

Diversity-Measurable Anomaly Detection

Wenrui Liu^{1,2}, Hong Chang^{1,2}, Bingpeng Ma², Shiguang Shan^{1,2}, Xilin Chen^{1,2}

¹Institute of Computing Technology, Chinese Academy of Sciences, China

²University of Chinese Academy of Sciences, China

wenrui.liu@vipl.ict.ac.cn, {changhong, sgshan, xlchen}@ict.ac.cn, bpma@ucas.ac.cn

Abstract

Reconstruction-based anomaly detection models achieve their purpose by suppressing the generalization ability for anomaly. However, diverse normal patterns are consequently not well reconstructed as well. Although some efforts have been made to alleviate this problem by modeling sample diversity, they suffer from shortcut learning due to undesired transmission of abnormal information. In this paper, to better handle the tradeoff problem, we propose **Diversity-Measurable Anomaly Detection (DMAD)** framework to enhance reconstruction diversity while avoid the undesired generalization on anomalies. To this end, we design **Pyramid Deformation Module (PDM)**, which models diverse normals and measures the severity of anomaly by estimating multi-scale deformation fields from reconstructed reference to original input. Integrated with an information compression module, PDM essentially decouples deformation from prototypical embedding and makes the final anomaly score more reliable. Experimental results on both surveillance videos and industrial images demonstrate the effectiveness of our method. In addition, DMAD works equally well in front of contaminated data and anomaly-like normal samples.

1. Introduction

Visual anomaly detection is a fundamental and important problem in computer vision community, with wide applications in video surveillance and industrial inspection. It aims to detect outliers from seen classes and novel patterns from unseen classes. This task is very challenging because abnormal data is diversely distributed and expensive to collect. So we have to construct models based on only normal samples under unsupervised setting, targeting at high discrimination between normal and abnormal samples.

During the past decade, reconstruction-based methods have achieved great progress in anomaly detection. These methods use Autoencoders (AEs) [8, 9, 20, 24, 26, 29, 40] or Generative Adversarial Networks (GANs) [1, 19, 32] to re-

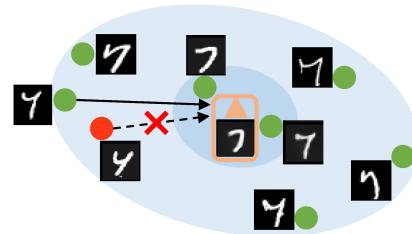


Figure 1. Illustration of difficulty in anomaly detection in MNIST dataset. The prototype is indicated by orange triangle and the anomaly by red point. In this case, the anomaly can hardly be detected based on reconstruction error or distance in high-dimensional feature space. Our solution is illustrated in Fig. 2.

construct the normal counterparts from any input images or video frames. AE-based methods firstly compress the inputs to discard the information beyond normal prototypes, and then decode the embedding to reconstruct the inputs. According to the estimated reconstruction error, the anomalies can be detected.

However, the performance of reconstruction-based methods for anomaly detection has long been limited by a tough problem, *i.e. the tradeoff between reconstructing diverse normals and detecting unknown anomalies*. In order to discriminate anomalies more easily, previous works [8] imposes more constraints to suppress abnormal information during autoencoding, which leads to high reconstruction error for diverse normal instances. For example, in Figs. 1 and 2g, the severely deformed normal (a.k.a. anomaly-like) sample “7” has even higher error than the abnormal sample “4”. To better reconstruct diverse normals, each query vector correspond to multiple prototypes in the memory, which may be combined into abnormal embedding even if abnormal projection is far away from the prototype. As a consequence, anomalies that distribute in low likelihood area between prototypical embedding are difficult to identify from diverse normals. MNAD [29] introduces skip-connection for diverse reconstruction and additional constraints to get round the incorrect combination problem. But the latter forces model transmit more unrestrained in-

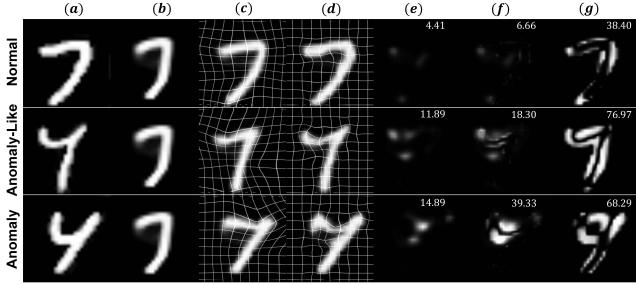


Figure 2. Illustration of our diversity-measurable method in addressing the detection difficulty. Numbers in white are anomaly scores. **a)** Original input; **b)** Reconstructed reference; **c)** Coarse deformation; **d)** Fine deformation; **e)** Measurement of diversity¹; **f)** Deformation-augmented error map assigns lower anomaly score to the anomaly-like sample than the true anomaly; **g)** Pixel-wise reconstruction error yields incorrect anomaly scores.

formation with abnormal part by skip-connection, resulting in shortcut learning and undesired reconstruction of anomalies.

A key to address the above tradeoff problem is to find a proper measurement of diversity that normal and abnormal samples have, which is positively correlated with the severity of anomaly. With such a measure, we do not need to fight against imperfect reconstruction of normals or undesired reconstruction of anomalies, because anomalies can be detected more accurately by the diversity measure together with the reconstruction error. Note that pixel-wise reconstruction error is not an ideal measurement of diversity, because the high-error region often confuses anomalies with diverse normals, *e.g.* normals with structural deformation and anomalies with colors close to the background may yield unreliable reconstruction error.

In this paper, we propose a **Diversity-Measurable Anomaly Detection (DMAD)** framework to enhance the measurability of reconstruction diversity so as to measure abnormality more accurately. Our basic idea is to decouple the reconstruction into compact representation of prototypical normals and measurable deformations of more diverse normals and anomalies. The under-estimated reconstruction error can be compensated by the diversity, which can be properly measured. To this end, the DMAD framework includes a **Pyramid Deformation Module (PDM)** to model and measure the diversity and an Information Compression Module (ICM) to learn the prototypical normal patterns.

Inspired by [4, 15], we assume anomalies (*e.g.* in video surveillance) can be represented as significant deformation of appearances, including positional changes and fine motions. In contrast, diverse normal samples can be represented as weaker deformations thus easily distinguished

¹In this case, we only count fine deformation because deformations in position and angle are considered as normal. In real-world experiments, we consider both coarse and fine deformations.

from the abnormal ones. Therefore, we design PDM to model the diversity of normals as well as the severity of anomalies. More specifically, PDM learns hierarchical two-dimensional deformation fields (Fig. 2c,d) that describe the pixel-level transformation direction and distance from reference (Fig. 2b, which is reconstructed from prototypes in memory) to original input. In ICM, we learn compressed representation as sparse prototypes. As a result, a single memory item is enough to represent each normal cluster. This is more compact than other memory-based works which require multiple memory items. Integrating PDM with ICM, DMAD essentially decouples the deformation information (Fig. 2e) from class prototypes and makes the final anomaly score more discriminative (Fig. 2f).

We evaluate our anomaly detection framework in scenarios of video surveillance and industrial defect detection. To apply DMAD in the latter scenario, we propose a variant of PDM, PPDM, to deal with the false positive issue in texture reconstruction. Extensive experimental results verify the efficacy of our approach. Moreover, our method works well even in front of contaminated data and anomaly-like normals. The main contributions of our work are as follows:

- We introduce diversity-measurable anomaly detection framework which allows reconstruction-based models to achieve better tradeoff between reconstructing diverse normals and detecting unknown anomalies.
- We propose pyramid deformation module to implement diversity measurement, in which the deformation information is explicitly separated from compact class prototypes and the resulting diversity measure is positively correlated to abnormality.
- Our approach outperforms previous works on video anomaly detection and industrial defect detection, and works well in front of contaminated data and anomaly-like normals, demonstrating its broad suitability and robustness.

2. Related Work

Anomaly detection. Reconstruction-based methods model the distribution of normal data and assign anomalies with high reconstruction error, because models trained with only normal data cannot reconstruct anomalies. Some works use autoencoder to detect anomalies, such as convolutional autoencoder [9] and the variants [24, 40]. Other approaches introduce additional constraints or memory to make the model more discriminative. For example, sparse coding [25, 39] reduces representation redundancy with regularization; memory-augmented autoencoder (MemAE) [8] memorizes the normal patterns appearing in training dataset with external memory bank; variational autoencoder [10, 35] assumes a prior distribution of normal data to constrain the nonlinear representation capacity; HF²VAD [20]

uses CVAE [33] to capture the correlation among motions. Frame prediction [19] assumes that abnormal samples in videos cannot be represented by past frames that do not contain unseen information and forces the model to encode changes among different frames. In addition, autoencoders can also be combined with external object detectors [7, 14, 20] to capture background-invariant appearance.

Although these methods generally work well, they often have difficulty in discriminating abnormal samples from anomaly-like normals due to the tradeoff between reconstruction and discrimination. In our framework, pyramid deformation module and information compression module are leveraged to address the trade-off problem, significantly improving the performance in anomaly detection.

Memory network. Generative models map continuously in feature space and they may assign higher probabilities to anomalies than normal ones [28]. Recent research [8, 11, 20, 29] has explored the application of discrete external memory to generate seen normal patterns even if the input is abnormal. MemAE [8] proposes a memory-augmented autoencoder, which uses information from encoder as a query and obtains a normal pattern retrieved from memory module. MNAD [29] introduces skip-connection to alleviate the problem that diverse normal patterns may yield high reconstruction errors. HF²VAD [20] extends the memory module to multi-level memory, and uses additional estimators [3, 13] to explicitly model the motion information and filter the noise from background. DAAD [11] uses block-wise memory to increase the specificity of memory.

In most existing works, the memory module outputs a linear combination of memory items which may lead to undesired reconstruction of normal-like anomalies. And they do not explicitly consider that instances corresponding to the same memory grid may be located at different positions of the receptive field. In our approach, we alleviate these problems via compressing embedding into a single memory item to ensure that the output is absolutely normal.

Transformation modeling. The transformation between video frames is a key clue for anomaly detection. Some methods [7, 14, 19, 20] use external object detectors or optical-flow estimators to model the motion information implicitly or explicitly. Among the transformation estimators, optical-flow estimation [13] is widely studied, which aims to estimate the pixel-wise motions between consecutive frames. In addition, correspondence modeling between pair-wise images is also applied to 2D geometric matching [30] and 3D image registration [36] to obtain image deformation fields. STN [15] learns to transform the original images themselves to benefit identification algorithm. RegGAD [12] uses feature-level affine registration to relocate features without changing the embeddings.

However, pairwise transformation modeling only focus on pattern changes thus cannot detect static anomalies. And

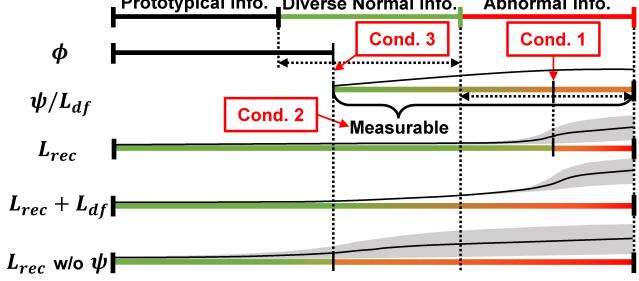


Figure 3. Three conditions for DMAD framework. The colors of lines represent the measurements of diversity (severity of anomaly); The gray area represents uncertainty in the measure; Double dashed arrows indicate possible ranges of the boundary. The last two lines denote DMAD and traditional reconstruction-based method respectively.

pretrained transformation estimators introduce generalization error in different scenarios. In contrast, we address these problems by separating and measuring the transformation from prototypical memory in end-to-end training.

3. Diversity-Measurable Anomaly Detection

In this section, we first analyze the objective of reconstruction-based anomaly detection and propose a diversity-measurable framework to address the tradeoff problem in existing works. Then, we introduce information compression module and pyramid deformation module (PDM) as an implementation of the framework. Finally, we explain the training and inference process and how to apply the framework to defect detection with a variant of PDM.

3.1. The framework

Given input x , autoencoder-based methods aim to model normal data distribution by minimizing the following reconstruction loss ($\|\cdot\|_2$ is just one type of reconstruction loss):

$$L = \|x - g(\phi(f(x), z))\|_2 + \gamma_1 R_1(\phi), \quad (1)$$

with respect to encoder $f(\cdot)$, decoder $g(\cdot)$, latent variable z (referring to discrete memory items in these cases) and a constrained feature mapping function $\phi(\cdot)$ corresponding to its constraint $R_1(\phi)$. Skip-connection and concatenation $[\cdot, \cdot]$ are introduced to generate diverse normal patterns [29]:

$$L = \|x - g([\phi(f(x), z), f(x)])\|_2 + \gamma_1 R_1(\phi). \quad (2)$$

However, due to diverse data distribution, previous methods have to face the conflict between representing diverse normals and detecting anomalies. The intrinsic reason lies in that the encoding of diversity $[\cdot, f(x)]$ contains redundant information that cannot be measured accurately.

In this work, we propose a principled framework, **Diversity-Measurable Anomaly Detection (DMAD)**, to alleviate the conflict. The basic idea is to restrict the anomaly

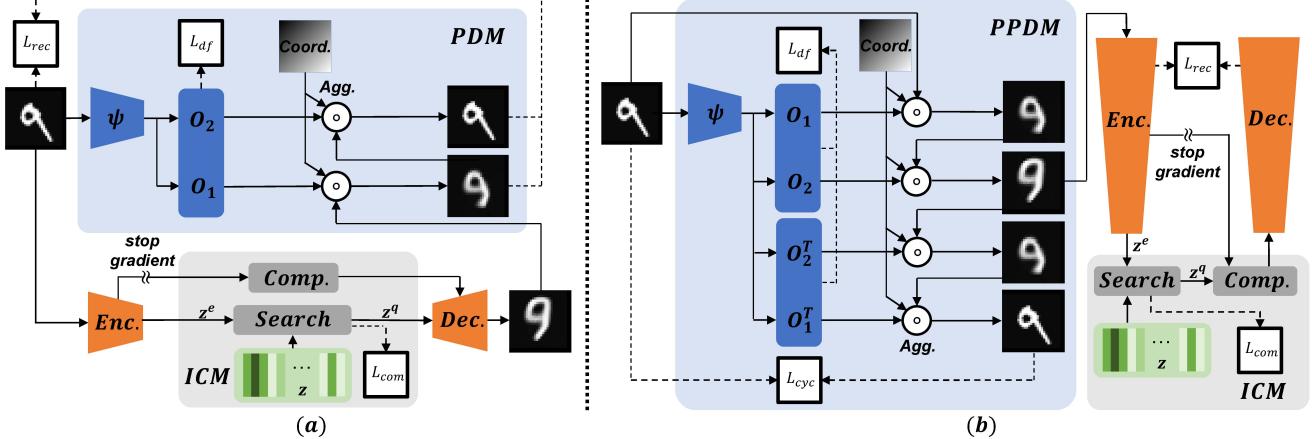


Figure 4. Two versions of diversity-measurable anomaly detection framework. Multi-scale pyramid deformation fields are estimated as \mathbf{O} and the reverse processes are \mathbf{O}^T . **a)** PDM version computes forward deformation \mathbf{O} after reconstruction. **b)** PPDM version employs cycle-consistent forward-backward deformations and the forward deformation is applied on the input.

information transmitted to $g(\cdot)$ while measuring and modeling the diversity of the remaining part. To this end, we design information compression module $\phi(\cdot)$ and diversity-aware module $\psi(\cdot)$ under DMAD framework:

$$L = \|x - g(\phi(f(x), z)) \circ \psi(x)\|_2 + \gamma_1 R_1(\phi) + \gamma_2 R_2(\psi), \quad (3)$$

where \circ refers to aggregation operator. With appropriate design of $\phi(\cdot)$, $\psi(\cdot)$ and the constraints, optimization of the reconstruction loss can improve the compactness of feature embedding. So that diverse representations are mapped via $\phi(\cdot)$ to compact prototypes in the memory. The diversity of input x relative to its reconstruction is represented by $\psi(\cdot)$. The under-estimated reconstruction error can be compensated by the diversity measured in $\psi(\cdot)$, a key factor contributing to accurate anomaly scores.

This framework can achieve our target on the premise that the following conditions (Fig. 3) are met: 1. $\psi(\cdot)$ can learn all diverse information from prototypical patterns to any normal inputs to ensure normal samples do not yield high anomaly score; 2. Deformation generated by $\psi(\cdot)$ is positively correlated to diversity measure; 3. Prototypical information about x represented by $\psi(\cdot)$ needs to be minimized. In the following subsections, we explain how to design modules to fulfill these conditions.

3.2. Information compression module

According to [35], we adapt VQ-Layer as an information compression module to learn $\phi(\cdot)$ given embedding $f(x) \in R^{D \times H' \times W'}$ as a query $z^e = f(x)$ and memory $z \in R^{D \times N}$. Then, we quantize z^e into a single-memory feature cube $z^q \in R^{D \times H' \times W'}$ by seeking for memory item with minimum $L2$ distance (“Search” in Fig. 4):

$$z_{h,w}^q = \operatorname{argmin}_{z_n} \|z_{h,w}^e - z_n\|_2, \quad (4)$$

where z_n is the n^{th} memory item, $h \in \{1, \dots, H'\}$, $w \in \{1, \dots, W'\}$ indicates the same location in both z^q and z^e . The compression loss L_{com} with stop-gradient operator $SG(\cdot)$ that updates its parameters separately [35] is combined by a hyperparameter β :

$$L_{com} = \|SG(z^e) - z^q\|_2 + \beta \|z^e - SG(z^q)\|_2. \quad (5)$$

Skip-connection with low information capacity (Fig. 4 “Comp.”) is also allowed to further improve the reconstruction quality without bringing excessive generalization (*i.e.* Conv-Layer with stop-gradient operator for intermediate features whose reduction factor is 16 or larger).

3.3. Pyramid deformation module

We categorize the unknown anomalies into the following three types: unseen class (*e.g.* novel objects), global anomaly (*e.g.* unexpected movement) and local anomaly (*e.g.* strange behavior and workpiece damage) of seen class. The unseen class is easy to be detected based on reconstruction result, but the latter two types are usually confused with normal ones. To discriminate these anomalies from normal ones, we represent the diversity using measurable deformation between reconstructed reference and original input, so that slight deformations occur in normals while drastic deformations occur in anomalies.

Inspired by STN and DCN [4,15], we introduce **Pyramid Deformation Module (PDM)** which explicitly learn deformation fields with hierarchical scales to model the motion, behavior and defect of different anomaly types, as shown in Fig. 4a. Specifically, after feature extraction, $\psi(\cdot)$ uses K -heads to compute offsets $\mathbf{O} = \{O_1, \dots, O_K\}$, corresponding to K coarse-to-fine deformations:

$$\psi(x) = Up(h(PE(x))) = \mathbf{O}, \quad (6)$$

where $PE(\cdot)$ is positional embedding operator [18], $h : R^{C \times H \times W} \rightarrow R^{2 \times \{H_1 \times W_1, \dots, H_K \times W_K\}}$ is the deformation estimator that generates offset vectors, $Up(\cdot)$ is upsampling function that resizes the outputs of K -heads to the same size with the original image. In our experiments, we set $K = 2$, with O_1 used to estimate coarse deformation (*e.g.* corresponding to the position of pedestrians or the placement of workpieces) and O_2 used to estimate fine deformation (*e.g.* corresponding to pedestrian behaviors or workpiece details).

Considering that quantized embedding without positional information may lead to inaccurate reconstruction, we also introduce position embedding operator for the decoder $g(\cdot)$. We then aggregate \mathbf{O} onto the reconstructed reference $g(PE(z^q))$, and obtain $\tilde{x}_k (k = 1, \dots, K)$ which is calibrated by the k^{th} layer of deformation fields:

$$\tilde{x}_k = g(PE(z^q)) \circ O_1 \cdots \circ O_k, \quad (7)$$

where \circ is grid-sampling function with a reference coordinate in this implementation (“Agg.” and “Coord.” in Fig. 4). However, minimizing the unconstrained reconstruction loss with respect to $\psi(x)$ may cause degenerate solution of the encoder $f(\cdot)$. To address this problem, we add constraint using smoothness loss via gradient operation and strength loss as:

$$L_{df} = \sum_k \|\nabla O_k\|_1 + \|O_k\|_2. \quad (8)$$

3.4. Foreground-background selection

Storing the background information in memory will break the compactness of embedding and require plenty of memory items. Besides, the deformation estimation should not be applied to the background. Some approaches use external estimators to remove interference from background, but the generalization in different scenarios cannot be guaranteed and extra noise will inevitably be introduced. Benefiting from the strong prior of fixed-view videos, we model the background with a learnable template x_{bg} and generate a binary mask to indicate whether a pixel belongs to foreground or background with $f_m(\cdot)$. The final reconstruction \hat{x}_k of k^{th} head is:

$$\hat{x}_k = f_m(x)\tilde{x}_k + (1 - f_m(x))x_{bg}. \quad (9)$$

3.5. Training and inference

Training phase. Once we obtain the reconstruction \hat{x}_k , we can calculate reconstruction loss L_{rec} as:

$$L_{rec} = \sum_k Dis(x, \hat{x}_k), \quad (10)$$

where $Dis(\cdot)$ refers to a distance function in sample space. Reminding of the optimization objective in Eq. 3, we implement the two constraints using L_{com} and L_{df} respectively.

Finally, training is performed by minimizing the overall loss:

$$L_{all} = L_{rec} + \gamma_1 L_{com} + \gamma_2 L_{df}. \quad (11)$$

Inference phase. In the inference phase, we use \mathbf{O} and the reconstruction loss to calculate the error maps of input sample x :

$$A_{rec} = Dis(x, \hat{x}_K), \quad (12)$$

$$A_{df} = \sum_k \|O_k\|_2. \quad (13)$$

Image-level anomaly score is computed based on local maxima:

$$Score_I = \max(A_{rec} \otimes k^*) + \alpha \max(A_{df} \otimes k^*). \quad (14)$$

where \otimes is convolution operator and k^* is convolution kernel for anomaly maps. α is a tradeoff parameter. That is, the reconstruction loss and deformation jointly determine the anomaly score, which is much more effective than traditional reconstruction-based methods, as illustrated in Fig. 3.

3.6. Variant of PDM

Modified framework with Pre-PDM. The DMAD framework together with ICM and PDM modules proposed above are suitable in many anomaly detection scenarios including video surveillance. However, for industrial defect detection, texture reconstruction may be harmful (such as spots on “Pill”), and we should reconstruct high-level semantic features instead. Since PDM does not work in high-dimensional feature space and interferes with the training process, we propose variant of PDM, *Pre-PDM (PPDM)*, as a solution. PPDM works in sample space, and is applied to the input sample rather than the reconstructed one. Eq. (3) is modified naturally as follows:

$$L = \|x \circ \psi'(x) - g(\phi(f(x \circ \psi'(x)), z))\|_2 + \gamma_1 R_1(\phi) + \gamma_2 R_2(\psi'). \quad (15)$$

Since we do not reconstruct the original samples, reconstruction loss cannot constrain PPDM to maintain information diversity. In order to prevent $x \circ \psi'(x)$ in Eq. (15) from shortcut learning, we propose to add the inversion of forward deformation, backward deformation \mathbf{O}^T , based on cycle-consistency principle to maintain the diversity of appearance information:

$$\psi'(x) = Up(h'(PE(x))) = \{\mathbf{O}, \mathbf{O}^T\}. \quad (16)$$

Training phase. The cycle-consistency losses L_{cyc} and constraint for forward-backward deformation L_{df}^+ are:

$$L_{cyc} = \|x - x \circ O_1 \cdots \circ O_K \circ O_K^T \cdots \circ O_1^T\|_2, \quad (17)$$

$$L_{df}^+ = \sum_k \|\nabla O_k\|_1 + \|\nabla O_k^T\|_1 + \|O_k\|_2 + \|O_k^T\|_2. \quad (18)$$

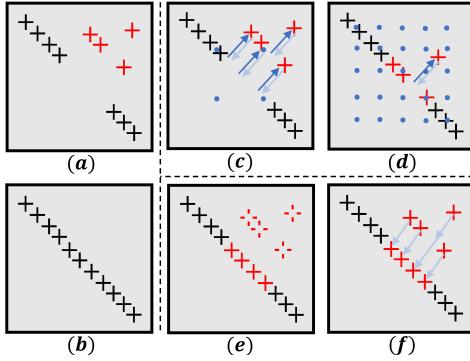


Figure 5. The process of forward and backward deformation: **a**) The pattern of original image. Red indicates high error regions and black indicates low error regions; **b**) The normal pattern; **c**) The coarse estimation O_1 (dark blue) and O_1^T (light blue); **d**) The fine estimation O_2 (dark blue) and O_2^T (light blue); **e**) Using \hat{x} simply causes high anomaly score at the wrong location (solid cross). Ground-truth is shown as hollow cross; **f**) $\hat{x} \circ O_2^T \circ O_1^T$ adjusts the high reconstruction error to original position.

So we train our anomaly detection model with PPDM by minimizing the following loss:

$$L_{all}^+ = L_{rec} + \gamma_1 L_{com} + \gamma_2 L_{df}^+ + \gamma_3 L_{cyc}. \quad (19)$$

The modified DMAD framework (*i.e.* with PPDM and Reverse Distribution [6]) is shown in Fig. 4b.

Inference phase. The result of PPDM is not aligned with the position of original input, which may reduce the performance of anomaly localization. In order to obtain the anomaly map of real location, we use backward deformation to perform inverse sampling (Fig. 5):

$$A_{rec}^+ = Dis(x, \hat{x}_K) \circ O_K^T \cdots \circ O_1^T, \quad (20)$$

$$\begin{aligned} A_{df}^+ &= \sum_k \|O_k \circ O_{k+1} \cdots \circ O_K\|_2 \\ &\quad + \sum_k \|O_k^T \circ O_{k-1}^T \cdots \circ O_1^T\|_2. \end{aligned} \quad (21)$$

The image-level anomaly score $Score_I^+$ and pixel-level anomaly score $Score_P^+$ are calculated respectively as:

$$Score_I^+ = \max(A_{rec}^+ \otimes k^*) + \alpha \max(A_{df}^+ \otimes k^*), \quad (22)$$

$$Score_P^+ = A_{rec}^+ + \alpha A_{df}^+. \quad (23)$$

4. Experiments and Analysis

Firstly, we perform toy experiment on MNIST to illustrate our approach. Then the quantitative and qualitative results of two versions of our DMAD framework are reported in video anomaly detection and industrial surface defect detection respectively². Finally, we conduct ablation experiments and analyze the results.

²Our code will be available at <https://github.com/FlappyPeggy/DMAD>.

	Normal: “1, 3, 5, 7, 9”	Abnormal: “0, 2, 4, 6, 8”
Method / Input	1 3 5 7 9 0 2 4 6 8	
Mem. Only	1 3 5 7 9 5 3 9 3 9	
+ Skip	1 3 5 7 9 0 2 4 6 8	
+ PDM	1 3 5 7 9 5 3 9 9 3	1 3 5 7 9 5 3 9 8 3
	1 3 5 7 9 5 3 9 8 3	

Figure 6. Visualization of toy experiment. The last three rows are: reference reconstructed from memory, reference after coarse deformation, and final output after applying fine deformation.

4.1. Datasets

Surveillance Videos. Ped2 [27], Avenue [23] and ShanghaiTech [25] are fixed-view videos. The anomalies include driving, cycling, running, throwing stuff, etc. Mutual occlusion, anomaly-like behavior, contaminated data and different scenes occur frequently in these datasets.

Industrial Images. MVTec [2] contains 15 types of industrial images, which are divided into 5 types of textures and 10 types of objects. The defects include crack, scratch, etc. The normal workpiece has different positions, angles and textures. It is used for detection and localization task.

4.2. Toy experiment

As shown in Fig. 6, we perform a toy experiment on MNIST dataset [16] with setting analogous to out-of-distribution (OOD) detection (*i.e.* training on “1, 3, 5, 7, 9” and test on all classes). Our model searches separate memory items for each digital category to reconstruct it into class-specific reference and uses the deformation fields from PDM to adjust it hierarchically. When tested with seen and unseen classes, the model adjusts reconstructed references to normal inputs, but fails on the abnormal ones.

In contrast, memory network without diversity-aware module cannot guarantee the intra-class compactness and the reconstruction diversity, which misdirects the model to obtain dataset-optimal “average memory”, leading to fuzzy reconstruction and lower discrimination ability. The model with full-channel skip-connection suffers from short-cut learning and reconstruct anomalies successfully which weakens the ability to identify anomalies.

4.3. Implementation details

The input images are resized into 256×256 and normalized to the value in range of $[-1, 1]$. According to frame-pred strategy [19], the history length is set to 4 in video anomaly detection and 0 for images. After feature extraction with a backbone, PDM and PPDM obtain different heads by stacking stride-2 convolution layers. Then

we use $Tanh$ as the activation function of output layer and clipping function to ensure the value of deformation is between $[-1, 1]$. Unless otherwise noted, the architecture of autoencoder complies with the settings of MNAD [29] and RD [6] for PDM and PPDM respectively. The function $Dis(\cdot)$ in Eq. (10) is a combination of MSE loss and $Grad$ loss for reconstruction in sample space, and COS loss for deep features. We set $(\gamma_1, \gamma_2) = (1, 0.25)$ for PDM, $(1, 1)$ for PPDM (γ_3 is discussed in Sec. 4.5), and $\beta = 0.25$ as in [35]. During post-processing, we use average kernel for surveillance videos and Gaussian kernel with $\sigma = 4$ for industrial images as in [6]. Frame-difference method is applied to remove static anomalies for Avenue, because our method detect all anomalies which may be labeled as normal ones in [23]. And we use the mask calculated in Eq. (9) to normalize anomaly map in [25] due to the scale changes. Besides, we set $\alpha = 0.2$ for three video tasks and an additional option $\alpha = 0.05$ for defect detection, depending on whether the defect includes geometrical changes. The model is optimized by AdamW [21] and the learning rate is $2e-4$ and $5e-3$ as in [6, 29] respectively, being decayed by CosineAnnealingLR [22] strategy. We adopt 60, 60, 60, 10, 400 epochs for MNIST, Ped2, Avenue, Shanghai and MVTec respectively with batch size of 8.

4.4. Main results

Surveillance Videos. We compare our method with SOTA works on video anomaly detection in Tab. 1. Our method outperforms comparative approaches even though we neither use external estimators nor remove abnormal frames in training data. In addition, if we detect global offset for camera jitter in [23], there will be extra 0.1% gain.

The qualitative results are shown in Fig. 7. We find that: anomalies are over-reconstructed in (b); normals are not reconstructed well in (c); (d) greatly improves normal reconstruction with slightly less inhibition ability for anomalies.

Industrial Images. Anomaly detection results on MVTec are shown in Tab. 2 and localization results are shown in Tab. 3. With the deformation from PPDM, texture anomalies are detected with high performance and our method outperforms SOTA methods in both detection and localization tasks without memorizing an enormous number of embedding from training data.

4.5. Ablation study

As shown in Tab. 4, single-output memory module without PDM suppresses diverse normal patterns seriously, while separate PDM without memory provides comparable performance gains as the previous SOTA works because the module “*Comp.*” serves as an information compression module instead. The number of multi-scale deformation fields also have a modest effect on performance. We suggest that “ K ” should at least make the scale of control grid

Table 1. Video anomaly detection results on Ped2 [27], Avenue [23] and Shanghai [25]. We calculate $AUC(\%)$ with all frames together. Numbers in bold indicate the best performance and the underlined ones are the second best. $^+$ indicates that we reproduce the result due to higher performance or absence of implementation.

Methods	Ped2 [27]	Avenue [23]	Shanghai [25]
Conv2D [9]	90.0	70.2	-
Conv3D [40]	91.2	77.1	-
ConvLSTM [24]	88.1	77.0	-
FramePred [19]	95.4	84.9	72.8
ConvVQ ⁺	90.2	84.3	-
MemAE [8]	94.1	83.3	71.2
MPN [26]	96.9	89.5	73.8
MNAD ⁺ [29]	97.8	88.5	70.5
HF ² VAD [20]	<u>99.3</u>	<u>91.1</u>	<u>76.2</u>
Ours	99.7	92.8	78.8

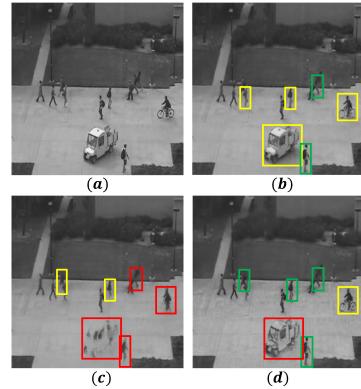


Figure 7. Qualitative results on Ped2. **a**) Original image; **b**) Memory network with skip-connection [29]; **c**) PDM with VQ-Layer [35]; **d**) PDM with ICM. Green box indicates negative, red box indicates positive, and yellow box indicates weak positive.

cover the size of base elements (e.g. pedestrian limbs). Besides, the foreground-background selection module further improves the compactness of memory embedding. Moreover, if any constraint for PDM is missing, the abnormal information will be transmitted and cause shortcut learning. Especially, the cycle-consistency constraint L_{cyc} is also a necessary part for PPDM to avoid degenerated solution (-1.7%), because the feature reconstruction error can be minimized by eliminating all necessary information.

As shown in Tab. 5, the proposed method is robust to hyperparameter γ_3 in Eq. (19). On condition that degenerated solution will not be formed, weakening the constraint makes model restore images from reference with less detail more easily and perceive the position of anomalies more accurately by transforming anomalies to normal patterns which benefits localization task. On the contrary, strengthening the constraint alleviates shortcut learning and improves image-level result by maintaining more abnormal details.

Table 2. Image-level $AUC(\%)$ of anomaly detection on MVTec [2].

Class \ Methods	GN [1]	PSVDD [37]	DAAD [11]	CutPaste [17]	PaDiM [5]	PatchCore [31]	DRAEM [38]	RD [6]	Ours
Avg _{Textures}	77.5	94.5	91.0	97.5	98.8	99.0	99.1	<u>99.5</u>	99.9
Avg _{Objects}	75.5	90.8	88.8	95.5	93.8	<u>99.2</u>	97.4	98.0	99.3
Avg _{All}	76.2	92.1	89.5	96.1	95.5	<u>99.1</u>	98.0	98.5	99.5

Table 3. Pixel-level $AUC(\%)$ of anomaly localization on MVTec [2].

Class \ Methods	PSVDD [37]	CutPaste [17]	PaDiM [5]	PatchCore [31]	DRAEM [38]	TMAE [34]	RD [6]	Ours
Avg _{Textures}	93.7	96.3	96.9	97.6	97.9	93.8	97.7	<u>97.8</u>
Avg _{Objects}	96.7	95.8	97.8	98.4	97.0	94.0	97.9	<u>98.3</u>
Avg _{All}	95.7	96.0	97.5	<u>98.1</u>	97.3	93.9	97.8	98.2

Table 4. Ablation study of proposed module and loss on Ped2 [27]. From left to right: without PDM; with only one head of PDM ($K=1$); without memory module; without background template; removing strength constraint; removing smoothness constraint.

w/o	PDM	$K > 1$	Mem	x_{bg}	$\ O\ _2$	$\ \nabla O\ _2$
AUC%	-9.4	-0.4	-1.8	-1.6	-1.4	-1.4

Table 5. Ablation study of deformation constraint on MVTec [2].

Task \ γ_3	0	0.1	0.25	0.5	1	2
Det.	97.9	99.4	99.4	<u>99.5</u>	99.6	99.4
Loc.	96.6	<u>98.0</u>	98.2	<u>98.0</u>	97.7	97.6

4.6. Discussion

Contaminated data assumption. Assuming that training data only contains completely normal data is unrealistic, because the workload of natural data cleansing is considerable, even has same cost as data annotations. We mix training data with abnormal events in Ped2 to simulate contaminated data and find performance drop of MNAD [29] is -3.7% while ours is -1.8% . The reason why DMAD is less affected may be that PDM can encode and transmit residual abnormal representation of mixed anomalies and anomaly-like normal samples, thus strictly maintaining representation compactness of the main encoder and low generalization ability on abnormal samples.

Two DMAD implementations. As introduced earlier, our DMAD framework is implemented by two versions using PDM and PPDM respectively, corresponding to different detection targets. As shown in Fig. 8, PDM learns measurable quantization error (caused by ICM) from memory embedding to diverse patterns which enhances the potential of ICM to keep intra-class compactness without generating unmeasurable reconstruction loss. Unlike PDM, PPDM is partially in charge of the information compression, *i.e.* PPDM removes the diverse appearance by using the reverse deformation process from inputs to reference.

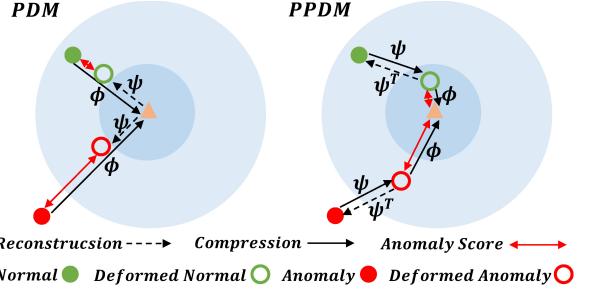


Figure 8. Illustration of two DMAD implementations with PDM and PPDM respectively.

5. Conclusion

In this paper, we present a reconstruction-based diversity-measurable anomaly detection framework, which simultaneously enhances anomaly discrimination and reconstruction diversity. Pyramid deformation module is proposed to be used together with information compression module for this purpose. PDM models multi-scale transformation fields from reference to original input explicitly without relying on external estimators. Therefore, diverse normal patterns can be reconstructed and anomaly severity can be measured accurately. Empirical studies on both videos and images benchmarks show the effectiveness and applicability of our work. In future research, we will further explore diversity-aware models for anomaly detection.

Limitations. Our method focuses on anomaly with measurable geometrical diversity, the most common type in anomaly detection. However, as for anomaly with other kind of diversities, *e.g.* colors, the proposed diversity measure may not be positively correlated to anomaly severity.

Acknowledgement. This work is partially supported by National Key R&D Program of China (No. 2018AAA0102402), and National Natural Science Foundation of China (NSFC): 61976203 and 62276246.

References

- [1] Samet Akcay, Amir Atapour Abarghouei, and Toby P. Breckon. Gandomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision*, pages 622–637, 2018. [1](#), [8](#)
- [2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtac AD - A comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019. [6](#), [8](#)
- [3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. [3](#)
- [4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *IEEE International Conference on Computer Vision*, pages 764–773, 2017. [2](#), [4](#)
- [5] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *In International Conference on Pattern Recognition*, pages 475–489, 2020. [8](#)
- [6] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9737–9746, 2022. [6](#), [7](#), [8](#)
- [7] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12742–12752, 2021. [3](#)
- [8] 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 *2019 IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019. [1](#), [2](#), [3](#), [7](#)
- [9] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–742, 2016. [1](#), [2](#), [7](#)
- [10] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations*, 2017. [2](#)
- [11] Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 8771–8780, 2021. [3](#), [8](#)
- [12] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael W. Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. *arXiv preprint arXiv:2207.07361*, 2022. [3](#)
- [13] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margaret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1647–1655, 2017. [3](#)
- [14] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019. [3](#)
- [15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 2017–2025, 2015. [2](#), [3](#), [4](#)
- [16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, pages 2278–2324, 1998. [6](#)
- [17] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. [8](#)
- [18] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pages 9628–9639, 2018. [5](#)
- [19] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - A new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018. [1](#), [3](#), [6](#), [7](#)
- [20] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 13568–13577, 2021. [1](#), [2](#), [3](#), [7](#)
- [21] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 2017. [7](#)
- [22] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations*, 2017. [7](#)
- [23] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 FPS in MATLAB. In *IEEE International Conference on Computer Vision*, pages 2720–2727, 2013. [6](#), [7](#)
- [24] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional LSTM for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo*, pages 439–444, 2017. [1](#), [2](#), [7](#)
- [25] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked RNN frame-

- work. In *IEEE International Conference on Computer Vision*, pages 341–349, 2017. 2, 6, 7
- [26] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15425–15434, 2021. 1, 7
- [27] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010. 6, 7, 8
- [28] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don’t know? In *7th International Conference on Learning Representations*, 2019. 3
- [29] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14360–14369, 2020. 1, 3, 7, 8
- [30] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2017. 3
- [31] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter V. Gehler. Towards total recall in industrial anomaly detection. *arXiv preprint arXiv:2106.08265*, 2021. 8
- [32] 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 *Information Processing in Medical Imaging - 25th International Conference*, pages 146–157, 2017. 1
- [33] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 3483–3491, 2015. 3
- [34] Daniel Stanley Tan, Yi-Chun Chen, Trista Pei-Chun Chen, and Wei-Chao Chen. Trustmae: A noise-resilient defect classification framework using memory-augmented auto-encoders with trust regions. In *IEEE Winter Conference on Applications of Computer Vision*, pages 276–285, 2021. 8
- [35] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 6306–6315, 2017. 2, 4, 7
- [36] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7639–7648, 2021. 3
- [37] Jihun Yi and Sungroh Yoon. Patch SVDD: patch-level SVDD for anomaly detection and segmentation. In *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision*, pages 375–390, 2020. 8
- [38] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Dræm - A discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 8310–8319, 2021. 8
- [39] Bin Zhao, Li Fei-Fei, and Eric P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3320, 2011. 2
- [40] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1933–1941, 2017. 1, 2, 7