. Due to using random seeds, different level anomaly regions may overlap each other. Thus, directly using the masks as the ground truth for anomaly segmentation task may bring label inconsistency. Therefore, we refine and quantify the segmentation mask into 3 defect levels based on the structure similarity index measure (SSIM) [29] between the anomaly-free image I and synthesized anomaly image A.

3.2. Anomaly reconstruction

As shown in Fig.3, our anomaly reconstruction sub-network receives an input image and outputs reconstructed

normal image and corresponding JND map. The reconstruction sub-network consists of an encoder and a decoder. Both of them are composed of alternating basic and DCSA blocks. The basic block extracts different levels of features with two consecutive 3x3 convolution layer followed by normalization and ReLU activation layers. Considering the unified training paradigm, instance normalization [28] is more suitable than the broadly used batch normalization [12] in anomaly reconstruction. The reason is that batch normalization [12] normalizes using the information from the whole batch, while instance normalization [28] normalizes each feature map on its own. Under unified setting, each batch of training data consists of different classes that complicates the reconstruction problem. Instance normalization discards instance-specific contrast information and reduces undesirable instance-level variation. Batch normalization, meanwhile, lacks the ability to address the class inconsistency. Therefore, we use instance normalization in each basic block rather than batch normalization. Experimental results in Table 3 show that the performance is improved by swapping batch normalization with instance normalization.

The feature maps extracted by the basic block are further integrated by the DCSA block that is inspired by the popular CBAM [30]. As shown in orange dash block of Fig.3, the DCSA block consists of channel and spatial attention sequential sub-modules. With average pooling and two consecutive 1x1 convolutions, the channel attention explores the correlations among feature channels. The correlations are add back to original feature maps to emphasize the important feature channels. The spatial attention further investigates the spatial correlation. It first aggregates spatial information by concatenating both average and maximum pooling of afore refined feature maps. Then, it spots the spatial importance of the aggregated features with a standard and a dilated 3x3 convolution-norm-relu paths. It balances the portion of dilated path by a factor that is set as 0.1. The spatial attention map is obtained by summing up two paths of features. By adding the spatial attention map, the important regions in feature maps are highlighted. As shown in Table 3, the performance is further improved by using DCSA blocks.

3.3. Anomaly localization

After reconstructing the normal and corresponding JND maps, we use a localization sub-network to highlight the anomaly regions. Instead of directly concatenating the reconstructed maps with input image, we firstly use a DiffNeck module to fully explore the difference between them. The pink dash block of Fig.3 illustrates the detailed structure of DiffNeck. Overall, DiffNeck extracts reconstruction difference among reconstructed normal map (Rec), reconstructed JND map (JND) and input image (Inp) in both separate and unified ways. That is, DiffNeck uses basic blocks to extract the concatenated features of (Rec, Inp), (JND, Inp) and (Rec, JND, Inp) respectively. Then, all levels of differences are summed up and form a final descriptor that is fed to the localization sub-network. DiffNeck also balances the portion of the separate path by a factor that is set as 0.1.

Similar with reconstruction sub-network, the localization sub-network also follows the encoder-decoder structure but with 6 scales (e1-e5 in Fig.3) and skip connections for corresponding scales. Scale-1 and scale-2 use standard convolution in all blocks. To use larger resolution feature maps and receptive field, the basic and DCSA blocks of the encoder contain dilated convolution layer from scale-3 to scale-6 with 2, 4, 8, 16 dilation rates respectively. Different with reconstruction sub-network, we use batch normalization in all blocks. Since we synthesize more realistic anomalies by considering easy, medium and hard level defects, localization sub-network predicts 3 levels of anomalous locations. To get the final pixel-level anomaly localization map, we combine them together by a sequence processing of SoftMax, smoothness and summing up. By doing this, the false alarm noise can be suppressed, and the weak true decision can be strengthened.

The total target for training OmniAL is defined as

$$\begin{aligned} L =& L_2(J, J_r) + L_{ssim}(J, J_r) \\ &+ L_2(I, I_r) + L_{ssim}(I, I_r) + L_{fl}(S, S_s) \end{aligned} \tag{3}$$

Where J is the ground truth JND map calculated from the anomaly-free image I, J_r and I_r are reconstructed JND map and normal image, S is the anomaly localization ground truth. Unlike JNLD [38], for both JND and normal reconstruction, we not only use MSE loss to supervise the pixel-to-pixel recovering but also the SSIM loss to yield plausible local consistency. To handle the unbalance of different types, we use focal loss [15] to supervise the predicted anomaly localization S_s.

4. Experimental results

4.1. Datasets and metrics

To demonstrate the effectiveness of proposed OmniAL, we conduct extensive experiments on the challenging MVTecAD [1] and VisA [39] datasets.

MVTecAD [1] dataset contains 10 object and 5 texture industrial products, such as bottle and leather. It consists of 3,629 normal images for training and 1,725 images for testing. There are 1,258 anomaly images of the testing set with pixel-level labelled various types of defects and the rest are normal images. Each class contains 60 to 320 color images with the resolution ranges from 700x700 to 1024x1024 pixels. In the testing set, defective appearance varies in different sizes, shapes and types, and most cases only contain a small fraction of anomalous pixels.

Table 1. Image AUROC comparison of our method OmniAL with the state-of-the-art on MVTecAD [1].

Category	Padim [6]	CutPaste [14]	MKD [23]	Draem [35]	JNLD [38]	UniAD [33]	OmniAL
	Unified / Separate						
bottle	97.9 / 99.9	67.9 / 98.2	98.7 / 99.4	94.6 / 99.1	99.1 / 96.3	99.7 / **100**	**100** / 99.4
cable	70.9 / 92.7	69.2 / 81.2	78.2 / 89.2	61.8 / 94.7	90.6 / **98.8**	95.2 / 97.6	**98.2** / 97.6
capsule	73.4 / 91.3	63.0 / 98.2	68.3 / 80.5	70.2 / **98.5**	74.4 / 85.0	86.9 / 85.3	**95.2** / 92.4
carpet	93.8 / 99.8	93.6 / 93.9	69.8 / 79.3	95.9 / 95.5	77.1 / 97.8	99.8 / **99.9**	**98.7** / 99.6
grid	73.9 / 96.7	93.2 / **100**	83.8 / 98.0	98.1 / 99.9	98.6 / **100**	98.2 / 98.5	**99.9** / **100**
hazelnut	85.5 / 92.0	80.9 / 98.3	97.1 / 98.4	95.1 / **100**	90.8 / **100**	**99.8** / 99.9	95.6 / 98.0
leather	99.9 / **100**	93.4 / **100**	93.6 / 95.1	99.9 / **100**	97.0 / **100**	**100** / **100**	99.0 / 97.6
metal nut	88.0 / 98.7	60.0 / **99.9**	64.9 / 73.6	88.9 / 98.7	93.3 / 99.6	**99.2** / 99.0	**99.2** / **99.9**
pill	68.8 / 93.3	71.4 / 94.9	79.7 / 82.7	69.0 / 98.9	82.7 / 94.6	93.7 / 88.3	**97.2** / **97.7**
screw	56.9 / 85.8	85.2 / 88.7	75.6 / 83.3	93.3 / 93.9	81.8 / **95.8**	87.5 / 91.9	**88.0** / 81.0
tile	93.3 / 98.1	88.6 / 94.6	89.5 / 91.6	98.3 / 99.6	99.2 / **100**	99.3 / 99.0	99.6 / **100**
toothbrush	95.3 / 96.1	63.9 / 99.4	75.3 / 92.2	82.8 / **100**	**100** / **100**	94.2 / 95.0	**100** / **100**
transistor	86.6 / 97.4	57.9 / 96.1	73.4 / 85.6	83.9 / 93.1	90.3 / 93.0	**99.8** / **100**	93.8 / 93.8
wood	72.1 / 96.5	80.4 / 99.1	93.4 / 94.3	**99.8** / 99.1	91.9 / **99.6**	98.6 / 97.9	93.2 / 98.7
zipper	79.7 / 90.3	93.5 / 99.9	87.4 / 93.2	99.1 / **100**	99.8 / 99.8	95.8 / 96.7	**100** / **100**
average	84.2 / 95.5	77.5 / 96.1	81.9 / 87.8	88.7 / **98.0**	91.3 / 97.4	96.5 / 96.6	**97.2** / 97.0

Table 2. Pixel AUROC comparison of our method OmniAL with the state-of-the-art on MVTecAD [1].

Category	Padim [6]	PSVDD [32]	MKD [23]	Draem [35]	JNLD [38]	UniAD [33]	OmniAL
	Unified / Separate						
bottle	96.1 / 98.2	86.7 / 98.1	91.8 / 96.3	87.4 / **99.1**	94.8 / 99.0	98.1 / 98.1	**99.2** / 99.0
cable	81.0 / 96.7	62.2 / 96.8	89.3 / 82.4	70.4 / 94.7	76.4 / **97.7**	**97.3** / 96.8	**97.3** / 97.1
capsule	96.9 / 98.6	83.1 / 95.8	88.3 / 95.9	49.2 / 94.3	57.0 / 92.7	**98.5** / **97.9**	96.9 / 92.2
carpet	97.6 / 99.0	78.6 / 92.6	95.5 / 95.6	95.2 / 95.5	93.7 / 99.0	98.5 / 98.0	**99.4** / **99.6**
grid	71.0 / 97.1	70.8 / 96.2	82.3 / 91.8	99.0 / 99.7	96.9 / **99.7**	98.2 / 98.5	**99.4** / 99.6
hazelnut	96.3 / 98.1	97.4 / 97.5	91.2 / 94.6	96.0 / **99.7**	85.9 / 99.4	96.5 / 94.6	**98.4** / 98.6
leather	84.8 / 99.0	93.5 / 97.4	96.7 / 98.1	98.6 / 98.6	87.0 / 99.5	98.8 / 98.3	**99.3** / **99.7**
metal nut	84.8 / 97.3	96.0 / 98.0	64.2 / 86.4	72.6 / **99.5**	97.4 / **99.5**	94.8 / 95.7	**99.1** / 99.1
pill	87.7 / 95.7	96.5 / 95.1	69.7 / 89.6	90.0 / 97.6	91.2 / 96.6	95.0 / 95.1	**98.9** / **98.6**
screw	94.1 / 98.4	74.3 / 95.7	92.1 / 96.0	89.3 / 97.6	87.0 / **99.7**	**98.3** / 97.4	98.0 / 97.2
tile	80.5 / 94.1	92.1 / 91.4	85.3 / 82.8	98.1 / 99.2	94.7 / **99.6**	91.8 / 91.8	**99.0** / 99.4
toothbrush	95.6 / 98.8	98.0 / 98.1	88.9 / 96.1	94.4 / 98.1	98.6 / 98.8	98.4 / 97.8	**99.4** / **99.2**
transistor	92.3 / 97.6	78.5 / 97.0	71.7 / 76.5	73.1 / 90.9	83.6 / 92.1	**97.9** / **98.7**	93.3 / 91.7
wood	89.1 / 94.1	80.7 / 90.8	80.5 / 84.8	96.2 / 96.4	88.7 / 96.3	93.2 / 93.4	**97.4** / **96.9**
zipper	94.8 / 98.4	95.1 / 95.1	86.1 / 93.9	96.9 / 98.8	95.3 / 99.4	96.8 / 96.0	**99.5** / **99.7**
average	89.5 / 97.4	85.6 / 95.7	84.9 / 97.0	87.1 / 97.3	88.6 / **97.9**	96.8 / 96.6	**98.3** / 97.8

Table 3. Ablation study on MVTecAD [1]. PA: Panel-guided anomaly synthesis, Rec: Reconstruction sub-network, Seg: Segmentation sub-network, BB: Basic Block, BN: Batch Norm, IN: Instance Norm, DC: Dilated Convolution, CA: Channel Attention, DSA: Dilated Spatial Attention, I: Image-level classification, P: Pixel-level localization.

PA	Rec-BB			DCSA		Seg-BB		DiffNeck	I-AUROC	P-AUROC	P-AP
-	BN	IN	DC	CA	DSA	BN	DC	-	Unified/Separate		
-	√	-	-	√	-	√	-	-	86.7/**98.8**	86.4/**98.4**	44.2/**75.0**
√	√	-	-	√	-	√	-	-	94.1/—	95.9/—	68.3/—
√	-	√	-	√	-	√	-	-	95.8/—	97.0/—	69.1/—
√	-	√	√	√	-	√	√	-	96.9/97.3	97.6/97.7	72.8/72.9
√	-	√	√	√	√	√	√	-	94.7/96.5	96.9/97.8	69.8/74.9
√	-	√	√	√	-	√	√	√	96.8/96.3	98.2/98.0	72.7/74.1
√	-	√	√	√	√	√	√	√	**97.2**/97.0	**98.3**/97.8	**73.4**/73.5

Table 4. Ablation study on With/Without panel-guided anomaly synthesis for unified training on MVTecAD [1].

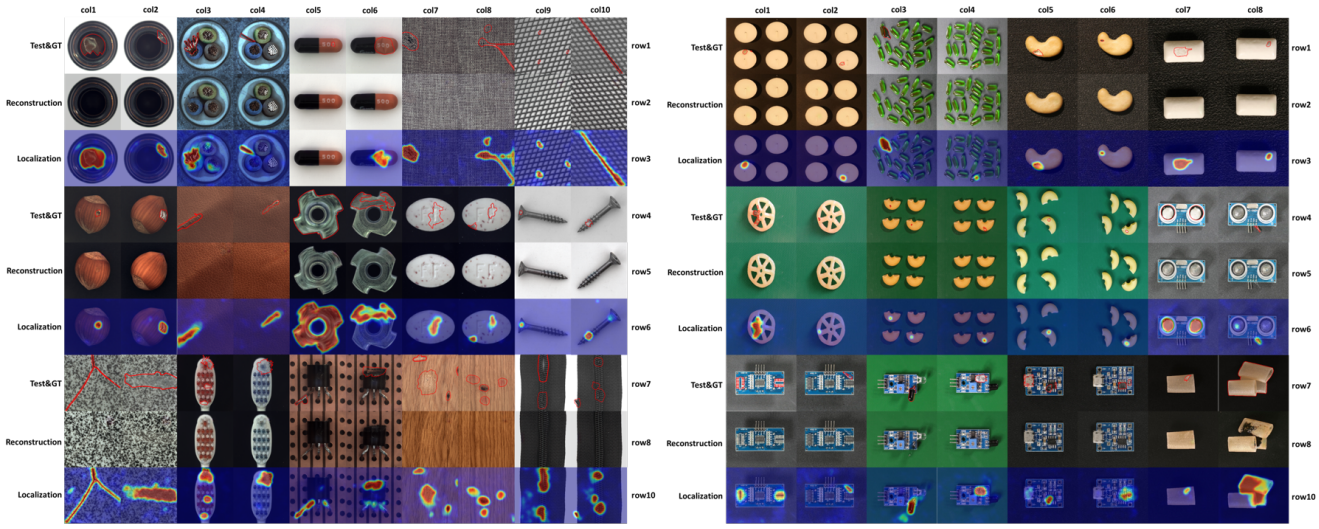
	Draem [35]	JNLD [38]	OmniAL
I-AUROC	92.5/88.7	92.9/**91.3**	**97.2**/86.6
P-AUROC	94.5/87.1	95.6/88.6	**98.3**/**93.7**
Pixel-AP	57.4/49.4	63.1/46.6	**73.4**/**54.4**

VisA [39] dataset consists of 10,821 high-resolution color images (9,621 normal and 1,200 anomalous samples) covering 12 objects in 3 domains, including complex structure, multiple instances and single instances. 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. All images were acquired using a 4,000x6,000 high-resolution RGB sensor. Example of each category and our corresponding anomaly reconstruction and localization results are shown in Fig.5a(MVTecAD) and Fig.5b(VisA).

Metrics For anomaly detection evaluation, the most common used metrics are Area Under the Receiver Operating Characteristic curve (AUROC) in both image-level and pixel-level. However, the pixel-AUROC is biased in favour of large anomalous regions and does not well reflect the pixel-level anomalous localization performance. Therefore, we additionally introduce the average precision (AP) to evaluate pixel-level anomaly localization performance. For reconstruction quality evaluation, we use peak signal to noise ratio (PSNR) and SSIM [29] that are commonly used for evaluating signal fidelity.

Table 5. Performance comparison of our method OmniAL with the state-of-the-art on VisA [39]. (Separate/Unified)

	Category	Padim+SPD [39]		Draem [35]			JNLD [38]			OmniAL		
		I-AUC	P-AUC	I-AUC	P-AUC	P-AP	I-AUC	P-AUC	P-AP	I-AUC	P-AUC	P-AP
Complex structure	PCB1	92.7	97.7	71.3/83.9	98.6/94.0	60.4/38.5	82.0/**82.9**	96.4/**98.0**	**72.8**/77.8	**96.6**/77.7	**98.7**/97.6	63.5/**76.9**
	PCB2	87.9	**97.2**	89.7/81.7	92.5/94.1	3.5/13.3	96.3/79.1	91.9/**95.0**	**31.2**/**37.7**	**99.4**/81.0	83.2/93.9	2.8/31.6
	PCB3	85.4	96.7	73.1/87.7	93.8/94.1	18.7/17.9	**96.9**/**90.1**	95.3/**98.5**	43.4/**46.8**	**96.9**/88.1	**98.4**/94.7	56.9/41.4
	PCB4	**99.1**	89.2	91.3/87.1	95.8/72.3	32.7/13.1	94.8/**96.2**	96.1/**97.5**	37.4/29.7	97.4/95.3	**98.5**/97.1	**38.4**/**33.2**
Multiple instances	Macaroni1	85.7	98.8	70.3/68.6	95.8/89.8	8.2/8.0	94.3/90.5	98.8/93.3	**25.9**/**14.2**	**96.9**/**92.6**	**98.9**/**98.6**	7.6/7.1
	Macaroni2	70.8	96.0	71.3/60.3	94.1/83.2	**25.4**/19.7	86.5/71.3	92.9/92.1	17.2/6.1	**89.9**/**75.2**	**99.1**/**97.9**	11.4/**9.2**
	Capsules	68.1	86.3	77.3/89.6	93.7/96.6	20.2/30.1	**89.1**/**91.4**	**98.9**/**99.6**	38.6/**55.6**	87.9/90.6	98.6/99.4	**62.9**/52.4
	Candles	**89.1**	**97.3**	82.3/70.2	87.0/82.6	27.9/12.6	**89.1**/85.4	94.8/94.5	25.9/**27.4**	85.1/**86.8**	90.5/**95.8**	**29.2**/24.6
Single instance	Cashew	90.5	86.1	94.2/67.3	94.7/68.5	41.2/7.0	96.0/82.5	96.3/94.1	43.7/39.4	**97.1**/**88.6**	**98.9**/**95.0**	**77.3**/**47.4**
	Chewing gum	**99.3**	96.9	93.4/90.0	97.5/92.7	40.9/59.0	98.5/96.0	**99.4**/98.9	74.7/**81.6**	94.9/**96.4**	98.7/**99.0**	**82.9**/79.5
	Fryum	89.8	88.0	**100**/86.2	**97.5**/83.2	40.9/26.4	93.2/91.9	95.8/90.0	**42.9**/30.4	97.0/**94.6**	89.3/**92.1**	28.3/**34.4**
	Pipe fryum	95.6	95.4	94.1/87.1	81.8/72.3	23.7/13.1	**96.0**/**87.5**	97.0/92.5	44.1/31.8	91.4/86.1	**99.1**/**98.2**	**69.1**/**62.6**
	Mean	87.8	93.8	84.1/80.5	88.8/87.0	25.4/20.8	93.0/87.1	**96.1**/95.2	41.5/39.9	**94.2**/**87.8**	96.0/**96.6**	**44.2**/**41.7**



(a). 15 categories of MVTecAD dataset (b). 12 categories of VisA dataset

Figure 5. Qualitative illustration of our anomaly detection results. From top to bottom, the test images overlap with ground truth contours, the reconstructions and the anomaly localization map overlays are shown in rows.

4.2. Implementation

To compare with existing methods, we train OmniAL both in unified and separate paradigm with a batch size of 20/12 images having size of 256x256/256x320 for MVTec/VisA and pixel value range of [0, 1] . A single model is trained with batches that contain all-class samples for unified paradigm. On the contrary, the separate models are trained with corresponding class samples respectively. The Adam optimizer has an initial learning rate of 1e-4. To alleviate over-fitting, during training, the anomaly texture is alternatively selected either from augmented normal images or from the DTD [4] dataset.

4.3. Comparison and performance

Performance on MVTecAD. We compare our proposed method with state-of-the-art methods recently reported on MVTecAD for both unified and separate models in anomaly detection and localization. Table 1 and 2 show our quantitative comparison with the state-of-the-art methods on the task of anomaly detection and localization. UniAD [33] is designed for the unified training paradigm while the others are for the separate training scheme. The results of unified model of Draem [35] and JNLD [38] are got by re-training their models under the paradigm of one model for N classes. The others are reported from UniAD [33]. As shown in Table 1 and 2, for both image-level and pixel-level anomaly detection, most of the existing methods' performance drop drastically when the paradigm switch from training N separate models to a unified model. Comparing with the transformer-based method UniAD [33], our CNN-based method achieves 0.7% and 1.5% higher performance on image-AUROC and pixel-AUROC. Moreover, as shown in Table 4, OmniAL(73.4) surpasses SOTA reconstruction-based methods Draem(49.4) [35] and JNLD(46.6) [38] in pixel-AP with more than 24%.

Performance on VisA. Comparing with MVTecAD, VisA is more difficult since it considers more complex structure and multiple misaligned instances scenes. Table 5 shows the superior performance of OmniAL comparing with the baseline separate method SPD [39] that proposes the VisA dataset and two reconstruction-based methods under the unified setting. The results of both separate and uni-

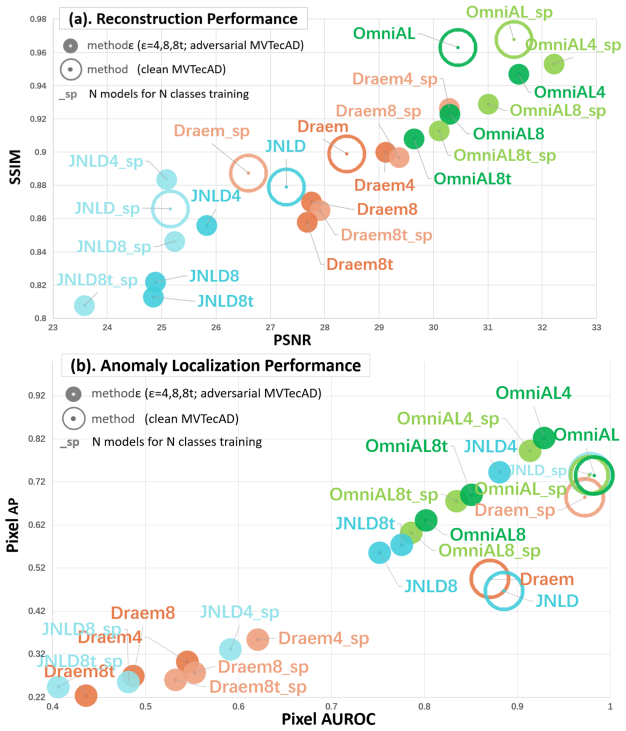


Figure 6. Robustness comparison of our method OmniAL with the state-of-the-art against adversarial attack.

fied model of Draem [35] and JNLD [38] are got by re-training on VisA. For separate training, OmniAL surpasses the best of them with 1.2% image-AUROC, 2.7% pixel-AP and similar pixel-AUROC. For unified training, OmniAL also surpasses the best of them with 0.7% image-AUROC, 1.4% pixel-AUROC and 1.8% pixel-AP.

As shown in Fig.5, our reconstruction and localization achieve good performance in most categories. Even in the multi-instance scene, OmniAL still successfully recovers the anomaly region back to normal. The reconstructed images have the high-fidelity appearance with the inputs in normal regions and recover the anomalies as close as the expectation. The quantity evaluation of reconstruction is reported in Fig.6a. The bigger circles indicate the reconstruction performance on the clean normal MVTecAD [1]. Our OmniAL surpasses the Draem [35] and JNLD [38] with big gaps in both PSNR and SSIM scores under the unified setting.

Adversarial robustness. Adversarial attacks are easily performed under the disguise of anomaly with additive noises and hardly arouse suspicion. Our anomaly synthesis method can be expediently expanded to generate adversarial anomaly samples with PGD [16] perturbations for robustness evaluation. To further evaluate the robustness of separate and unified models, we build up 3 adversarial datasets with PGD [16] perturbations $\epsilon = \{4/255, 8/255\}$ and $\epsilon = 8/255$ with targeted attack based on the anomaly-free training set from clean MVTecAD [1]. For realistic adversarial samples, the synthesized anomalies only appear in foreground of the object categories and the anomaly texture is cropped from the normal images. The generated adversarial datasets consist of 80% synthesized anomaly and 20% normal samples. Fig.6 illustrates the performance comparison in (a) reconstruction and (b) localization respectively. The names with '_sp' suffix indicate separate models while the others are unified models. OmniAL(green) achieves the best performance against all adversarial levels and has less degradation. All category adversarial samples and performance comparison are shown in supplementary material.

4.4. Ablation study

Table 3 demonstrates the effectiveness of proposed panel-guided anomaly synthesis, DCSA block and DiffNeck module for both unified and separate training paradigm. We build a baseline by only using channel attention in DCSA blocks and achieve the best performance in separate setting but the worse performance in unified setting. By using panel-guided anomaly synthesis, we achieve about 10% image-level and 24% pixel-level performance improvement comparing with the baseline. The overall performance is further improved by using instance normalization rather than batch normalization in the reconstruction basic blocks. With the help of dilated convolution layer in basic blocks, we can preserve high resolution feature maps and get further improvement. By further using dilated spatial attention, the separate models performance gain 2% but the unified model decreases 3% in pixel-AP. Alternatively, the model using DiffNeck module gains 1.2% for separate setting and 0.6% for unified setting. Finally, by combining dilated spatial attention and DiffNeck together, we achieve the balanced performance in separate and unified settings and get the OmniAL. As shown in Table 4, proposed panel-guided anomaly synthesis also boosts existing reconstruction-based methods with a large margin. More ablation studies are shown in supplementary material.

5. Conclusion

Considering the practical usages, we propose a unified CNN framework to localize anomalous for multiple classes with a single model. Extensive experiments on MVTecAD and VisA datasets verify the effectiveness of our proposed panel-guided anomaly synthesis, dilated channel and spatial attention (DCSA) block, and DiffNeck module. Especially, OmniAL produces more precisely anomaly reconstruction and localization results. The panel-guided anomaly synthesis can be easily used in existing methods and brings performance improvement under the unified training paradigm. Moreover, we make the first attempt to conduct a comprehensive study on the robustness of existing methods against different levels of adversarial for both separate and unified training paradigm.

References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec AD - A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, pages 9592–9600, 2019. Computer Vision Foundation / IEEE. 1, 2, 5, 6, 8

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*, pages 4182–4191, 2020. Computer Vision Foundation / IEEE. 2

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15750–15758. Computer Vision Foundation / IEEE, 2021. 3

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. IEEE Computer Society. 4, 7

[5] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *CoRR*, abs/2005.02357. 3

[6] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *ICPR*, volume 12664 of *Lecture Notes in Computer Science*, pages 475–489, 2020. Springer. 2, 3, 6

[7] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7378–7388. IEEE, 2022. 3

[8] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):4505–4523, 2022. 3

[9] Denis A. Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *WACV*, pages 1819–1828, 2022. IEEE. 2

[10] Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 3

[11] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael W. Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. *CoRR*, abs/2207.07361, 2022. 3

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015. 5

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2017–2025, 2015. 3

[14] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, pages 9664–9674, 2021. Computer Vision Foundation / IEEE. 2, 3, 6

[15] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *CVPR*, pages 2999–3007, 2017. IEEE Computer Society. 5

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR, 2018*. OpenReview.net. 8

[17] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936. 3

[18] Masoud PourReza, Bahram Mohammadi, Mostafa Khaki, Samir Bouindour, Hichem Snoussi, and Mohammad Sabokrou. G2D: generate to detect anomaly. In *WACV*, pages 2002–2011, 2021. IEEE. 2

[19] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, pages 14318–14328, 2022. 2, 3

[20] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *WACV*, pages 1906–1915, 2021. IEEE. 2

[21] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *WACV*, pages 1829–1838, 2022. IEEE. 2

[22] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *CoRR*, abs/2110.14051. 1, 3

[23] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14902–14912. Computer Vision Foundation / IEEE, 2021. 3, 6

[24] 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. *Int. J. Comput. Vis.*, 128(2):336–359, 2020. 3

[25] Jianbing Shen, Ying Zhao, Shuicheng Yan, and Xuelong Li. Exposure fusion using boosting laplacian pyramid. *IEEE Trans. Cybern.*, 44(9):1579–1590, 2014. 4

[26] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *CoRR*, abs/2207.10298. 1, 3

[27] Chin-Chia Tsai, Tsung-Hsuan Wu, and Shang-Hong Lai. Multi-scale patch-based representation learning for image anomaly detection and segmentation. In *WACV*, pages 3065–3073, 2022. IEEE. 2

[28] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016. 5

[29] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 4, 6

[30] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 3–19. Springer, 2018. 5

[31] Jinjian Wu, Guangming Shi, Weisi Lin, Anmin Liu, and Fei Qi. Just noticeable difference estimation for images with free-energy principle. *IEEE Trans. Multim.*, 15(7):1705–1710, 2013. 3, 4

[32] Jihun Yi and Sungroh Yoon. Patch SVDD: patch-level SVDD for anomaly detection and segmentation. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi, editors, *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part VI*, volume 12627 of *Lecture Notes in Computer Science*, pages 375–390. Springer, 2020. 3, 6

[33] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *CoRR*, abs/2206.03687, 2022. 3, 6, 7

[34] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6022–6031. IEEE, 2019. 3

[35] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Dræm - A discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, pages 8310–8319, 2021. IEEE. 2, 3, 6, 7, 8

[36] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognit.*, 112:107706, 2021. 2

[37] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. DSR - A dual subspace re-projection network for surface anomaly detection. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pages 539–554. Springer, 2022. 3

[38] Ying Zhao. Just noticeable learning for unsupervised anomaly localization and detection. In *ICME*, pages In Press, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[39] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. *CoRR*, abs/2207.14315, 2022. 2, 3, 5, 6, 7