Glancing at the Patch: Anomaly Localization with Global and Local Feature Comparison

Shenzhi Wang[1] Liwei Wu[1] Lei Cui[1] Yujun Shen[2]
[1] SenseTime Research [2] The Chinese University of Hong Kong
{wangshenzhi, wuliwei, cuilei}@sensetime.com sy116@ie.cuhk.edu.hk

Abstract

Anomaly localization, with the purpose to segment the anomalous regions within images, is challenging due to the large variety of anomaly types. Existing methods typically train deep models by treating the entire image as a whole yet put little effort into learning the local distribution, which is vital for this pixel-precise task. In this work, we propose an unsupervised patch-based approach that gives due consideration to both the global and local information. More concretely, we employ a Local-Net and Global-Net to extract features from any individual patch and its surrounding respectively. Global-Net is trained with the purpose to mimic the local feature such that we can easily detect an abnormal patch when its feature mismatches that from the context. We further introduce an Inconsistency Anomaly Detection (IAD) head and a Distortion Anomaly Detection (DAD) head to sufficiently spot the discrepancy between global and local features. A scoring function derived from the multi-head design facilitates high-precision anomaly localization. Extensive experiments on a couple of real-world datasets suggest that our approach outperforms state-of-the-art competitors by a sufficiently large margin.

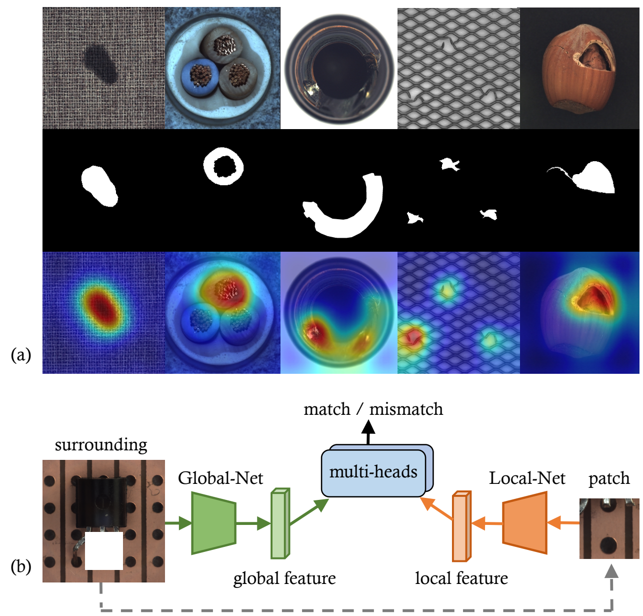


Figure 1. **(a) Anomaly localization results** where our approach can precisely segment the anomalous regions. From top to bottom: abnormal samples, ground-truth, and anomaly score maps produced by our algorithm. **(b) Concept diagram of global and local feature comparison.** Local-Net and Global-Net are employed to extract features from a patch and its surrounding respectively. Multiple anomaly detection heads are designed to determine whether the global and local features match or not.

1. Introduction

Anomaly detection has received broad attention in recent years due to its wide applications in industrial inspection [6, 7, 48, 11, 10, 32], medical diagnosis [51, 5, 46, 42], and surveillance [27, 30, 36]. Its primary goal is to identify anomalies from normal samples, usually treated as a bi-classification problem. Considering the ambiguous definition of anomaly types as well as the great imbalance between sufficient normal data and scarce abnormal data, a common practice is to learn the distribution of anomaly-free data and then use it as a criterion to detect outliers [43, 38, 42, 41, 14, 48].

Recent development of deep neural networks has significantly advanced this task with a more powerful capability in representation learning [7, 48, 28]. Most existing algorithms deploy deep models to spot anomalies at the instance level (*i.e.*, abnormal images belongs to different categories against normal ones) by extracting global feature from the entire image [14, 38, 49, 22, 37, 43, 2]. However, in many real cases, the anomalies simply differ from the regular data at some local areas [6, 32], as shown in Fig. 1a. From this perspective, localizing the anomalous regions at the pixel level is far more practical.

To solve this pixel-precise task, one feasible solution is to adopt generative models, such as Generative Adversarial Networks (GANs) [42, 41, 47] and Auto-Encoders

(AEs) [8, 11, 12, 51], which can produce images with per-pixel generation. But these approaches still treat each training image as a whole, omitting the learning of local information. An alternative solution is to pick patches from the image and then perform inspection on every patch to see whether there exists a defect [7, 10]. But this kind of approach does not consider the correlation between the patch and its surrounding. As a result, they can only detect the anomalous patches that have some flaws inside, but fail to handle the ordinary patches that are placed in the wrong position, like the second example shown in Fig. 1a where the top cable should be in green color.

In this work, we propose a novel framework for unsupervised anomaly localization with due consideration to *both the global and local information*. At the training stage, given a normal image, we randomly crop a patch and introduce a Local-Net and Global-Net to extract features from this patch and its surrounding respectively, as shown in Fig. 1b. Concretely, we develop the global feature to match the local feature, encouraging the Global-Net to conjecture the feature of the missing patch based on the context. For this purpose, we learn Global-Net jointly with an Inconsistency Anomaly Detection (IAD) head and a Distortion Anomaly Detection (DAD) head, leading to a fused metric to better measure the similarity between the global and local features. At the inference stage, a scoring function developed from the multi-head feature comparison is capable of producing an adequate score map from a test image to help localize anomalous regions, as shown in Fig. 1a. In summary, our contributions are:

- We propose a novel unsupervised anomaly localization approach by collecting both the global and local information from training data. In particular, the local feature extracted from an image patch is regressed by the global feature extracted from its surrounding.
- We introduce the multi-head feature comparison where the IAD head targets at spotting the mismatch between patches and surroundings while the DAD head aims to detect subtle defects occurring in the patch. We further derive a scoring function from this multi-head design, facilitating high-precision anomaly localization.
- We achieve state-of-the-art performances on a couple of real-world datasets, significantly surpassing existing methods. For example, on the recent MVTec AD dataset [6], which is specifically designed to benchmark the anomaly localization task, we beat the second competitor by 4.7% improvement (*i.e.*, from 91.4% to 96.1%) under the per-region-overlap (PRO) metric [6].

2. Related Work

We summarize existing anomaly detection methods into two categories: compress-based and reconstruct-based.

Compress-based. Compress-based methods typically project raw images [7, 10, 22, 48] or high-dimensional features [38, 43] into a low-dimensional feature space, where normal and abnormal examples are much more distinguishable. For this purpose, SVDD [38] considers a distance-minimize objective, which extracts the shared feature from normal examples while avoiding them to be the same. Based on [38], Yi and Yoon [48] propose a patch-based SVDD that contains multiple cores rather than a single core in [38], enabling anomaly segmentation. Bergmann *et al.* [7] utilize a pre-trained teacher network to embed image patches and estimate the anomaly score with a collection of student networks. Kwon *et al.* [22] argue that normal and abnormal images are more distinguishable at the backward gradient space. Some other methods [42, 41, 34] discover the underlying data distribution with the help of Generative Adversarial Networks (GANs). After training a GAN on normal data, they determine whether a test sample is anomalous based on the discriminator output.

Reconstruct-based. Reconstruct-based approaches assume that normal images can be described by a unified distribution in image space. They commonly utilize Auto-Encoders [40, 11, 20, 50, 17, 14, 36, 32] or GANs [49, 46, 37, 39, 5] to learn the distribution underlying normal data and then make the decision based on whether a test sample can be well recovered or not. To improve the reconstruction sharpness, prior work [11, 12] introduces the skip connections that provide the decoder with more spatial information. Gong *et al.* [14] and Park *et al.* [36] believe that limiting the generalization ability is vital in finding novel images, and hence propose a memory-based auto-encoder to reconstruct images from features. Xia *et al.* [46] learn the image reconstruction from the image segmentation map instead of the original input. Pseudo-anomalies are also widely used to improve the performance of anomaly detection. Zaheer *et al.* [49] take the images reconstructed by old generators as another kind of anomalies. Huang *et al.* [20] apply data augmentation to improve the attribute restoration ability of the reconstruction model.

Discussion. Different from the above approaches, which learn representation either from a patch (local) or from the entire image (global), our algorithm gives due consideration to both the global and the local information and makes decision based on the comparison between local and global features. A recent work [35] also proposes to aggregate local and global information for anomaly detection. In particular, it employs the embedding learned from the compress-based method as the global information and treat the image recovered by the reconstructed-based method as the local information. Differently, we extract the local information from an image patch and the global information from its surrounding, forming an one-on-one matching.

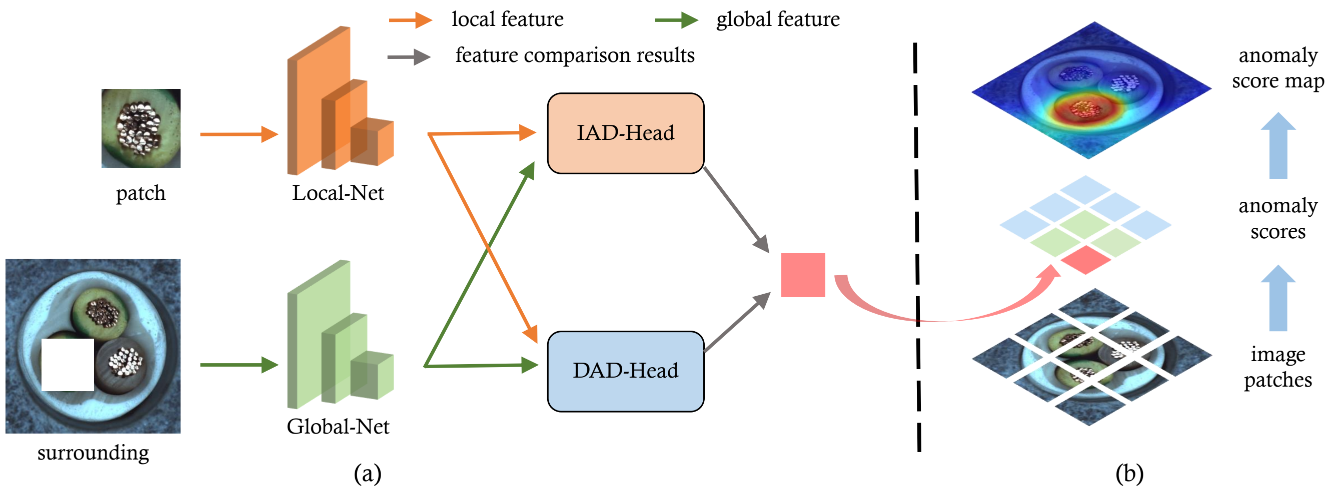


Figure 2. **Anomaly Localization Framework.** (a) At the training stage, a Local-Net and Global-Net are employed to extract features from an image patch and its surrounding respectively. The Global-Net is jointly learned with an Inconsistency Anomaly Detection (IAD) head and a Distortion Anomaly Detection (DAD) head to mimic the output from the Local-Net. (b) At the inference stage, a scoring function is developed based on the feature comparison results produced by the IAD-head and DAD-head. Anomaly scores corresponding to different patches are aggregated together into an anomaly score map for anomaly localization.

3. Method

Given a training set of normal images $\{\mathbf{I_1}, \mathbf{I_2}, \dots, \mathbf{I_n}\}$, and a test set containing both anomaly and anomaly-free images $\{\mathbf{I'_1}, \mathbf{I'_2}, \dots, \mathbf{I'_m}\}$, the goal is to identify test images as normal or abnormal, and further localize the anomalous regions in abnormal samples.

In the following sections, we first discuss how to extract the local and global features from an image (Sec. 3.1), then explain the Inconsistency Anomaly Detection head (IAD-head) and the Distortion Anomaly Detection head (DAD-head) used for feature comparison (Sec. 3.2), and finally introduce the way to produce an anomaly score map from a test image at the inference stage (Sec. 3.3).

3.1. Local and Global Feature Extraction

In this part, we introduce how to extract the local and the global feature from a patch and its surrounding, as shown in Fig. 2a.

Local Feature Extraction. We use Local-Net, a light-weight neural network, to embed image patches into local features. Since shallow networks distilled from deep networks trained on classification tasks show promising results in anomaly detection [7, 33, 32], we distill Local-Net from pre-trained ResNet-18 [19]. Concretely, Local-Net is first distilled on ImageNet [13], and then fine-tuned on a particular training set. Knowledge distillation loss [7] and compactness loss [44] are utilized in distillation and fine-tuning. Here, the knowledge distillation loss is defined as

$$l_k = ||\mathcal{D}(\mathcal{L}(\mathbf{p})) - \mathcal{R}(\mathbf{p})||_2^2, \tag{1}$$

where $\mathbf{p}$ is the image patch and $||\cdot||_2$ denotes the ℓ_2 norm. $\mathcal{L}(\cdot)$ and $\mathcal{R}(\cdot)$ stand for Local-Net and the teacher model (*i.e.*, the pre-trained ResNet-18) respectively. $\mathcal{D}(\cdot)$ is a decoder to ensure $\mathcal{L}(\cdot)$ and $\mathcal{R}(\cdot)$ to have same output dimension.

The compactness loss is formulated as

$$l_c = \sum_{i \neq j} c_{ij}, \tag{2}$$

where c_{ij} represents the (i, j) entry in the correlation matrix over the Local-Net outputs $\mathcal{L}(\mathbf{p})$ within a mini-batch.

Overall, the Local-Net is optimized with

$$l_{local} = \lambda_k l_k + \lambda_c l_c, \tag{3}$$

where λ_k and λ_c are loss weights to balance different terms. After distillation and fine-tuning, the local feature $\mathbf{Z}_l$ can be extracted by the Local-Net as

$$\mathbf{Z}_l = \mathcal{L}(\mathbf{p}). \tag{4}$$

Note that the learning of the Local-Net is referred as pre-training in our framework. During the training of Global-Net and DAD-head, as discussed below, parameters of the Local-Net are fixed.

Global Feature Extraction. Another deep model, named Global-Net, is employed to extract the global feature from the surrounding of the patch $\mathbf{I} \backslash \mathbf{p}$. To prevent the local feature from disturbing the global feature, we apply partial convolution [25] to our Global-Net. Specifically, for every convolutional layer in the Global-Net, the convolution operation at every location is formulated as

$$x' = \begin{cases} \mathbf{W^T}(\mathbf{X} \odot \mathbf{M})\frac{\text{sum}(\mathbf{1})}{\text{sum}(\mathbf{M})} + b, & \text{if sum}(\mathbf{M}) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Here, $\odot$ represents the element-wise product, $\mathbf{X}$ denotes the input feature map, and $\mathbf{M}$ is the binary mask in current layer. For every pooling layer, feature map is updated as the normal pooling, and $\mathbf{M}$ is refreshed with the binary version of the mask's pooling result. The initial mask $\mathbf{M_0}$ is defined as a binary matrix where the patch's pixels are zero and the others are one. Therefore, Global-Net $\mathcal{G}(\cdot)$ can extract global feature $\mathbf{Z}_g$ without peeking at the patch, as formulated below:

$$\mathbf{Z}_g = \mathcal{G}(\mathbf{I}, \mathbf{M_0}). \tag{6}$$

3.2. Anomaly Detection Heads

In this part, we introduce two anomaly detection heads, *i.e.*, the Inconsistency Anomaly Detection head (IAD-head) and the Distortion Anomaly Detection head (DAD-head). As shown in Fig. 2a, IAD-head and DAD-head accept the local feature and the global feature extracted from the patch and its surrounding and make comparison between these two features.

Inconsistency Anomaly Detection Head. Inconsistency anomaly detection head (IAD-head) is designed to detect the inconsistency between the local feature $\mathbf{Z}_l$ and the global feature $\mathbf{Z}_g$ with

$$l_{\text{IAD}} = \frac{1}{n}||\mathbf{Z}_l - \mathbf{Z}_g||_2^2, \tag{7}$$

where n is the dimension of both the local feature and the global feature.

We assume that in normal images local and global features are consistent, while in abnormal images the situation is the contrary. Therefore, in the training process, l_{IAD} is utilized as a loss to close the distance between $\mathbf{Z}_l$ and $\mathbf{Z}_g$. During inference, l_{IAD} serves as a scoring function to indicate the global-local inconsistency lying in the patch, which will be discussed in Sec. 3.3.

Distortion Anomaly Detection Head. Distortion anomaly detection head (DAD-head) is a trainable head, which aims to detect the distortions in images, *e.g.*, bent grids and cut carpets. Compared with the IAD-head that focuses on the mismatch between the patch and its surrounding, the DAD-head is capable of spotting tiny defects localized in the patch. Concretely, the DAD-head exploits a number of fully-connected layers to determine whether distortions exist in the patch. In addition to the original patch $\mathbf{p}$, we introduce a negative patch $\mathbf{p}^-$ following [28, 11], which is generated by adding a random small stain on $\mathbf{p}$. The reason for constructing negative patches in this way is to maintain the majority of the patch and introduce only tiny differences, encouraging the DAD-head to spot small distortions. The features extracted from $\mathbf{p}$ and $\mathbf{p}^-$ are equiprobably fed into the DAD-head together with the global feature $\mathbf{Z}_g$. Then the DAD-head determines whether the input local feature is $\mathbf{Z}_l$ or $\mathbf{Z}_l^-$ by producing a positive probability

$$p = \mathcal{C}(\mathbf{Z}^*, \mathbf{Z}_g), \tag{8}$$

where $\mathcal{C}(\cdot, \cdot)$ is the classification network in the DAD-head. $\mathbf{Z}^*$ can be either local feature $\mathbf{Z}_l$ or negative local feature $\mathbf{Z}_l^-$. During training, the classifier in the DAD-head is supervised by a cross-entropy loss as

$$l_{\text{DAD}} = -(y\log(p) + (1-y)\log(1-p)), \tag{9}$$

where y is the target output of the classifier, *i.e.*, 0 for the positive patch and 1 for the negative patch.

Training Objective. The total training loss for the Global-Net and the DAD-head is

$$l = l_{\text{IAD}} + \lambda_t l_{\text{DAD}}, \tag{10}$$

where λ_t is a loss weight to balance different energies. Intuitively, l_{IAD} guides the Global-Net to imagine the local distribution, while l_{DAD} encourages the Global-Net to learn a more distinguishable representation. Meanwhile, the DAD-head is trained to find the subtle differences between normal and distorted patches.

3.3. Anomaly Localization

Scoring Function. At the inference stage, we feed local feature $\mathbf{Z}_l$ and global feature $\mathbf{Z}_g$ into the IAD-head to generate the inconsistency anomaly score

$$s_{\text{IAD}} = \frac{1}{n}||\mathbf{Z}_l - \mathbf{Z}_g||_2^2. \tag{11}$$

We also feed them into the DAD-head to produce the distortion anomaly score

$$s_{\text{DAD}} = 1 - \mathcal{C}(\mathbf{Z}_l, \mathbf{Z}_g). \tag{12}$$

Finally, our scoring function integrates these two scores:

$$s = \lambda_s s_{\text{IAD}} + (1-\lambda_s)s_{\text{DAD}}, \tag{13}$$

where λ_s is a hyper-parameter to balance the inconsistency anomaly score s_{IAD} and the distortion anomaly score s_{DAD}. We set $\lambda_s = 0.8$ in our experiments. Detailed study on λ_s can be found in Sec. 4.4.

Anomaly Score Map. With the scoring function to assign the anomaly score to a particular patch, we further propose a pipeline to aggregate the anomaly scores for different patches into an anomaly score map. Concretely, we generate image patches one after another, and organize them in a raster-scan order, as shown in Fig. 2b. Overlap is allowed between two adjacent image patches. We assign an anomaly score for each patch with the scoring function in Eq. (13), and construct an anomaly score map for the entire image with the inverse distance weighted (IDW) interpolation.

4. Experiments

In this section, we study our model's performance on pixel-level anomaly detection task (Sec. 4.2), and image-level one-class classification task (Sec. 4.3). Qualitative and quantitative results show our approach attains state-of-the-art performance compared with other methods.

4.1. Datasets

MVTec AD. MVTec AD [6] is a real-world industrial image anomaly detection dataset with 5354 high-resolution images in 15 categories. The training set has 3629 normal images, and the test set contains 1725 normal or abnormal images. The ground truth in the test set includes both labels and anomaly masks. We follow the original dataset split of MVTec AD, *i.e.*, use only anomaly-free images in training, and test on both normal and abnormal images.

CIFAR-10. CIFAR-10 [21] includes 60000 tiny images with 10 classes. In each class, 5000 images are used for training, and the other 1000 images for testing. We follow the protocol in GradCon [22] to split dataset for one-class classification task. Specifically, based on the original training-test split of CIFAR-10, we construct the training set from images of one class as inliers, and build the test set from inlier images and the same number of outlier images randomly sampled from other classes.

4.2. Pixel-level Anomaly Localization

We evaluate our approach's localization ability on the pixel-level anomaly detection task. Both qualitative and quantitative results are provided.

Setup. For pre-training Local-Net, following [7], we first distill Local-Net from pre-trained ResNet-18 on ImageNet [13], and then fine-tune it into the specific category in MVTec AD with the same loss as that in distillation. When training Global-Net and DAD-head, we randomly crop patch $\mathbf{p}$ from the image, and add some random stains on the patch to produce $\mathbf{p}^-$. Then Global-Net and DAD-head are trained with the loss in Eq. (10). At inference stage, image patches are cropped in a roster-scan order, and the anomaly score s for each patch is estimated according to Eq. (13) with $\lambda_s = 0.8$. Finally, anomaly score maps of images are constructed as discussed in Sec. 3.3. More implementation details can be found in *Supplementary Material*.

Baselines. We have two parts of competitors. The first part is baselines in [6, 7], including the 1-NN classifier [3], the One-Class SVM (OCSVM) [43], the K-Means classifier [29], deterministic autoencoder with l_2-reconstruction error as the anomaly score (l_2-AE) [16] , variational autoencoder with reconstruction probability as the anomaly score (VAE) [4], CNN-Feature Dictionary (CNN-FD) [32], the SSIM-Autoencoder (SSIM-AE) [8] and AnoGAN [42]. The results for above methods are all reported in [6, 7]. The second part is recently peer-reviewed models, including teacher-student (TS) [7], Visually Explained Variational Autoencoder (VAVAE) [26], Superpixel Masking and Inpainting (SMAI) [23], Gradient Descent Reconstruction with VAEs (GDR) [12], Encoding Structure-Texture Relation with P-Net for AD (P-Net) [51]. The results for models in the second part are reported in the original papers.

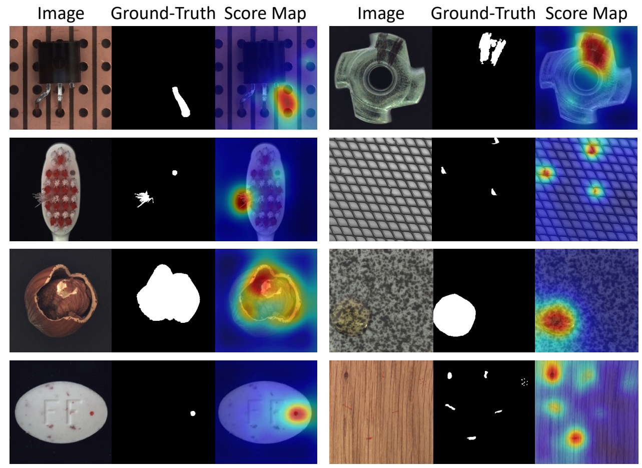


Figure 3. Qualitative anomaly localization results on MVTec AD dataset [6]. For each example, the images from left to right are the defective image, the ground-truth, and the anomaly score map produced by our algorithm. Zoom in for details.

Qualitative Results. Fig. 3 shows our qualitative results in MVTec AD [6]. Our approach satisfactorily addresses all kinds of anomalies and further locates the subtle defects. More anomaly localization results can be found in *Supplementary Material*.

To further illustrate the importance of the global and local feature comparison in anomaly localization, in Fig. 4 we qualitatively compare our model with TS [7], which only utilizes the local feature of the patch. When encountering some hard anomalies which seems totally normal in any single patch, *e.g.*, misplacement, swapping and bend in Fig. 4, local feature based models cannot detect the anomalies, while our method handles them excellently.

Quantitative Results in PRO. We follow the protocol in [6, 7], and use the per-region-overlap (PRO) as the evaluation metric. Unlike other per-pixel metrics, PRO weights ground-truth regions equally regardless of region sizes [7]. Specifically, as described in TS [7], we increase the threshold until the average per-pixel false positive rate reaches 30%. For each threshold, we calculate the PRO value, *i.e.*, the average ratio of the area detected as anomalous in each anomaly connected component. And the final metric is the normalized area under the PRO curve.

We compare our results with all the baselines that report the PRO metric. Tab. 1 shows the comparison results. Our method outperforms the other methods by a wide margin.

Table 1. Comparison results among different anomaly detection methods in the **pixel-level anomaly localization task on MVTec AD dataset** [6]. Competitors include 1-NN [3], OC-SVM [43], K-Means [29], l_2-AE [16], VAE [4], SSIM-AE [8], AnoGAN [42], CNN-FD [32] and TS [7]. The results of baselines are borrowed from [6, 7]. **Per-region-overlap (PRO)** [7] is used as the evaluation metric.

	Category	1-NN	OC-SVM	K-Means	l_2-AE	VAE	SSIM-AE	AnoGAN	CNN-FD	TS	Ours
Texture	Carpet	0.512	0.355	0.253	0.456	0.501	0.647	0.204	0.469	0.879	**0.977**
	Grid	0.228	0.125	0.107	0.582	0.224	0.849	0.226	0.183	**0.952**	0.932
	Leather	0.446	0.306	0.308	0.819	0.635	0.561	0.378	0.641	**0.945**	0.909
	Tile	0.822	0.722	0.779	0.897	0.870	0.175	0.177	0.797	**0.946**	0.883
	Wood	0.502	0.336	0.411	0.727	0.628	0.605	0.386	0.621	0.911	**0.941**
Object	Bottle	0.898	0.850	0.495	0.910	0.897	0.834	0.620	0.742	0.931	**0.968**
	Cable	0.806	0.431	0.513	0.825	0.654	0.478	0.383	0.558	0.818	**0.980**
	Capsule	0.631	0.554	0.387	0.862	0.526	0.860	0.306	0.306	**0.968**	0.960
	Hazelnut	0.861	0.616	0.698	0.917	0.878	0.916	0.698	0.844	**0.965**	0.962
	Metal Nut	0.705	0.319	0.351	0.830	0.576	0.603	0.320	0.358	0.942	**0.967**
	Pill	0.725	0.544	0.514	0.893	0.769	0.830	0.776	0.460	0.961	**0.978**
	Screw	0.604	0.644	0.550	0.754	0.559	0.887	0.466	0.277	0.942	**1.000**
	Toothbrush	0.675	0.538	0.337	0.822	0.693	0.784	0.749	0.151	0.933	**0.961**
	Transistor	0.680	0.496	0.399	0.728	0.626	0.725	0.549	0.628	0.666	**0.999**
	Zipper	0.512	0.355	0.253	0.839	0.549	0.665	0.467	0.703	0.951	**0.992**
	Mean	0.640	0.479	0.423	0.790	0.639	0.694	0.443	0.515	0.914	**0.961**

Specifically, in comparison with TS [7], the state-of-the-art competitor, our model exceeds greatly in categories such as the cable and transistor. This is identical with our observation in Fig. 4, for TS fails in the anomaly cases involved with global and local feature comparison.

Quantitative Results in Pixel-level AUROC. We further provide quantitative results on pixel-level AUROC [6]. The comparison with all the baselines that report pixel-level AUROC is shown in Tab. 2. Our model substantially surpasses other methods in pixel-wise AUROC ($\geq$ 2%). This conclusion is consistent with that in PRO results above.

Discussion. Recall that, for each category in the MVTec AD dataset, we train a separate model only on the normal data with exactly the same configuration (*e.g.*, hyper-parameters like learning rate, training epochs, *etc.*). Results in Tab. 1 and Tab. 2 show that our approach is generalizable to various types of data, suggesting strong robustness.

4.3. Image-level Anomaly Detection

To further prove that our method is able to handle various tasks in anomaly detection, our model is applied to image-level anomaly detection task. Here we present the results of unsupervised one-class classification on CIFAR-10 [21].[1]

Setup. During pre-training, the distillation process is the same as that in Sec. 4.2, and when fine-tuning on each category of CIFAR-10, we resize each image into the patch size, which functions as the image patch in Sec. 4.2.

During training Global-Net and DAD-head, every single image is reshaped to the patch size and the image size, denoted as $\mathbf{I_L}$ and $\mathbf{I_G}$, respectively. Then we input $\mathbf{I_L}$ and $\mathbf{I_G}$ into Local-Net and Global-Net to generate the local and global feature. The negative local feature is produced by feeding a random different image (resized into the patch size) into Local-Net. The other procedures are the same as those in the experiments on MVTec AD. Further details can be found in *Supplementary Material*.

[1]Our approach also excels in image-level anomaly detection task on MVTec AD [6], which is discussed in *Supplementary Material*.

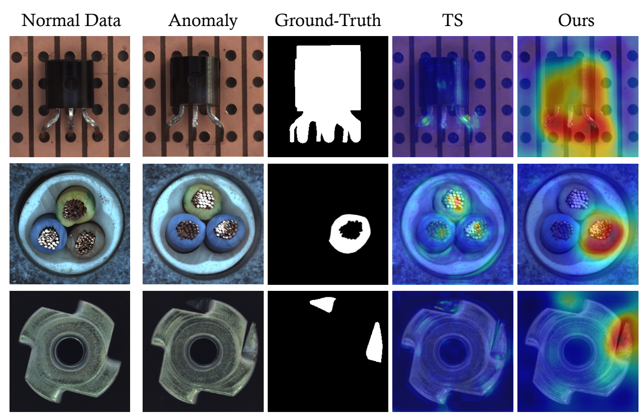


Figure 4. Qualitative comparisons on MVTec AD dataset [6] between TS [7], which considers only local patches for anomaly localization, and our approach. Through adequately utilizing both the global and the local information, we manage to identify the transistor shift (first row), the cable swap (second row), and metal nut bend (third row) anomalies that are omitted by TS.

Baselines. Our competitors include OC-SVM [43], KDE [9], l_2-AE [16], VAE [4], pixelCNN [45], LSA [1], AnoGAN [42], DSVDD [38], OCGAN [37], and GradCon [22]. The results of baselines are borrowed from [37, 22].

Quantitative Results in AUROC. Image-level AUROC comparisons with baselines on CIFAR-10 are shown in

Table 2. Comparison results among different anomaly detection methods in the **pixel-level anomaly localization task on MVTec AD dataset** [6]. Competitors include SSIM-AE [8], l_2-AE [16], AnoGAN [42], CNN-FD [32], VEVAE [26], SMAI [23], GDR [12] and P-Net [51]. The results of SSIM-AE, l_2-AE, AnoGAN and CNN-FD are borrowed from the MVTec AD paper [6], and the results of VEVAE, SMAI, GDR and P-Net are reported in their original papers. **Pixel-level AUROC** is utilized as the evaluation metric.

	Category	SSIM-AE	l_2-AE	AnoGAN	CNN-FD	VEVAE	SMAI	GDR	P-Net	Ours
Texture	Carpet	0.87	0.59	0.54	0.72	0.78	0.88	0.74	0.57	**0.96**
	Grid	0.94	0.90	0.58	0.59	0.73	0.97	0.96	**0.98**	0.78
	Leather	0.78	0.75	0.64	0.87	**0.95**	0.86	0.93	0.89	0.90
	Tile	0.59	0.51	0.5	0.93	0.80	0.62	0.65	**0.97**	0.80
	Wood	0.73	0.73	0.62	0.91	0.77	0.80	0.84	**0.98**	0.81
Object	Bottle	0.93	0.86	0.86	0.78	0.87	0.86	0.92	**0.99**	0.93
	Cable	0.82	0.86	0.78	0.79	0.90	0.92	0.91	0.70	**0.94**
	Capsule	**0.94**	0.88	0.84	0.84	0.74	0.93	0.92	0.84	0.90
	Hazelnut	0.97	0.95	0.87	0.72	**0.98**	0.97	**0.98**	0.97	0.84
	Metal Nut	0.89	0.86	0.76	0.82	**0.94**	0.92	0.91	0.79	0.91
	Pill	0.91	0.85	0.87	0.68	0.83	0.92	**0.93**	0.91	**0.93**
	Screw	0.96	0.96	0.8	0.87	0.97	0.96	0.95	**1.00**	0.96
	Toothbrush	0.92	0.93	0.90	0.77	0.94	0.96	**0.99**	**0.99**	0.96
	Transistor	0.90	0.86	0.80	0.66	0.93	0.85	0.92	0.82	**1.00**
	Zipper	0.88	0.77	0.78	0.76	0.78	0.90	0.87	0.90	**0.99**
	Mean	0.86	0.82	0.74	0.78	0.86	0.89	0.89	0.89	**0.91**

Tab. 3. Our method considerably exceeds the second state-of-the-art, GradCon [22], by 4.1% in image-level AUROC. The results show that our model is adaptive to different settings of anomaly detection.

4.4. Analysis on Multi-head Feature Comparison

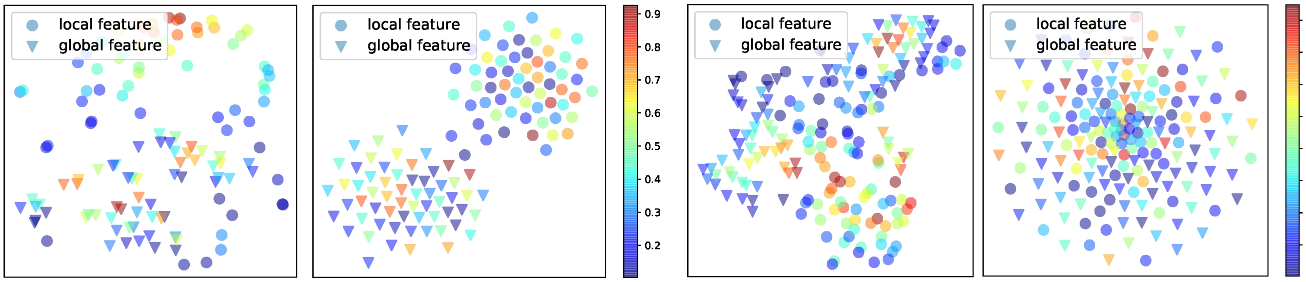
In this section, we first conduct ablation study on IAD-head and DAD-head, and further analyze the different discriminative ability in view of these two heads.

Ablation Study. To evaluate the effectiveness of joint scoring function of the IAD-head and DAD-head, *i.e.*, Eq. (13), we vary λ_s from 0.0 to 1.0, and calculate the PRO metric on each category. According to Eq. (13), larger λ_s results in greater proportion of s_{IAD} in the synthetic anomaly score s, while the proportion of s_{DAD} decreases accordingly. In the extreme case, s will degenerate into s_{DAD} or s_{IAD} if λ_s equals 0 or 1, respectively.

Tab. 4 shows the PRO results of our method under different values of λ_s. Overall, multi-head scoring function performs better than single-head ones, with 4.9% and 0.7% increase compared to single DAD-head and single IAD-head scoring function, respectively. More concretely, categories of textures (*e.g.*, carpet, grid and wood) enjoy greater improvement than those of objects after applying multi-head scoring function. We infer this is because much more repetitive patterns are contained in textures than in objects, and multi-head strategy is conducive to a more distinguishable feature representation, enhancing our model's anomaly localization ability greatly on textures.

However, it should be noticed that in Tab. 4, the overall performances of single DAD-head are worse than those of single IAD-head for both texture and object classes. The reason might be that the performance of DAD-head is highly correlative with the way to construct negative patches. That is, if the construction approach is similar to the real anomalies, DAD-head might perform better, otherwise might perform worse. In our experiment, to involve *no* priori knowledge, we utilize a simple negative patch construction approach and apply the same method to all the different classes, yet DAD-head still shows great potential in cooperation with IAD-head. To further improve performance of DAD-head, we encourage users to modify the negative patch construction way according to the real-world scenarios.

Feature Visualization. To better understand the different discriminative abilities in view of IAD-head and DAD-head, we visualize the global and local features under the metrics of IAD-head and DAD-head. Fig. 5 shows the feature visualization for the texture and object. We randomly crop 800 patches from both texture-type (*i.e.*, carpet) and object-type (*i.e.*, cable) defective images, and utilize t-SNE [31] to visualize the features of the abnormal patches and their surroundings. As shown in Fig. 5a, compared with IAD-head, DAD-head is more discriminative on the texture, presenting a clear boundary between anomalies' local and global features. On the contrary, Fig. 5b illustrates that IAD-head seperates abnormal local and global features better than DAD-head on the object. Different discriminative abilities of IAD-head and DAD-head ensure excellent performance of the multi-head anomaly detection mechanism in various situations.



IAD-Head DAD-Head

(a) Texture feature visualization

IAD-Head DAD-Head

(b) Object feature visualization

Figure 5. **Visualization of the global and local features** under the metrics provided by the two anomaly detection heads, *i.e.*, IAD-head and DAD-head. Abnormal patches selected from 800 patches in MVTec AD dataset [6], together with their surroundings, are used for both texture-type (*i.e.*, carpet) and object-type (*i.e.*, cable) visualization. The color map indicates the ratio of the ground-truth anomalous area to each individual patch. Triangles and circles stand for the global and local features respectively. It turns out that texture features are more distinguishable in the view of DAD-head while object features can be better differentiated by IAD-head.

Table 3. Comparison results among different one-class classification methods in the **image-level anomaly detection task on CIFAR-10** [21]. Competitors include OC-SVM [43], KDE [9], l_2-AE [16], VAE [4], PixelCNN [45], LSA [1], AnoGAN [42], DSVDD [38], OCGAN [37] and GradCon [22]. The results of baselines are borrowed from [22, 37]. **Image-level AUROC** is utilized as the evaluation metric.

Normal Class	OC-SVM	KDE	l_2-AE	VAE	PixelCNN	LSA	AnoGAN	DSVDD	OCGAN	GradCon	Ours
Airplane	0.630	0.658	0.411	0.634	0.788	0.735	0.671	0.617	0.757	0.760	**0.791**
Automobile	0.440	0.520	0.478	0.442	0.428	0.580	0.547	0.659	0.531	0.598	**0.703**
Bird	0.649	0.657	0.616	0.640	0.617	0.690	0.529	0.508	0.640	0.648	**0.675**
Cat	0.487	0.497	0.562	0.497	0.574	0.542	0.545	0.591	**0.620**	0.586	0.561
Deer	0.735	0.727	0.728	0.743	0.511	**0.761**	0.651	0.609	0.723	0.733	0.739
Dog	0.500	0.496	0.513	0.515	0.571	0.546	0.603	**0.657**	0.620	0.603	0.638
Frog	0.725	**0.758**	0.688	0.745	0.422	0.751	0.585	0.677	0.723	0.684	0.732
Horse	0.533	0.564	0.497	0.527	0.454	0.535	0.625	0.673	0.575	0.567	**0.674**
Ship	0.649	0.680	0.487	0.674	0.715	0.717	0.758	0.759	**0.820**	0.784	0.814
Truck	0.508	0.540	0.378	0.416	0.426	0.548	0.665	0.731	0.554	0.678	**0.722**
Mean	0.586	0.610	0.536	0.583	0.551	0.641	0.618	0.648	0.657	0.664	**0.705**

Table 4. **Results of ablation study on the multi-head scoring function**. λ_s in Eq. (13) varies from 0.0 to 1.0 with step 0.2. The categories having better performance with multi-head scoring functions than with single-head ones is highlighted in boldface. **Per-region-overlap (PRO)** is used as the evaluation metric.

Category	0.0	0.2	0.4	0.6	0.8	1.0
Carpet	0.963	0.966	0.968	0.972	**0.977**	0.965
Grid	0.868	0.882	0.907	**0.932**	**0.932**	0.894
Leather	0.899	0.902	0.909	**0.911**	0.909	0.896
Tile	**0.951**	0.920	0.905	0.895	0.883	0.874
Wood	0.832	0.900	0.925	**0.941**	**0.941**	0.920
Bottle	0.928	0.956	0.965	0.966	**0.968**	0.965
Cable	0.941	0.980	**0.991**	0.989	0.980	0.961
Capsule	0.843	0.884	0.910	0.937	0.960	**0.978**
Hazelnut	0.937	0.937	0.949	0.957	0.962	**0.964**
Metal Nut	0.921	0.930	0.941	0.956	0.967	**0.971**
Pill	0.875	0.935	0.958	0.971	**0.978**	**0.978**
Screw	0.948	0.986	0.996	0.999	**1.000**	**1.000**
Toothbrush	0.897	0.913	0.934	0.948	0.961	**0.966**
Transistor	0.883	0.935	0.972	0.993	**0.999**	0.997
Zipper	**0.995**	**0.995**	**0.995**	0.994	0.992	0.974
Mean	0.912	0.935	0.948	0.957	**0.961**	0.954

5. Conclusion and Discussion

In this work, we propose an unsupervised anomaly localization approach with due consideration to both the global and the local information from an image. Two anomaly detection heads are introduced to sufficiently spot the discrepancy between global and local features. With the scoring function developed from such multi-head design, we achieve high-precision anomaly localization, significantly surpassing state-of-the-art alternatives.

However, there still remains some future work worth exploration. On one hand, our approach uses a fixed patch size regardless of the anomaly type. To further improve the robustness under various anomaly scales, techniques such as score map averaging [7] and feature pyramid [24] could be considered. On the other hand, Local-Net in our work is distilled from a deep model pre-trained on ImageNet. Self-supervised learning methods [18, 15] might be of benefit to getting a better representation on specific datasets. Nevertheless, our approach sheds light on a promising direction by relating the individual patches with their surroundings for the anomaly localization task.

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 6, 8

[2] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian Conf. Comput. Vis.*, 2018. 1

[3] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 1992. 5, 6

[4] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2015. 5, 6, 8

[5] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI Brainlesion Workshop*, 2018. 1, 2

[6] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2, 5, 6, 7, 8

[7] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 2, 3, 5, 6, 8

[8] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018. 2, 5, 6, 7

[9] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006. 6, 8

[10] Wen-Hsuan Chu and Kris M Kitani. *Neural Batch Sampling with Reinforcement Learning for Semi-Supervised Anomaly Detection*. PhD thesis, Carnegie Mellon University Pittsburgh, 2020. 1, 2

[11] Anne-Sophie Collin and Christophe De Vleeschouwer. Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise. *arXiv preprint arXiv:2008.12977*, 2020. 1, 2, 4

[12] David Dehaene, Oriel Frigo, Sébastien Combrexelle, and Pierre Eline. Iterative energy-based projection on a normal data manifold for anomaly localization. *arXiv preprint arXiv:2002.03734*, 2020. 2, 5, 7

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. 3, 5

[14] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Int. Conf. Comput. Vis.*, 2019. 1, 2

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 8

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2006. 5, 6, 7, 8

[17] Simon Hawkins, Hongxing He, Graham Williams, and Rohan Baxter. Outlier detection using replicator neural networks. In *International Conference on Data Warehousing and Knowledge Discovery*, 2002. 2

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 8

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3

[20] Chaoqin Huang, Fei Ye, Jinkun Cao, Maosen Li, Ya Zhang, and Cewu Lu. Attribute restoration framework for anomaly detection. *arXiv preprint arXiv:1911.10676*, 2019. 2

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 6, 8

[22] Gukyeong Kwon, Mohit Prabhushankar, Dogancan Temel, and Ghassan AlRegib. Backpropagated gradient representations for anomaly detection. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2, 5, 6, 7, 8

[23] Zhenyu Li, Ning Li, Kaitao Jiang, Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Superpixel masking and inpainting for self-supervised anomaly detection. In *Brit. Mach. Vis. Conf.*, 2020. 5, 7

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 8

[25] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Eur. Conf. Comput. Vis.*, 2018. 3

[26] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. Towards visually explaining variational autoencoders. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 5, 7

[27] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1

[28] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. *arXiv preprint arXiv:2007.01760*, 2020. 1, 4

[29] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 1982. 5, 6

[30] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Int. Conf. Comput. Vis.*, 2013. 1

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008. 7

[32] Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Anomaly detection in nanofibrous materials by cnn-based self-similarity. *Sensors*, 2018. 1, 2, 3, 5, 6, 7

[33] Tiago S Nazare, Rodrigo F de Mello, and Moacir A Ponti. Are pre-trained cnns good feature extractors for anomaly detection in surveillance videos? *arXiv preprint arXiv:1811.08495*, 2018. 3

[34] Duc Tam Nguyen, Zhongyu Lou, Michael Klar, and Thomas Brox. Anomaly detection with multiple-hypotheses predictions. In *Int. Conf. Mach. Learn.*, 2019. 2

[35] Lihai Nie, Laiping Zhao, and Keqiu Li. Glad: Global and local anomaly detection. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020. 2

[36] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 2

[37] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2, 6, 8

[38] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Int. Conf. Mach. Learn.*, 2018. 1, 2, 6, 8

[39] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[40] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, 2014. 2

[41] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 2019. 1, 2

[42] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, 2017. 1, 2, 5, 6, 7, 8

[43] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Adv. Neural Inform. Process. Syst.*, 1999. 1, 2, 5, 6, 8

[44] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 3

[45] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Adv. Neural Inform. Process. Syst.*, 2016. 6, 8

[46] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. *arXiv preprint arXiv:2003.08440*, 2020. 1, 2

[47] Sri Kalyan Yarlagadda, David Güera, Paolo Bestagini, Fengqing Maggie Zhu, Stefano Tubaro, and Edward J Delp. Satellite image forgery detection and localization using gan and one-class classifier. *Electronic Imaging*, 2018. 1

[48] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Asian Conf. Comput. Vis.*, 2020. 1, 2

[49] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 2

[50] Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017. 2

[51] Kang Zhou, Yuting Xiao, Jianlong Yang, Jun Cheng, Wen Liu, Weixin Luo, Zaiwang Gu, Jiang Liu, and Shenghua Gao. Encoding structure-texture relation with p-net for anomaly detection in retinal images. *arXiv preprint arXiv:2008.03632*, 2020. 1, 2, 5, 7