Focus the Discrepancy: Intra- and Inter-Correlation Learning for Image Anomaly Detection

Xincheng Yao¹, Ruoqi Li¹, Zefeng Qian¹, Yan Luo¹, Chongyang Zhang¹,²*
¹School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
²MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{i-dover, nilponi, zefeng_qian, luoyan_bb, sunny_zhang}@sjtu.edu.cn¹

Abstract

Humans recognize anomalies through two aspects: larger patch-wise representation discrepancies and weaker patch-to-normal-patch correlations. However, the previous AD methods didn’t sufficiently combine the two complementary aspects to design AD models. To this end, we find that Transformer can ideally satisfy the two aspects as its great power in the unified modeling of patch-wise representations and patch-to-patch correlations. In this paper, we propose a novel AD framework: FOcus-the-Discrepancy (FOD), which can simultaneously spot the patch-wise, intra- and inter-discrepancies of anomalies. The major characteristic of our method is that we renovate the self-attention maps in transformers to Intra-Inter-Correlation (I2Correlation). The I2Correlation contains a two-branch structure to first explicitly establish intra- and inter-image correlations, and then fuses the features of two-branch to spotlight the abnormal patterns. To learn the intra- and inter-correlations adaptively, we propose the RBF-kernel-based target-correlations as learning targets for self-supervised learning. Besides, we introduce an entropy constraint strategy to solve the mode collapse issue in optimization and further amplify the normal-abnormal distinguishability. Extensive experiments on three unsupervised real-world AD benchmarks show the superior performance of our approach. Code will be available at https://github.com/xcyao00/FOD.

1. Introduction

The goal of anomaly detection (AD) is to distinguish an instance containing anomalous patterns from those normal samples and further localize those anomalous regions. Anomalies are defined as opposite to normal samples and are usually rare, which means that we need to tackle AD tasks under the unsupervised setting with only normal samples accessible. The core idea of most unsupervised AD methods is to compare with normal samples to distinguish anomalies [46, 35, 60, 11, 28, 58]. Even for humans, we also recognize anomalies in this way, specifically through three discrepancies, i.e., 1. patch patterns that differentiate from the normal visuals; 2. image regions that destroy textures or structures; 3. novel appearances that deviate from our accumulated knowledge of normality. Namely, anomalous patches usually have three characteristics: their patch-wise representations are different from the normal visuals; they are different from most patches within one image; they deviate from our accumulated knowledge of normality. These views intrinsically reveal that humans’ recognition of anomalies depends on two aspects: patch-wise representations (1) and intra- and inter-image correlations (2, 3).

Previous methods mainly follow the former aspect to learn distinguishable representations or reconstructions,
such as reconstruct-based methods [6, 57, 1] and knowledge distillation AD models [4, 41]. The goal of these methods is to generate reconstructed samples or feature representations, and larger patch-wise representation discrepancies can appear in the abnormal patches. However, only the patch-wise representation discrepancies are insufficient for detecting more complex anomalies (e.g., rows 3 and 4 in Figure 1), since the patch-wise errors can’t provide comprehensive descriptions of the spatial context. Other mainstream AD methods, such as embedding-based [11, 35] and one-class-classification-based (OCC) [39, 60] methods, are much similar to the latter aspect. These methods achieve anomaly detection by measuring the distances between the features of test samples and normal features. Compared with the non-learnable feature distances, the explicit intra- and inter-image correlations in our method are more effective to detect diverse anomalies (see Table 1, 2, 3). Moreover, patch-wise representation discrepancies and intra- and inter-correlation discrepancies are complementary, and can be combined to develop more powerful AD models.

Recently, with the self-attention mechanism and long-range modeling ability, transformers [50] have significantly renovated many computer vision tasks [14, 24, 7, 67, 56] and recently popular language-vision multimodal tasks [30, 22]. Transformers have shown great power in the unified modeling of patch-wise representations and patch-to-patch correlations. Transformers are quite suitable for AD tasks as their modeling ability can satisfy the two aspects of anomaly recognition quite well. Some works [26, 18, 8, 62] also attempt to employ transformers to construct AD models. However, these methods only use transformers to extract vision features, which didn’t sufficiently adapt transformers’ long-range correlation modeling capability to AD tasks. Different from these works, we explicitly exploit transformers’ self-attention maps to establish the intra- and inter-image correlations. The correlation distribution of each patch can provide more informative descriptions of the spatial context, which can reveal more intricate and semantic anomaly patterns.

In this paper, motivated by humans’ anomaly recognition process, we propose a novel AD framework: FOCus-the-Discrepancy (FOD), which can exploit transformers’ unified modeling ability to simultaneously spot the patch-wise, intra- and inter-discrepancies. Our key designs are composed of three recognition branches: the patch-wise discrepancy branch is to reconstruct the input patch features for distinguishing simple anomalies; the intra-correlation branch is to explicitly model patch-to-patch correlations in one image for distinguishing hard global anomalies (e.g., row 1 in Figure 1); the inter-correlation branch is to explicitly learn inter-image correlations with known normal patterns from the whole normal training set. To implement the intra- and inter-correlation branches, we adapt Transformer and renovate the self-attention mechanism to the I2Correlation, which contains a two-branch structure to first separately model the intra- and inter-correlation distribution of each image patch, and then fuse the features of two-branch to spotlight the abnormal patterns. To learn the intra- and inter-correlations adaptively, we propose the RBF-kernel-based target-correlations as learning targets for self-supervised learning, the RBF kernel is used to present the neighborhood continuity of each image patch. Besides, an entropy constraint strategy is applied in the two branches, which can solve the mode collapse issue in optimization and further amplify the normal-abnormal distinguishability.

In summary, we make the following main contributions:

1. We propose a novel AD framework: FOD, which can effectively detect anomalies by simultaneously spotting the patch-wise, intra- and inter-discrepancies.

2. We renovate the self-attention mechanism to the I2Correlation, which can explicitly establish intra- and inter-correlations in a self-supervised way with the target-correlations. An entropy constraint strategy is proposed to further amplify the normal-abnormal distinguishability.

3. Our method can achieve SOTA results on three real-world AD datasets, this shows our method can more effectively determine anomalies from complementary views.

2. Related Work

Anomaly Detection. In this paper, we divide the mainstream AD methods into five categories: reconstruction-based, embedding-based, OCC-based methods, knowledge distillation and normalizing flow AD models. The reconstruction-based methods are the most popular AD methods and also widely studied, where the assumption is that models trained by normal samples only can reconstruct normal regions but fail in abnormal regions. Many previous works attempt to train AutoEncoders [6, 28, 57, 8, 17, 64], Variational AutoEncoders [23] and GANS [42, 1, 29, 40] to reconstruct the input images. Overfitting to the input images is the most serious issue of these methods, which means that the anomalies are also well reconstructed [62].

Recently, some embedding-based methods [2, 33, 9, 11, 35] show better AD performance by using ImageNet pre-trained networks as feature extractors. In [33], the authors fit a multivariate Gaussian to model the image-level features for further Mahalanobis distance measurement. PaDiM [11] extends the above method to localize pixel-level anomalies. PatchCore [35] extends on this line by utilizing locally aggregated, mid-level features and introducing greedy coreset subsampling to form nominal feature banks. However, these methods directly utilize pre-trained networks without any adaptation to the target dataset. Some works [31, 10, 32] attempt to adapt pre-trained features to the target data distribution. There are also some other methods for using pre-trained networks by freezing them and only
optimizing a sub-network, e.g., via knowledge distillation [4, 41, 12, 38], or normalizing flows [36, 15, 37, 63].

OCC is another classical AD paradigm, the earliest works are mainly to extend the OCC models such as OC-SVM [44] or SVDD [47, 39] for anomaly detection. Recently, in [60], a patch-based SVDD that contains multiple cores rather than a single core in [39] is proposed to enable anomaly localization. In [25], a Fully Convolutional Data Description combined with receptive field upsampling is proposed to generate anomaly maps. In [48], the authors further extend the PatchSVDD [60] model by the proposed multi-scale patch-based representation learning method.

Transformer-based Anomaly Detection Methods. Recently, transformers [50] have shown great power in modeling long-range dependencies. For image anomaly detection, some works [26, 18, 62, 8, 59] also attempt to exploit transformers to design AD models. However, most of these methods only use transformers as backbones to extract vision features, and don’t effectively adapt the long-range modeling capacity of transformers to AD tasks. Unlike the previous usage of transformers, we explicitly exploit the self-attention maps of transformers to establish intra- and inter-correlations. Our work shares some similarities with a recent work [20]. But we point out that our work has some significant differences from [20]: 1) Different Insight: [20] aims to unify the pointwise representation capability, which are more suitable for anomaly detection by knowledge distillation [4, 41, 12, 38], or normalizing flows [36, 15, 37, 63].

2) Novel Method: we propose entropy constraint, inter-correlation branch, and I2Correlation by effectively combining intra- and inter-correlation branches. 3) Different Task: [20] focuses on time-series AD, whereas we focus on image AD.

3. Approach

3.1. Model Overview

Figure 2 overviews our proposed approach. The model consists of three branches: patch-wise discrepancy branch, intra-correlation branch, and inter-correlation branch. The input 2D image is first sent into a pre-trained backbone to extract multi-scale feature maps. At each feature level, we construct a subsequent transformer network to exploit transformers to design AD models. However, most of these methods only use transformers as backbones to extract vision features, and don’t effectively adapt the long-range modeling capacity of transformers to AD tasks. Unlike the previous usage of transformers, we explicitly exploit the self-attention maps of transformers to establish intra- and inter-correlations. Our work shares some similarities with a recent work [20]. But we point out that our work has some significant differences from [20]: 1) Different Insight: [20] aims to unify the pointwise representation capability, which are more suitable for anomaly detection by knowledge distillation [4, 41, 12, 38], or normalizing flows [36, 15, 37, 63].

2) Novel Method: we propose entropy constraint, inter-correlation branch, and I2Correlation by effectively combining intra- and inter-correlation branches. 3) Different Task: [20] focuses on time-series AD, whereas we focus on image AD.

3.2. Patch-Wise Reconstruction Discrepancy

Adopted from previous reconstruction-based AD methods, we employ feature reconstruction as our patch-wise recognition view for simplicity. In our approach, we construct a transformer network to reconstruct the input features. With the long-range dependency modeling ability, the features reconstructed by Transformer can have larger effective receptive fields [13], which are more conducive to detect hard global anomalies and logical anomalies. Moreover, with the further introduced inter-correlation branch (sec.3.4), the features can even perceive normal regions of images from the whole normal training set. Therefore, the reconstructed features generated by our model have better global perception and more discriminative semantic representation capability, which are more suitable for anomaly detection by patch-wise representation discrepancies.

Learning Objective. We can utilize classical reconstruction losses as the learning objective. We combine the ℓ2 distance and cosine distance to measure the feature distances between the reconstructed features X̂ and the input features X. The loss function is defined as follows:

\[ \mathcal{L}_1 = \|X̂ - X\|_2 + \left(1 - \frac{X̂ \cdot X}{\|X̂\|_2 \|X\|_2}\right) \]

3.3. Intra-Correlation Learning

The intra-correlation learning branch aims to learn informative patch-to-patch correlations from the input patch sequence adaptively. As shown in the right part of Figure 2, we explicitly take advantage of the self-attention maps of transformers as intra-correlation matrices. Formally, the intra-correlation matrix of the kth layer is calculated by:

\[ S_k^q = \text{Softmax} \left( \frac{Q_k^q (K_k^q)^T}{\sqrt{d_m}} \right) \]

\[ [Q_k^q, K_k^q, V_k^q] = X_{k-1}[W_k^Q, W_k^K, W_k^V] \]

where \(Q_k^q, K_k^q, V_k^q \in \mathbb{R}^{N \times d_m}\) represent the query, key, and value of the kth layer. \(W_k^Q, W_k^K, W_k^V \in \mathbb{R}^{d_m \times d_m}\) represent the learnable projection matrices for \(Q_k^q, K_k^q, V_k^q\), respectively. \(S_k^q \in \mathbb{R}^{N \times N}\) denotes the learned intra-image patch-to-patch correlations. Since \(\text{Softmax}(\cdot)\) can convert the values in the similarity map into range [0, 1] along the horizontal axis, each row of \(S_k^q\) can represent a discrete correlation distribution for each corresponding image patch. But different from the vanilla transformers, we further introduce target correlation matrices as the learning objective to explicitly optimize the intra-image correlations.

6805
Target Correlation. Inspired by the contrastive learning method BarlowTwins [66], we can construct a target correlation matrix $T \in \mathbb{R}^{N \times N}$ as the learning target. The objective function of BarlowTwins measures the correlation matrix between the embeddings of two identical networks fed with distorted versions of a batch of samples, and tries to make this matrix close to the identity. In our work, the role of the target correlation is to introduce a prior correlation of patches as a pretext learning target, where each patch can be highly correlated to itself and also correlated to its neighborhood patches and the correlation decreases with the increase of distance. This allows us to optimize the intra- and inter-correlations in a self-supervised way. To this end, we use the radial basis function (RBF) to construct the target correlation matrix. We further adopt two learnable kernel variances $\sigma_x$ and $\sigma_y$ for horizontal and vertical axes to make the target correlation of each patch adapt to the specific pattern of itself. The target correlation matrix of the $k$th layer is defined as:

$$T_k^\theta = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{||x_{ij} - x_{ij'}||^2}{2(\sigma_x^2 + \sigma_y^2)}\right)$$

$$i, i' \in \{1, \ldots, H\}; j, j' \in \{1, \ldots, W\}$$

(5)

where $||x_{ij} - x_{ij'}||^2$ means the Square Euclidean distance between point $x_{ij}$ and $x_{ij'}$, $i, i'$ and $j, j'$ means vertical and horizontal coordinates, respectively.

Next, we need to measure the distance between target- and intra-correlation distributions. This can usually be achieved by calculating the KL divergence. We can obtain a KL divergence value from each level of the network. Thus, we average all KL divergence values to combine the patch-to-patch correlations from multi-layer features into a more informative measure as follows:

$$\text{Div}(T^\theta, S^\theta) = \frac{1}{K} \sum_{k=1}^{K} \left( K \text{L}(T_k^\theta \| S_k^\theta) + K \text{L}(S_k^\theta \| T_k^\theta) \right)$$

(6)

where $K \text{L}(\cdot \| \cdot) \in \mathbb{R}^N$ and its each element means the KL divergence between two discrete distributions corresponding to each row of $T_k^\theta$ and $S_k^\theta$.

Due to the rarity of anomalies and the dominance of normal patterns, the normal patches should build strong correlations with most patches in the whole image, while the weights of abnormal correlation distributions are harder to distribute to most patches and are more likely to concentrate on the adjacent image patches due to the neighborhood continuity. Since normal and abnormal patches have different correlation distributions, this is a distinguishable criterion for anomaly detection. Note that the intra-correlation branch explicitly exploits the spatial dependencies of each image patch, which are more informative than the patch-wise representations for anomaly detection.

Entropy constraint. Since the normal image patterns are usually diverse, the learned correlation distributions of the normal patches may also easily concentrate on the adjacent patches, which can cause the distinguishability between normal and abnormal to be downscaled. To address...
this, we further introduce an entropy constraint item for making normal patches establish strong associations with most normal patches in the whole image as much as possible. The entropy constraint item is defined as:

$$\text{Ent}(S^g) = \frac{1}{K} \sum_{k=1}^{K} \left( \sum_{i=1}^{N} \sum_{j=1}^{N} (-S_{g(k,i,j)}^g \log(S_{g(k,i,j)}^g)) \right)$$ (7)

We will maximize the entropy constraint item. The loss function for the intra-correlation branch is defined as:

$$\mathcal{L}_g = \lambda_1 \text{Div}(T^g, S^g) - \lambda_2 \text{Ent}(S^g)$$ (8)

where $\lambda_1$ and $\lambda_2$ are used to trade off the loss items. The optimization of intra-correlation $S^g$ is actually an alternating process with the guidance of target correlation $T^g$ (see App. A.1), ultimately resulting that each normal patch can establish strong correlations with most normal patches.

3.4. Inter-Correlation Learning

Through the intra-correlation branch, we can establish patch-to-patch correlations within a single image. However, an image usually doesn’t contain all possible normal patterns, which may cause it difficult to distinguish some ambiguous abnormal patches (see rows 2, 3, 4 in Figure 1) only through the intra-correlations. To address this, we should effectively take advantage of the known normal patterns, which may cause it difficult to distinguish some normal patterns. Specifically, we propose to establish inter-correlations with the whole normal training set. In this branch, the features of each patch establish a discrete inter-correlation distribution with the reference features extracted from all normal samples (see Figure 2). The inter-correlation matrix of the $k$th layer is similar to the corresponding intra-correlation, and is defined as:

$$\text{Inter Correlation} : S^g_k = \text{Softmax}\left(Q^g_k(K^g_k)^T/\sqrt{d_m}\right)$$

$$Q^g_k, K^g_k, V^g_k = X_{k-1}W^{Q,k}, X_fW^{K,k}, X_fW^{V,k}$$ (9)

where $X_f \in \mathbb{R}^{N_c \times d_e}$ represents the external reference features, $N_c$ is the length of the reference features and $d_e$ is the feature dimension. $W^{Q,k} \in \mathbb{R}^{d_m \times d_m}$ and $W^{K,k}, W^{V,k} \in \mathbb{R}^{d_e \times d_m}$ are learnable matrices for $Q^g_k, K^g_k, V^g_k$. $S^g_k \in \mathbb{R}^{N \times N}$ denotes the learned inter-image correlations.

**Loss.** The loss function for the inter-correlation branch has the opposite optimization direction to $\mathcal{L}_g$. Because the external reference features contain more comprehensive normal patterns, normal patches can establish stronger correlations with the closest reference normal patterns, instead of establishing strong correlations with most reference patterns. By contrast, it shall be harder for anomalous patches to establish strong correlations with any of the reference patterns. So the inter-correlation distributions of anomalies are more dispersed, while the normal inter-correlation distributions are more likely to be concentrated. To this end, we maximize the KL divergence and minimize the entropy item in the training process. The practical optimization strategy is also opposite to the intra-correlation branch (see App. A.1 for details). The loss function is defined as:

$$\mathcal{L}_c = -\lambda_1 \text{Div}(T^c, S^c) + \lambda_2 \text{Ent}(S^c)$$ (10)

**External Reference Features.** External reference features are used for providing accumulated knowledge of normality for the inter-correlation learning branch. Thus, these features should represent all possible normal patterns of all normal samples from the whole normal training set. To this end, we can employ many methods to generate the reference features, such as sampling key features by coreset subsampling algorithm [35], generating prototype features by memory module [28], or learning codebook features through vector quantization [65] or sparse coding techniques [55]. However, because the RBF-kernel in $T^c$ is position-sensitive, our reference features are better to preserve the positional information. The ablation results in App.B.2 show that the methods that can’t preserve the position information perform worse. From comprehensive ablation studies, we find that using patch-wise averaged features as the external reference features is a simple but effective way. We think that features extracted by deep neural networks are highly redundant [16, 52], and different normal patterns generally correspond to larger activation values at different channels in the feature vector [61]. So feature averaging will not eliminate some rare normal patterns [11], these patterns may be preserved at specific channels. And averaging can greatly reduce the feature redundancy, making the obtained reference features more representative.

Formally, for position $(i, j)$, we first extract the set of patch features at $(i, j)$, $X_{ij} = \{x_{ij}^k\}, k \in [1, N]$ from the $N$ normal training images. Then, the reference features at position $(i, j)$ is computed as $x_{ij}^c = \frac{1}{N} \sum_{k=1}^{N} x_{ij}^k$. The final external reference features are composed of averaged features at all locations and then flattened into 1D: $X_f = \text{Flatten}(\{x_{ij}^c\})$.

3.5. I2Correlation

We further combine the intra-correlation and inter-correlation branches to form the I2Correlation block. The output features of the intra- and inter-correlation branches are defined as: $Z_{g,k}^g = S_{g,k}^g V_{g,k}^g$ and $Z_{c,k}^g = S_{c,k}^g V_{c,k}^g$, respectively. Then, we use the residual feature $Z_{g,k}^{res} = Z_{g,k}^g - Z_{c,k}^g$ as the output of the I2Correlation block. The feature $Z_{g,k}^{res}$ is generated from the external reference features, which can contain rich normal patterns. Thus, by subtracting $Z_{c,k}^g$ from the feature $Z_{g,k}^g$, it is conducive to spotlight the abnormal patterns in the $Z_{g,k}^{res}$. This is beneficial for anomaly detection.
The total loss function consists of the three branch loss functions, and is combined as follows:

\[
L_{\text{total}} = L_i + L_o + L_c \tag{11}
\]

### 3.6. Anomaly Scoring

We utilize reconstruction errors as the baseline anomaly criterion and incorporate the normalized correlation distance into the reconstruction criterion. The final anomaly score of the \( i \)th patch in the input patch sequence is shown as follows:

\[
s_i = \left[ \|X_{i,:} - \hat{X}_{i,:}\|_2 + (1 - \frac{X_{i,:} \cdot \hat{X}_{i,:}}{\|X_{i,:}\|_2 \|\hat{X}_{i,:}\|_2}) \right] \odot \left(1 - \text{Softmax} \left(- \text{Div} (T_{i,:}, \mathcal{S}_{i,:})\right) \right) \tag{12}
\]

where \( \odot \) is the element-wise multiplication.

### 4. Experiments

#### 4.1. Experimental Setup

##### Datasets.
We extensively evaluate our approach on two widely used industrial AD datasets: the MVTecAD [3] and BTAD [26], and one recent challenging dataset: the MVTecAD-3D [5]. MVTecAD is established as a standard benchmark for evaluating unsupervised image anomaly detection methods. This dataset contains 5354 high-resolution images from 15 real-world categories. 5 classes consist of textures and the other 10 classes contain objects. A total of 73 different anomaly types are presented. BTAD is another popular benchmark for unsupervised image anomaly detection, which contains 2540 RGB images of three industrial products. All classes in this dataset belong to textures. MVTecAD-3D is recently proposed, which contains 4147 2D RGB images paired with high-resolution 3D point cloud scans from 10 real-world categories. Even though this dataset is mainly designed for 3D anomaly detection, most anomalies can also be detected only through RGB images without 3D point clouds. Since we focus on image anomaly detection, we only use RGB images of the MVTecAD-3D dataset. We refer to this subset as MVTec3D-RGB.

##### Evaluation Metrics.
The standard metric in anomaly detection, AUROC, is used to evaluate the performance of AD methods. Image-level AUROC is used for anomaly detection and a pixel-based AUROC for evaluating anomaly localization [3, 4, 11, 64].

##### Implementation Details.
We use EfficientNet-b6 [45] to extract two levels of feature maps with \( 8 \times 16 \times \) downsampling ratios, the pre-trained networks are from the timm library [53]. Then, we construct a subsequent transformer network (see Figure 2) at each feature level to reconstruct patch features and learn patch-to-patch correlations. The parameters of the feature extractor are frozen in the training process, only the parameters of the subsequent transformer networks are learnable. All the subsequent transformer networks in our model contain 3 layers. We set the hidden dimension \( d_{m} \) as \( 256, 512 \) and the number of heads as 8. The hyperparameters \( \lambda_1 \) and \( \lambda_2 \) are set as 0.5 and 0.5 to trade off the two parts of the \( L_{i} \) and \( L_{o} \) loss functions (see App. B.2 for hyperparameter sensitivity analysis). We use the Adam [19] optimizer with an initial learning rate of \( 1e^{-4} \). The total training epochs are set as 100 and the batch size is 1 by default. All the training and test images are resized and cropped to \( 256 \times 256 \) resolution from the original resolution.

#### 4.2. Main Results

##### SOTA Methods.
We extensively compare our method with those published SOTA methods in the past three years.

---

Table 1. Comparison of our method with the SOTA methods on the MVTecAD dataset.

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Venue</th>
<th>Image-level AUROC</th>
<th>Patch-level AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capet</td>
<td>DRAEM [63]</td>
<td>CVPR 2020</td>
<td>0.977</td>
<td>0.997</td>
</tr>
<tr>
<td>Grid</td>
<td>PaDiM [41]</td>
<td>ICCV 2021</td>
<td>0.956</td>
<td>0.996</td>
</tr>
<tr>
<td>Leather</td>
<td>UFR [39]</td>
<td>NeurIPS 2021</td>
<td>0.930</td>
<td></td>
</tr>
<tr>
<td>Tile</td>
<td>FCDD [25]</td>
<td>ICLR 2021</td>
<td>0.927</td>
<td></td>
</tr>
<tr>
<td>Wood</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.972</td>
<td>0.963</td>
</tr>
<tr>
<td>Bottle</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.975</td>
<td>0.963</td>
</tr>
<tr>
<td>Cable</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.976</td>
<td>0.968</td>
</tr>
<tr>
<td>Capule</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.977</td>
<td>1.000</td>
</tr>
<tr>
<td>Hazelnut</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>Metal nut</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>Pill</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>Screw</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>Toothbrush</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>Transistor</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>Zipper</td>
<td>PaDiM [42]</td>
<td>CVPR 2021</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>Mean</td>
<td>PaDiM [42]</td>
<td>CVPR 2021</td>
<td>0.999</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 2. Detailed image-level AUROCs on the MVTecAD dataset.

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Venue</th>
<th>Image-level AUROC</th>
<th>Patch-level AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carpet</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.972</td>
</tr>
<tr>
<td>Grid</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Leather</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Tile</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Wood</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Bottle</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Cable</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Capule</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Hazelnut</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Metal nut</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Pill</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Screw</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Toothbrush</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Transistor</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Zipper</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
<tr>
<td>Mean</td>
<td>PaDiM [41]</td>
<td>CVPR 2021</td>
<td>0.950</td>
<td>0.963</td>
</tr>
</tbody>
</table>
The comparison results on MVTecAD are shown in Table 1. Then, we select five reproducible methods to report the detailed results on MVTecAD; these methods are representative and SOTA AD methods in the mainstream categories of image anomaly detection as we discuss in Related Work, including: DRAEM [64], PatchSVDD [60], PatchCore [35], MKD [41], CFLOW [15]. For a fair comparison, we reproduce all these methods with the same backbone as in our model. Thus, despite using the unmodified code from the official repositories, we are not able to exactly reproduce the original results, but our numbers are very close. The detailed image-level AUROC results are shown in Table 2, and detailed pixel-level AUROC results are in the App. Table 6. Most of the methods in Table 1 don’t report results on the BTAD and MVTec3D-RGB datasets, so we reproduce the five representative methods on the two datasets for comparison. The image-level and pixel-level AUROC results on BTAD and MVTec3D-RGB are shown in Table 3. Additional detailed results are in the App. Table 7, 8.

Anomaly Detection. On MVTecAD, we set the SOTA performance on the mean detection AUROC, which is slightly higher than the best competitor, PatchCore [35]. Note that the results in Table 2 show that our method can achieve much better results than PatchCore when using the same backbone. What’s more, in addition to the pill, screw and toothbrush classes, our method achieves more than 99% AUROC in all other classes, while other methods only achieve more than 99% AUROC in most nine classes. This shows that our method is more stable and effective in real-world applications. In the classes (e.g. metal nut, screw, and transistor) with more global and logical anomalies, our method can achieve significantly better results than those methods depending on patch-wise discrepancy (e.g. DRAEM [64] and MKD [41]), and can also achieve better results compared with other methods. This verifies that the three discrepancies are complementary factors and our model can simultaneously spot these discrepancies to recognize harder global and logical anomalies. On BTAD, our FOD can achieve 96.0% mean detection AUROC, which can outperform the best competitor CFLOW [15] by a margin of 1.2%. On MVTec3D-RGB, we can outperform all previous SOTA methods by a margin of 3.3%. Note that this dataset is much more challenging than the MVTecAD dataset when comparing the best results (99.2% for MVTecAD vs. 88.4% AUROC for MVTec3D-RGB). This demonstrates the robustness of our method.

Anomaly Localization. Our method can achieve comparable results with the best competitors on all three datasets. Our method is slightly lower than CFLOW [15] and PatchCore [35] on the three datasets, but we observe that our model that considers the intra- and inter-correlations outperforms the vanilla reconstruction AD models (e.g. DRAEM [64] and MKD [41]) on all three datasets, which verifies the effectiveness of correlation modeling. Compared with patch-wise reconstruction methods, our method is more effective and robust to detect hard global and logical anomalies (see Figure 3).

Qualitative Results. Samples in Figure 1 show qualitative impressions of accurate anomaly localization from our method. It can be found that our approach can achieve the best anomaly score maps by combining the three discrepancies. Additional visualizations are in the App. Figure 6.

4.3 Ablation Study and Model Analysis

Ablation Study. To explain how our model works effectively, we further investigate the effect of the three key designs in our model: recognition views, entropy constraint and reference features. The quantitative results are shown in Table 4, more results can be found in App. Table 9. The entropy constraint is quite effective and necessary in the intra-correlation branch. Specifically, it brings a remarkable 7.7% averaged absolute AUROC promotion, which demonstrates that the entropy constraint strategy is really conducive to increase the distinguishability between abnormal and normal. Only utilizing pure reconstruction criterion or pure KL divergence can’t get the most superior detection results, the combined criterion can outperform each single criterion consistently by a margin of 1.0% and 9.8%. Thus the reconstruction errors and the intra- and inter-correlations can collaborate to improve detection performance. For the external reference features, we compare the simple mean features with more elaborate coreset features [35] (see App. Table 15 for more comparison methods). The results show that the simple mean features are more effective than the coreset features (0.923 vs. 0.836), which means that mean features are effective enough to represent all possible normal patterns and preserve the positional information. Finally, our proposed FOD surpasses the pure reconstruction Transformer by 4.0% absolute improvement. These verify that our proposed explicit correlation learning approach is effective.

Effect for Different Anomalies. To illustrate the effectiveness of our model intuitively, we explore the anomaly localization quality under different types of anomalies (see Figure 3). E.g., simple local anomalies (Column 1) are obviously different from normal visuals; hard global anomalies (Columns 2, 3) have less obvious visual appearances,
Humans recognize anomalies through two aspects: patch-wise representation discrepancies and weak patch-to-patch correlations. In this paper, we propose a novel AD framework: FOcus-the-Discrepancy, to simultaneously spot the patch-wise, intra- and inter-discrepancies. The patch-wise discrepancies and intra- and inter-correlations are complementary factors, and can be combined to detect more complex and diverse anomalies. The major characteristic of our method is that we renovate the self-attention maps in transformers to I2Correlation to explicitly establish intra- and inter-image correlations for AD modeling. The combination of explicit correlation learning and transformer architecture can match the core idea of anomaly detection quite well.

Acknowledgements

This work was supported in part by the National Natural Science Fund of China (61971281), the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Science and Technology Commission of Shanghai Municipality (22DZ2229005).
References


[6] Paul Bergmann, Sindy Löwe, Michael Fauser, and David Sattlegger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In ICCVTA, 2019. 2


[13] Xiaohan Ding, Xiangyu Zhang, Yizhuang Zhou, Jungong Han, Guiguang Ding, and Jian Sun. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In CVPR, 2022. 3


<table>
<thead>
<tr>
<th>Number</th>
<th>Author(s)</th>
<th>Title</th>
<th>Conference</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny</td>
<td>Barlow twins: Self-supervised learning via redundancy reduction</td>
<td>In ICML</td>
<td>2021</td>
</tr>
<tr>
<td>67</td>
<td>Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yunwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang</td>
<td>Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers</td>
<td>In CVPR</td>
<td>2021</td>
</tr>
<tr>
<td>68</td>
<td>Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer</td>
<td>Spot-the-difference self-supervised pre-training for anomaly detection and segmentation</td>
<td>In ECCV</td>
<td>2022</td>
</tr>
</tbody>
</table>

---

6813