# Filter or Compensate: Towards Invariant Representation from Distribution Shift for Anomaly Detection

**Zining Chen**[1], **Xingshuang Luo**[1], **Weiqiu Wang**[1], **Zhicheng Zhao**[1,2,3*], **Fei Su**[1,2,3], **Aidong Men**[1]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]Beijing Key Laboratory of Network System and Network Culture, China
[3]Key Laboratory of Intereactive Technology and Experience System, Ministry of Culture and Tourism, Beijing, China
chenzn@bupt.edu.cn, {luoxingshuang,wangweiqiu,zhaozc,sufei,menad}@bupt.edu.cn

## Abstract

Recent Anomaly Detection (AD) methods have achieved great success with In-Distribution (ID) data. However, real-world data often exhibits distribution shift, causing huge performance decay on traditional AD methods. From this perspective, few previous work has explored AD with distribution shift, and the distribution-invariant normality learning has been proposed based on the Reverse Distillation (RD) framework. However, we observe the misalignment issue between the teacher and the student network that causes detection failure, thereby propose FiCo, **Fi**lter or **Co**mpensate, to address the distribution shift issue in AD. FiCo firstly compensates the distribution-specific information to reduce the misalignment between the teacher and student network via the Distribution-Specific Compensation (DiSCo) module, and secondly filters all abnormal information to capture distribution-invariant normality with the Distribution-Invariant Filter (DiIFi) module. Extensive experiments on three different AD benchmarks demonstrate the effectiveness of FiCo, which outperforms all existing state-of-the-art (SOTA) methods, and even achieves better results on the ID scenario compared with RD-based methods.

**Code** — https://github.com/znchen666/FiCo

## Introduction

Anomaly detection (AD) has been extensively researched and plays a critical role in numerous applications. Its main objective is to identify anomalous patterns within large amounts of data. Real-world applications, such as manufacturing quality control (Bergmann et al. 2019), video surveillance (Liu et al. 2018), and medical monitoring (Schlegl et al. 2019), are in high demand for accurate and robust AD algorithms. In most scenarios, acquiring labeled anomaly data is challenging and expensive. As a result, unsupervised anomaly detection has become the prevailing focus of research. To address this issue, previous studies have made efforts from various aspects, such as reconstruction-based (Ristea et al. 2022; Zavrtanik, Kristan, and Skočaj 2022; Zhang et al. 2023; Zhang, Xu, and Zhou 2024), embedding-based (Roth et al. 2022; Yu et al. 2021; Huang et al. 2022; Liu et al. 2023; Zhu and Pang 2024; Lee and Choi 2024), and knowledge distillation-based (Bergmann et al. 2020; Salehi et al. 2021; Deng and Li 2022; Tien et al. 2023; Gu et al. 2023) approaches, etc., which have led to significant advancements recently.

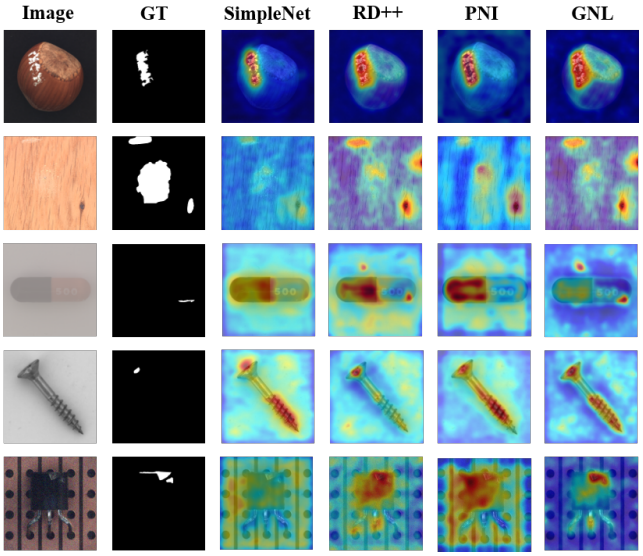



Figure 1: Anomaly map from different scenarios of SOTA AD methods (Bae, Lee, and Kim 2023; Liu et al. 2023; Tien et al. 2023; Cao, Zhu, and Pang 2023) on the MVTec benchmark (Bergmann et al. 2019). The image of each row represents a different scenario, including ID and four OOD (brightness, contrast, defocus blur and gaussian noise) scenarios.

These methods mostly assume training and test sets are In-Distribution (ID), thus the only purpose is to identify anomalies from normal ones without distribution shift. However, the assumption is not realistic in real-world scenarios, causing huge performance decay when confronting real-world Out-of-Distribution (OOD) data (Bae, Lee, and Kim 2023; Liu et al. 2023; Tien et al. 2023), as shown in Fig. 1. The test data possibly have both anomalous patterns and distribution shifts, and most methods show accurate anomaly map merely on the ID scenario without distribution shift, but

*Corresponding author

are negatively affected by the distribution shift from OOD scenarios. From this perspective, we investigate the research for resolving the distribution shift between ID and OOD data on various downstream applications, such as image classification (Zhou et al. 2021; Xu et al. 2021; Cha et al. 2021; Mahajan, Tople, and Sharma 2021; Lv et al. 2022; Chen et al. 2024), semantic segmentation (Choi et al. 2021; Zhao et al. 2022; Huang et al. 2023) and person re-identification (Liao and Shao 2020, 2022; Chen et al. 2023), etc. The methods can be mainly categorized into data augmentation (Xu et al. 2021; Zhao et al. 2022), domain-invariant learning (Mahajan, Tople, and Sharma 2021; Lv et al. 2022), and learning strategies (Cha et al. 2021; Liao and Shao 2022), where domain-invariant learning has been the mainstream and achieves competitive results with relatively low computational costs.

Then (Cao, Zhu, and Pang 2023) proposes the first work on AD under multiple OOD scenarios by designing the distribution-invariant normality learning. The method is based on the Reverse Distillation (RD) (Deng and Li 2022) framework, where the output from the teacher network flows to the student network via a one-class bottleneck module (OCBE). It learns invariant representation via consistency between multiple augmentations to filter distribution-specific information. However, we observe that there's information misalignment between the teacher and student network. Thus, the question emerges that "What information should be filtered or compensated to obtain the invariant representation in the RD framework?"

This paper revisits the efficient and effective RD framework from the perspective of invariant representation to tackle the distribution shift issue. We observe that one drawback in (Cao, Zhu, and Pang 2023) is the information loss on distribution-specific representation, while another drawback lies in the absence of an explicit mechanism to better learn distribution-invariant normality. Therefore, we firstly compensate for the distribution-specific information in the student network, while secondly filtering irrelevant information to achieve distribution-invariant normality. In conclusion, this paper proposes Filter or Compensate (FiCo) method to thoroughly explore the invariant representation to resolve the distribution shift, surpassing state-of-the-art (SOTA) methods on multiple AD benchmarks with a relatively large margin.

The main contributions of the paper can be summarized as follows,

- We propose FiCo for better invariant representation learning to address the distribution shift issue in AD task. Firstly, the Distribution-Invariant Filter (DiIFi) module is proposed to filter all abnormal information for distribution-invariant normality, including anomlous patterns and distribution-specific information.
- Secondly, the Distribution-Specific Compensation (DiSCo) module is designed to compensate for distribution-specific information, thereby reducing the misalignment between the teacher and student network. Consequently, merely the anomalous pattern is taken into account during the inference without the interference of the distribution-specific information.
- Extensive experiments on different AD benchmarks with OOD scenarios manifest the superior performance of our method compared with SOTA AD methods. FiCo not only achieves better performance on OOD scenarios, but also improves accuracy on ID scenario to surpass RD-based methods.

## Related Work

### Anomaly Detection

Recent anomaly detection methods have been dispersed into various categories, while the mainstream research can be coarsely categorized into reconstruction-based, embedding-based and knowledge distillation-based methods. For the reconstruction-based methods, AutoEncoder (AE) (Kingma and Welling 2013) and Generative Adversarial Network (GAN) (Goodfellow et al. 2020) are widely adopted as the generative models to reconstruct samples. Then, research has been expanded to various aspects to address the issue, including the memory module (Park, Noh, and Ham 2020; Hou et al. 2021; Gu et al. 2023), pseudo-anomaly augmentation (Li et al. 2021; Schlüter et al. 2022) and diffusion model (Wyatt et al. 2022; Zhang et al. 2023; Zhang, Xu, and Zhou 2024), etc. Embedding-based methods show strong improvement in recent literatures by simply using pretrained networks for feature extraction. These methods identify anomalies by the input feature embedding with the normal feature distribution via different standards (Cohen and Hoshen 2020; Yu et al. 2021). Meanwhile, different spatial feature are specifically designed for measurement (Defard et al. 2021; Roth et al. 2022; Bae, Lee, and Kim 2023; Yao et al. 2023). (Reiss et al. 2021; Deecke et al. 2021) introduce different modules for adaptation to the distribution of target dataset. Recently, text-based AD has emerged to leverage the capability of CLIP for textual knowledge with text prompts (Zhu and Pang 2024; Lee and Choi 2024).

Knowledge distillation is a promising solution for anomaly detection that the student network learns the anomaly-free feature from the teacher network and detects abnormal one based on the discrepancy. (Bergmann et al. 2020) ensembles multiple student networks for more discriminate feature. (Salehi et al. 2021) designs feature-level distillation at various layers of the pretrained expert network. (Deng and Li 2022) proposes the reverse distillation framework that the one-class embedding from the teacher network flows to the student network to restore multi-scale feature. (Tien et al. 2023) improves (Deng and Li 2022) on feature compactness by designing optimal transport loss, and anomalous signal suppression by simulating pseudo-anomaly samples. (Gu et al. 2023) designs the normality recall memory to store normal information based on RD framework to tackle "normality forgetting" issue. (Cao, Zhu, and Pang 2023) proposes distribution-invariant normality learning to tackle the distribution shift issue by introducing consistency loss on different augmented views.

### Out-of-Distribution Generalization

Out-of-Distribution (OOD) issue is essential in various downstream tasks, where methods can be roughly categorized into three aspects, including data augmentation (Zhou et al. 2021; Xu et al. 2021), domain-invariant learning (Mahajan, Tople, and Sharma 2021; Lv et al. 2022), and learning strategies (Cha et al. 2021; Liao and Shao 2022). Data augmentation has been a simple yet effective technique in OOD generalization by synthesizing novel images and features. Domain-invariant learning has been the mainstream solution for OOD generalization (Sun and Saenko 2016; Lv et al. 2022). Learning strategies such as meta learning (Li et al. 2018a), adversarial learning (Li et al. 2018b), gradient optimization (Foret et al. 2020) also boost the research on fundamental training protocols.

However, the AD task differs from those downstream tasks that no class or domain label is available, which cannot meet the requirements with many aforementioned solutions. From this perspective, domain-invariant learning is plausible and promising for its simplicity and effectiveness. Thus, (Cao, Zhu, and Pang 2023) proposes the distribution-invariant normality learning by introducing common augmentations to filter distribution-specific information based on the RD (Deng and Li 2022) framework. Nevertheless, it neglects the mechanism of the RD framework with coarse consistency at different spatial levels, which harms the invariant representation from the student network.

## Preliminaries

### Task Description

Let $(x_s, y_s)$, $(x_t, y_t)$ denote the training and test samples with the label indicating anomalies, and suppose $\mathcal{X}_{id}$, $\mathcal{X}_{ood}$ are ID and OOD distributions. During the training process, the dataset merely contains normal samples with ID distribution, $\mathcal{I}_s = \{(x_s \in \mathcal{X}_{id} \mid y_s = 0)\}$. However, in the inference stage, the test dataset contains both normal and anomalous samples with different distribution, $\mathcal{I}_t = \{(x_t \in \mathcal{X}_{id} \cup \mathcal{X}_{ood} \mid y_t = 0, 1)\}$, where $0, 1$ denote normal and anomalous samples, respectively. The goal of the task is to train models on the training datasets with only normal and ID samples, and generalize well on the unseen test dataset with anomalies and distribution shifts. Note that no data from test dataset is available during the training process.

### Briefly Review on RD-based Methods

Reverse distillation for anomaly detection is first proposed in (Deng and Li 2022), which consists of a frozen pretrained teacher network $E(\cdot)$, a one-class embedding (OCBE) module $\phi(\cdot)$ and a student network $D(\cdot)$. Unlike previous methods on knowledge distillation, the output of the pretrained teacher network is sequentially passed through the OCBE module and the student network, which means the high-level semantic knowledge flows to the student first. During the training process, cosine similarity is adopted to formulate the loss function,

$$\mathcal{L}_{RD} = \sum_{k=1}^{K} \{1 - \frac{f^{E_k} \cdot f^{D_k}}{\|f^{E_k}\| \|f^{D_k}\|}\} \tag{1}$$

where $K$ is the total number of layers, $E_k$ and $D_k$ are the $k^{th}$ layer of the encoder and the decoder, and $f^{E_k}$, $f^{D_k}$ are the feature maps of $x_s$ from the $k^{th}$ block of the teacher and student network, respectively. During inference, the multi-scale similarities of representation are utilized for evaluation, where low similarity score indicates anomalies.

Afterwards, GNL (Cao, Zhu, and Pang 2023) proposes the distribution-invariant normality learning to tackle the distribution shift issue. It introduces multiple augmentations on the training sample to synthesize $x_s^n$, where $n$ is the $n^{th}$ augmented view of sample $x_s$. Then it designs consistency loss at both the OCBE module and the final output of the student network as $\mathcal{L}_{abs}$ and $\mathcal{L}_{lowf}$ to filter distribution-specific information,

$$\mathcal{L}_{abs} = \sum_{n=1}^{N} \{1 - \frac{(f^{\phi})^T \cdot f_n^{\phi}}{\|f^{\phi}\| \|f_n^{\phi}\|}\} \tag{2}$$

$$\mathcal{L}_{lowf} = \sum_{n=1}^{N} \{1 - \frac{(f^{D_1})^T \cdot (f_n^{D_1})}{\|f^{D_1}\| \|f_n^{D_1}\|}\} \tag{3}$$

where $f^{\phi}$, $f_n^{\phi}$ are the output representation of original image and the $n^{th}$ augmented image from the OCBE module, while $f^D$, $f_n^D$ are the final output from the student network. Then the final loss can be formulated as,

$$\mathcal{L}_{GNL} = \mathcal{L}_{RD} + \mathcal{L}_{abs} + \mathcal{L}_{lowf} \tag{4}$$

During inference, (Cao, Zhu, and Pang 2023) utilizes the existing Test-Time Augmentation (TTA) technique EFDM (Zhang et al. 2022) to minimize the discrepancy between the distribution of the test sample and the normal sample.

## Approach

Our method FiCo resolves the drawbacks of distribution-invariant normality learning in (Cao, Zhu, and Pang 2023) by designing additional modules and loss functions. The overall architecture is presented in Fig. 2.

### Distribution-Specific Compensation Module

GNL (Cao, Zhu, and Pang 2023) designs training objectives to align the teacher and student network for distribution-invariant normality, but neglects the distribution shifts from the OOD samples in the test dataset. As a result, the misalignment of distribution-specific information confuses the model to misclassify the distribution shifts as anomalies. Therefore, we aim to compensate for the distribution-specific information to prevent the misleading effect of the distribution shifts. Specifically, assume that the representation map from each block of the student network consists of distribution-invariant information and distribution-specific information. As the inference process of RD framework is to calculate the multi-scale similarities between the representations of the pre-trained teacher network and the student network, the misalignment on distribution-specific information causes the discrepancy which is prone to be recognized as anomalous patterns. Under this observation, we investigate how to compensate for distribution-specific information to guarantee the alignment between the teacher and student network when inferring on OOD samples. Therefore,

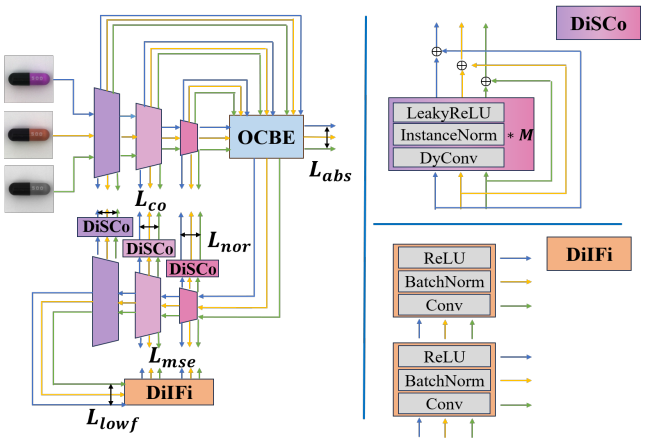



Figure 2: The overall architecture of our method FiCo, including detailed structure of DiSCo module and DiIFi module with designed losses. DiSCo module aims to compensate for the distribution-specific information by $\mathcal{L}_{Co}$ to prevent misalignment. DiIFi module attempts to filter abnormal patterns to obtain invariant normality with $\mathcal{L}_{Fi}$, including $\mathcal{L}_{lowf}, \mathcal{L}_{mse}, \mathcal{L}_{nor}$. $\mathcal{L}_{abs}$ indicates the consistency loss between original and augmented representation at OCBE module.

we propose the DiSCo module to take responsibility for preventing distribution-specific information loss.

Suppose the DiSCo module as $C_k(\cdot)$ which is inserted after each block of the student network. The DiSCo module receives the feature map from each block and reconstructs distribution-specific information. To prevent representation loss, the shortcut is adopted and the final output can be described as,

$$f_F^{D_k} = C_k(f^{D_k}) + f^{D_k}, \qquad f_{F,n}^{D_k} = C_k(f_n^{D_k}) + f_n^{D_k} \quad (5)$$

where $f_n^{D_k}$ denotes the output of the $n^{th}$ augmented sample from the $k^{th}$ block of the student network, and $f_F^{D_k}$ and $f_{F,n}^{D_k}$ are the output from original and augmented views after the shortcut on DiSCo modules. Note that if $n$ is omitted, the symbol represents the original image without augmentation. Then we design the loss function to ensure the valid compensation, which can be formulated as,

$$\mathcal{L}_{Co} = \sum_{k=1}^{K}\{1 - \frac{f^{E_k} \cdot f_F^{D_k}}{\|f^{E_k}\|\|f_F^{D_k}\|}\} + \alpha \sum_{n=1}^{N}\sum_{k=1}^{K}\{1 - \frac{f_n^{E_k} \cdot f_{F,n}^{D_k}}{\|f_n^{E_k}\|\|f_{F,n}^{D_k}\|}\} \quad (6)$$

where $f_n^{E_k}$ is the output of the $n^{th}$ augmented sample from the $k^{th}$ block of the teacher network and $\alpha$ is the hyper-parameter to balance the ratio.

Furthermore, for the constitution of the DiSCo module, considering the discrepancy between the training dataset and the OOD test dataset, DyConv (Chen et al. 2020) is utilized to introduce attention to differentiate diverse feature distribution, followed with InstanceNorm (Ulyanov, Vedaldi, and Lempitsky 2016) and LeakyReLU (Xu et al. 2015) for normalization and activation. The DiSCo module consists of $M$ blocks of DyConv, InstanceNorm, LeakyReLU and can be trained end-to-end in the student network.

## Distribution-Invariant Filter Module

RD framework and the following improvement have made assumptions that anomalous patterns should be constrained towards the student network, so that the student network merely reconstructs normal patterns (Deng and Li 2022; Tien et al. 2023). As a result, the discrepancy between the teacher and student network can be maximized when confronting anomalies. From this perspective, we perceive that filtering all abnormal information in the student network to obtain the invariant normality naturally promotes invariant representation learning. However, GNL (Cao, Zhu, and Pang 2023) designs $L_{lowf}$ at the final block of the student network for consistency on diverse augmented views of a single sample, but no explicit mechanisms on the previous blocks are exploited. Therefore, we attempt to incorporate the filter module at previous blocks of the student network to acquire invariant representation.

We analyze that the output of the final block is low-level information, including edges, colors, shapes, etc., where the consistency between different augmented views successfully captures invariant representation. However, the output from previous blocks are high-level semantics with distribution-specific information. As a result, directly applying the same function as $\mathcal{L}_{lowf}$ on previous blocks suffers from semantic information loss. Therefore, the distribution-invariant filter (DiIFi) module is designed to filter distribution-specific information from previous blocks, by imitating what the final DiSCo module $C_1(\cdot)$ recognizes as distribution-specific information. Specifically, let the DiIFi module as $I(\cdot)$ which consists of K-1 ConvBlocks (Convolution, BatchNorm, ReLU), and $I_k(\cdot)$ indicates the operation on the $k^{th}$ block to transform the distribution-specific information $C_1(f_n^{D_1}) \in \mathbb{R}^{C\times H\times W}$ learned from the final DiSCo module. The formulation of sequential transformation of distribution-specific information to previous blocks is,

$$f_{k,n}^{D_1} = \begin{cases} I_k(C_1(f_n^{D_1})) \in \mathbb{R}^{2C\times \frac{H}{2}\times \frac{W}{2}}, & k = 2 \\ I_k(f_{k-1,n}^{D_1}) \in \mathbb{R}^{2^{k-1}C\times \frac{H}{2^{k-1}}\times \frac{W}{2^{k-1}}}, & k > 2 \end{cases} \quad (7)$$

where $f_{k,n}^{D_1}$ denotes the transformed distribution-specific information of the $n^{th}$ augmented sample from the $k^{th}$ block of the student network. The DiIFi module attempts to align the transformed distribution-specific feature $f_{k,n}^{D_1}$ with the corresponding compensated feature $C_k(f_n^{D_k})$ from previous DiSCo modules. Therefore, the Mean Square Error (MSE) loss is adopted to minimize the discrepancy,

$$\mathcal{L}_{mse} = \sum_{k=2}^{K}(C_k(f^{D_k}) - f_k^{D_1})^2 + \sum_{n=1}^{N}\sum_{k=2}^{K}(C_k(f_n^{D_k}) - f_{k,n}^{D_1})^2 \quad (8)$$

where the first and the second item are operations on original and augmented views, respectively. The DiIFi module trained with $\mathcal{L}_{mse}$ has two-fold merits. Firstly, DiIFi module impels all previous DiSCo modules to mimic what the final DiSCo module learns. As the final representation

consists of affluent low-level information, the transformation from which can prevent previous DiSCo modules to learn biased distribution-specific information. Secondly, as the DiSCo modules are supervised by $\mathcal{L}_{Co}$ to compensate for distribution-specific information, the DiIFi module implicitly promotes residual blocks to filter all abnormal information, including distribution-specific and anomalous patterns to learn invariant normality, as shown in Fig. 4.

Lastly, to prevent the input of the DiIFi module from normality collapsing that the compensated distribution-specific information is biased from augmented representation, we also incorporate the consistency loss based on cosine similarity after the final DiSCo module,

$$\mathcal{L}_{nor} = \sum_{n=1}^{N} \{1 - \frac{(C_1(f^{D_1}))^T \cdot (C_1(f_n^{D_1}))}{\|C_1(f^{D_1})\| \|C_1(f_n^{D_1})\|}\} \quad (9)$$

The overall filter loss can be described as,

$$\mathcal{L}_{Fi} = \mathcal{L}_{lowf} + \beta \mathcal{L}_{mse} + \gamma \mathcal{L}_{nor} \quad (10)$$

where $\beta, \gamma$ are the balancing hyper-parameters.

## Training and Inference

**Training.** The whole network is trained end-to-end with $L_{FiCo}$,

$$\mathcal{L}_{FiCo} = \mathcal{L}_{Fi} + \mathcal{L}_{abs} + \mathcal{L}_{Co} \quad (11)$$

where we maintain the original framework from (Cao, Zhu, and Pang 2023) to insert additional modules, and simply replace the $L_{RD}, L_{lowf}$ with $\mathcal{L}_{Fi}, \mathcal{L}_{Co}$ for the filter and compensation process.

**Inference.** During the inference process, all other settings are remained identical with (Cao, Zhu, and Pang 2023), including the test-time augmentation EFDM (Zhang et al. 2022) and the calculation process of sample-level anomaly score. The only difference is the additional DiSCo modules after all $K$ blocks from the student network are remained to compensate for the distribution-specific information, while the DiIFi module is discarded.

# Experiments

## Benchmarks

Experiments are conducted on three AD benchmarks with distribution shifts (Cao, Zhu, and Pang 2023), including MVTec (Bergmann et al. 2019), PACS (Li et al. 2017) and CIFAR-10 (Krizhevsky, Hinton et al. 2009). MVTec is a widely-used industrial AD dataset with 15 categories, including 5 categories for texture anomalies and 10 categories for object anomalies. It consists of 5,354 images, including 3,629 normal images from the training set and 1,725 images from the test set with both normal and abnormal one. PACS is a prevalent dataset from OOD classification with a total of 9,991 images from seven classes and four domains. CIFAR-10 is used as the benchmark of the one-class novelty detection task, including 10 categories with 50,000 and 10,000 images from the training and test set. All datasets follow the same procedure in (Cao, Zhu, and Pang 2023). For MVTec and CIFAR-10, diverse visual corruptions are conducted to generate the OOD scenarios. For PACS, we merely use images on the common photo domain as the training set, and infer on different test sets from all domains.

## Implementation Details

The backbone is WideResNet50 (Zagoruyko and Komodakis 2016) as widely adopted and all the images for MVTec, PACS are resized to 256 $\times$ 256, while 32 $\times$ 32 for CIFAR-10. All the additional modules can be trained end-to-end with Adam optimizer (Kingma and Ba 2014), and the initial learning rate is set to 0.005. The hyper-parameters $\alpha, \beta, \gamma$ that control the balancing ratio of different additional losses are 0.05, 0.02, 1 for MVTec and CIFAR-10 dataset, while $\beta$ is set to 0.1 for PACS. The number of blocks $M$ in DiSCo module is set to 4 for all datasets. Other relevant hyper-parameters, such as the number of augmentations $N$, the style blending ratio in EFDM (Zhang et al. 2022), and detailed operations are all maintained the same as (Cao, Zhu, and Pang 2023) for fair comparison. For the evaluation metrics, the Area Under the Receiver Operator Curve (AUROC) on the sample-level is adopted (Cao, Zhu, and Pang 2023), which is a universal assessment between the normal and anomalous samples.

## Comparison with State-of-the-Art Methods

We compare our proposed method FiCo with recent methods on anomaly detection, including Deep SVDD (Ruff et al. 2018), f-AnoGAN (Schlegl et al. 2019), KD (Salehi et al. 2021), RD (Deng and Li 2022), PatchCore (Roth et al. 2022), RD++ (Tien et al. 2023), SimpleNet (Liu et al. 2023), PNI (Bae, Lee, and Kim 2023), GNL (Cao, Zhu, and Pang 2023), RealNet (Zhang, Xu, and Zhou 2024). The results are the average performance on all classes. Note that $^\dagger$ denotes our implementation and otherwise is the reported performance in (Cao, Zhu, and Pang 2023).

**MVTec.** As shown in Table 1, our method FiCo surpasses all the prevalent methods on the average AUROC. Compared with SOTA methods PNI and RealNet, FiCo merely shows a slight decrease of 0.84% and 0.87% on ID performance, but improves 19.40% and 7.33% on the average performance of all OOD scenarios. Moreover, compared with all the RD-based methods, FiCo not only shows superiority on most OOD scenarios, but also has improved the performance on ID scenario to achieve SOTA performance. Furthermore, FiCo exceeds GNL on ID scenario with 1.04% and all OOD scenarios with an average of 1.22% that demonstrates the effectiveness of our method.

**PACS.** Results in Table 2 show that our method FiCo has superior generalizablity when confronting real-world data with OOD scenarios. All the prevalent AD methods suffer from huge performance decay, especially on the sketch domain, such as 14.84% drop on SimpleNet compared with our method FiCo. Instead, FiCo is capable to resolve the distribution shift issue that improves performance on all scenarios compared with RD-based methods, with 8.14% and 4.03% average improvement on RD and RD++. Meanwhile, FiCo surpasses GNL on all scenarios with an average of 2.94% that demonstrate the significance of our additional modules and losses.

**CIFAR-10.** Table 3 presents the conventional one-class novelty detection task on CIFAR-10. One-class novelty detection is another form of anomaly detection where merely

| Method | ID | OOD | | | | Avg. |
|---|---|---|---|---|---|---|
| | Ori | Br | Co | Bl | No | |
| Deep SVDD | 70.0 | 55.2 | 50.1 | 68.8 | 59.1 | 60.6 |
| f-AnoGAN | 75.7 | 48.4 | 49.3 | 38.0 | 39.1 | 50.1 |
| KD | 85.5 | 83.8 | 64.0 | 84.2 | 82.0 | 79.9 |
| PatchCore† | 99.1 | 96.0 | 92.1 | 97.2 | 93.9 | 95.7 |
| SimpleNet† | 99.4 | 90.6 | 71.7 | 91.6 | 76.1 | 85.9 |
| PNI† | 99.6 | 87.8 | 67.6 | 90.2 | 66.1 | 82.3 |
| RealNet† | **99.7** | 92.3 | 95.4 | 95.6 | 76.7 | 91.9 |
| RD | 98.6 | 96.5 | 94.1 | **98.9** | 90.1 | 95.7 |
| RD++† | 98.7 | 96.1 | 95.2 | 98.2 | 84.4 | 94.5 |
| GNL | 98.0 | 97.4 | 97.5 | 97.8 | 94.1 | 97.0 |
| GNL† | 97.7 | 97.2 | 96.5 | 97.0 | 93.7 | 96.4 |
| **FiCo (ours)** | 98.8 | **97.9** | **97.9** | 98.5 | **95.0** | **97.6** |

Table 1: Comparison of state-of-the-art methods on sample-level AUROC for MVTec. "Ori,Br,Co,Bl,No" represents original, brightness, contrast, defocus blur and gaussian noise scenario.

| Method | ID | OOD | | | Avg. |
|---|---|---|---|---|---|
| | P | A | C | S | |
| Deep SVDD | 40.9 | 53.4 | 41.2 | 39.5 | 43.8 |
| f-AnoGAN | 61.3 | 50.2 | 52.4 | **63.8** | 56.9 |
| KD | 88.2 | 62.9 | 62.6 | 51.4 | 66.3 |
| PatchCore† | 77.5 | 57.5 | 56.5 | 52.1 | 60.9 |
| SimpleNet† | **91.6** | 62.3 | 54.8 | 47.5 | 64.1 |
| RD | 81.5 | 61.1 | 60.3 | 55.1 | 64.5 |
| RD++† | 86.9 | 61.7 | 65.2 | 60.6 | 68.6 |
| GNL | 87.7 | 65.6 | 68.0 | 62.4 | 70.9 |
| GNL† | 87.5 | 64.8 | 68.3 | 58.1 | 69.7 |
| **FiCo (ours)** | 89.7 | **67.6** | **70.9** | 62.3 | **72.6** |

Table 2: Comparison of state-of-the-art methods on sample-level AUROC for PACS. "P,A,C,S" represents photo, art painting, cartoon and sketch domain.

one class is regarded as the normal class and all other classes are abnormal counterparts. Our method FiCo also surpasses all methods on average AUROC. Especially compared with RD++ and GNL, FiCo shows superiority on both ID and OOD scenarios.

## Ablation Studies

**Performance of different components.** As our method FiCo consists of different additional modules and losses, we conduct ablation study on each part step-by-step to investigate the effectiveness. As shown in Table 4, we start from the re-implementation of GNL (Cao, Zhu, and Pang 2023) as the baseline method, and sequentially add DiSCo module with $\mathcal{L}_{Co}$, DiIFi module with $\mathcal{L}_{mse}$, and $\mathcal{L}_{nor}$. Note that except for $\mathcal{L}_{nor}$, other losses are omitted within the module design for abbreviation in Table 4. Results show that each part can positively improve performance with 0.45%, 1.59% and 0.90% on average. Specifically, after inserting the DiIFi module upon the DiSCo module, performance on all scenarios improves significantly for better invariant representation. Moreover, $\mathcal{L}_{nor}$ is capable to improve performance by preventing the normality collapsing.

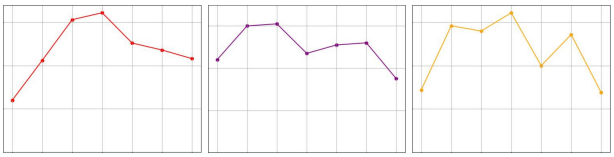
**Hyper-parameter Sensitivity.** The three hyper-parameters $\alpha, \beta, \gamma$ are designed to balance the ratio between different additional losses. Ablation study is conducted to analyze the sensitivity to demonstrate the practicality of our method. Fig. 3 shows the results of each hyper-parameter with the others fixed to the optimal value. The fluctuations of three hyper-parameters are low with merely 1.01%, 0.26% and 0.92%, and any combination of hyper-parameter values can surpass GNL (Cao, Zhu, and Pang 2023). The best performance is achieved when $\alpha = 0.05$, $\beta = 0.02$ and $\gamma = 1.0$.

| Method | ID | OOD | | | | Avg. |
|---|---|---|---|---|---|---|
| | Ori | Br | Co | Bl | No | |
| Deep SVDD | 64.6 | 59.1 | 55.9 | 62.1 | 54.5 | 59.3 |
| f-AnoGAN | 70.3 | 54.6 | 57.2 | 60.7 | 51.8 | 58.9 |
| KD | 84.2 | 75.9 | 64.4 | 63.5 | 56.9 | 69.0 |
| PatchCore† | 80.6 | 72.9 | 63.0 | 57.7 | 55.5 | 65.9 |
| RD | **84.6** | 75.9 | 65.3 | **66.7** | 58.8 | 70.3 |
| RD++† | 80.3 | 75.9 | 66.9 | 60.3 | 63.3 | 69.3 |
| GNL | 82.3 | 77.9 | 66.1 | 64.0 | 61.5 | 70.4 |
| GNL† | 79.2 | 76.9 | 67.5 | 63.2 | **64.6** | 70.3 |
| **FiCo (ours)** | 80.5 | **77.8** | **69.2** | 63.8 | 64.4 | **71.1** |

Table 3: Comparison of state-of-the-art methods on sample-level AUROC for CIFAR-10. "Ori,Br,Co,Bl,No" represents original, brightness, contrast, defocus blur and gaussian noise scenario.

| Method | ID | OOD | | | Avg. |
|---|---|---|---|---|---|
| | P | A | C | S | |
| GNL | 87.5 | 64.8 | 68.3 | 58.1 | 69.7 |
| DiSCo | 88.2 | 64.2 | 69.0 | 59.2 | 70.1 |
| DiSCo + DiIFi | 89.5 | 65.5 | 70.5 | 61.6 | 71.7 |
| **FiCo** | **89.7** | **67.6** | **70.9** | **62.3** | **72.6** |

Table 4: Effectiveness of different components on the PACS benchmark.



(a) $\alpha$ (b) $\beta$ (c) $\gamma$

Figure 3: Experimental results on hyper-parameters on the MVTec benchmark.

## Analysis

**Filter and Compensation Process.** Fig. 4 shows the anomaly map from the distribution-invariant representation $f^{D_k}$ for the filter process, and the final output $f_F^{D_k}$ for the compensation process. It can be clearly observed that the anomaly map from $f^{D_k}$ filters most of the abnormal information, including distribution-specific information and anomalous patterns. Thus, the discrepancy between $f^{D_k}$ and $f^{E_k}$ is maximized so that there exists more activated regions on both the real anomalous regions and distribution-specific regions, which validates the function of the DiIFi

module to filter abnormal information. However, to prevent distribution-specific information to be recognized as anomalous patterns, the compensation process should restitute the distribution-specific noise, i.e., style information, for alignment with $f^{E_k}$. Compared with $f^{D_k}$, $f_F^{D_k}$ merely concentrates on the anomalous regions and the activated distribution-specific regions are mostly eliminated. Furthermore, our method not only resolves the distribution shift issue on OOD scenarios, but also shows strong robustness on ID data during the filter and compensation process.

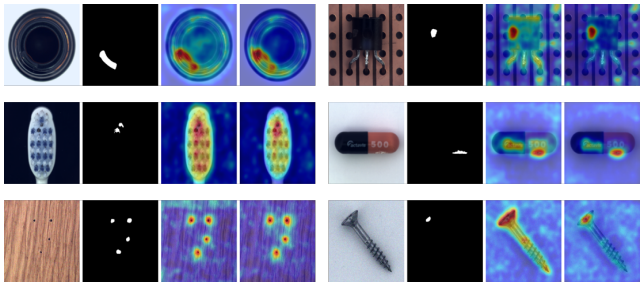


Figure 4: Anomaly map of $f^{D_k}$ and $f_F^{D_k}$ on the MVTec benchmark. Each row represents a different scenario, including ID, defocus blur and gaussian noise. For each scenario, two examples are shown with the original image, the groundtruth label, anomaly map from $f^{D_k}$ and from $f_F^{D_k}$.

**Key Difference for Anomaly Score.** Anomaly detection aims to enlarge the discrepancy on the distribution of anomaly scores between normal samples and anomalies. Thus, we visualize the distribution of sample-level anomaly scores on all images on the hardest scenario gaussian noise and the simplest scenario ID. Fig. 5 displays the comparison between FiCo and GNL (Cao, Zhu, and Pang 2023) to illustrate our merits. It can be observed that GNL fails on several cases that the overlap between the normal samples and anomalies is large. Instead, FiCo not only increases the discrepancy between normal samples and anomalies on OOD scenarios, but also shows strong adaptability on ID data. Furthermore, the anomaly scores for FiCo on normal samples are smaller than GNL that proves the effectiveness.

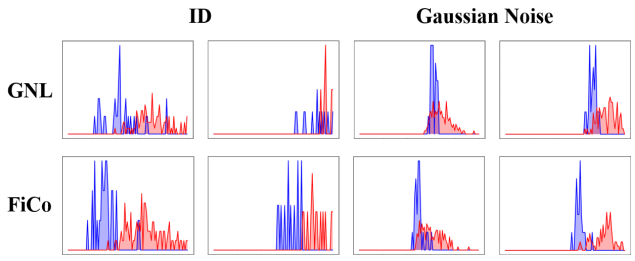



Figure 5: Anomaly scores of FiCo and GNL (Cao, Zhu, and Pang 2023) on ID, gaussian noise scenario from the MVTec benchmark. Each scenario consists of two examples from 'zipper', 'toothbrush', and 'pill', 'metal nut'. Color blue and red indicates distribution on normal samples and anomalous samples, respectively.

**Time Consumption.** We evaluate the total training and inference time on SOTA RD-based methods. As presented in Table 5, the total training time is approximately three times shorter than RD++ (Tien et al. 2023) for a single epoch, let alone RD++ uses 280 epochs while FiCo merely consumes 20 epochs. Meanwhile, our method consumes longer training time compared with GNL (Cao, Zhu, and Pang 2023) due to the additional modules but remains reasonable. For the inference time, our method FiCo takes slightly longer time than RD++ and GNL, but yields substantial improvement on all ID and OOD scenarios to achieve SOTA performance.

| Method | $T_{train}$ (s/epoch) | $T_{test}$ (ms/image) | Results (%) |
|---|---|---|---|
| RD++ | 59.6 | 31.3 | 95.8 |
| GNL | 13.3 | 40.8 | 95.2 |
| **FiCo (ours)** | 18.1 | 46.2 | 97.5 |

Table 5: Comparison of time consumption for RD-based methods on category 'toothbrush' in the MVTec benchmark on TESLA T4 GPU. $T_{train}$ and $T_{test}$ are training time per epoch and test time per image.

## Conclusion

This paper proposes FiCo to tackle the misalignment issue from the perspective of invariant representation for anomaly detection under distribution shifts. With the filter and compensation process, the model is capable to learn distribution-invariant normality and identify real anomalous patterns rather than distribution-specific information. The proposed DiSCo and DiIFi modules with novel training objectives are exclusively designed to address the misalignment issue, and can also be trained end-to-end without much extra costs. Experiments on several benchmarks for both industrial anomaly detection and one-class novelty detection in AD, show the strong generalizability and robustness of our method. Moreover, the visualizations are valid for the explanation of the whole process. In the future, the exploration of large vision-language models and generative models to resolve the distribution shift in anomaly detection worth more exploration.

## Acknowledgments

This work is supported by Chinese National Natural Science Foundation under Grants (62076033).

## References

Bae, J.; Lee, J.-H.; and Kim, S. 2023. Pni: industrial anomaly detection using position and neighborhood information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6373–6383.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9592–9600.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2020. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4183–4192.

Cao, T.; Zhu, J.; and Pang, G. 2023. Anomaly detection under distribution shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6511–6523.

Cha, J.; Chun, S.; Lee, K.; Cho, H.-C.; Park, S.; Lee, Y.; and Park, S. 2021. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34: 22405–22418.

Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; and Liu, Z. 2020. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11030–11039.

Chen, Z.; Wang, W.; Zhao, Z.; Su, F.; Men, A.; and Dong, Y. 2023. Cluster-instance normalization: A statistical relation-aware normalization for generalizable person re-identification. *IEEE Transactions on Multimedia*.

Chen, Z.; Wang, W.; Zhao, Z.; Su, F.; Men, A.; and Meng, H. 2024. PracticalDG: Perturbation Distillation on Vision-Language Models for Hybrid Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23501–23511.

Choi, S.; Jung, S.; Yun, H.; Kim, J. T.; Kim, S.; and Choo, J. 2021. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11580–11590.

Cohen, N.; and Hoshen, Y. 2020. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*.

Deecke, L.; Ruff, L.; Vandermeulen, R. A.; and Bilen, H. 2021. Transfer-based semantic anomaly detection. In *International Conference on Machine Learning*, 2546–2558. PMLR.

Defard, T.; Setkov, A.; Loesch, A.; and Audigier, R. 2021. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, 475–489. Springer.

Deng, H.; and Li, X. 2022. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9737–9746.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2020. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gu, Z.; Liu, L.; Chen, X.; Yi, R.; Zhang, J.; Wang, Y.; Wang, C.; Shu, A.; Jiang, G.; and Ma, L. 2023. Remembering Normality: Memory-guided Knowledge Distillation for Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16401–16409.

Hou, J.; Zhang, Y.; Zhong, Q.; Xie, D.; Pu, S.; and Zhou, H. 2021. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8791–8800.

Huang, C.; Guan, H.; Jiang, A.; Zhang, Y.; Spratling, M.; and Wang, Y.-F. 2022. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, 303–319. Springer.

Huang, W.; Chen, C.; Li, Y.; Li, J.; Li, C.; Song, F.; Yan, Y.; and Xiong, Z. 2023. Style Projected Clustering for Domain Generalized Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3061–3071.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lee, M.; and Choi, J. 2024. Text-Guided Variational Image Generation for Industrial Anomaly Detection and Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26519–26528.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9664–9674.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. 2018a. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, 5542–5550.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018b. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5400–5409.

Liao, S.; and Shao, L. 2020. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 456–474. Springer.

Liao, S.; and Shao, L. 2022. Graph sampling based deep metric learning for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7359–7368.

Liu, W.; Luo, W.; Lian, D.; and Gao, S. 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6536–6545.

Liu, Z.; Zhou, Y.; Xu, Y.; and Wang, Z. 2023. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20402–20411.

Lv, F.; Liang, J.; Li, S.; Zang, B.; Liu, C. H.; Wang, Z.; and Liu, D. 2022. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8046–8056.

Mahajan, D.; Tople, S.; and Sharma, A. 2021. Domain generalization using causal matching. In *International Conference on Machine Learning*, 7313–7324. PMLR.

Park, H.; Noh, J.; and Ham, B. 2020. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14372–14381.

Reiss, T.; Cohen, N.; Bergman, L.; and Hoshen, Y. 2021. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2806–2814.

Ristea, N.-C.; Madan, N.; Ionescu, R. T.; Nasrollahi, K.; Khan, F. S.; Moeslund, T. B.; and Shah, M. 2022. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13576–13586.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *International conference on machine learning*, 4393–4402. PMLR.

Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M. H.; and Rabiee, H. R. 2021. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14902–14912.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Langs, G.; and Schmidt-Erfurth, U. 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54: 30–44.

Schlüter, H. M.; Tan, J.; Hou, B.; and Kainz, B. 2022. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, 474–489. Springer.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, 443–450. Springer.

Tien, T. D.; Nguyen, A. T.; Tran, N. H.; Huy, T. D.; Duong, S.; Nguyen, C. D. T.; and Truong, S. Q. 2023. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24511–24520.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Wyatt, J.; Leach, A.; Schmon, S. M.; and Willcocks, C. G. 2022. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 650–656.

Xu, B.; Wang, N.; Chen, T.; and Li, M. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14383–14392.

Yao, X.; Li, R.; Qian, Z.; Luo, Y.; and Zhang, C. 2023. Focus the discrepancy: Intra-and inter-correlation learning for image anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6803–6813.

Yu, J.; Zheng, Y.; Wang, X.; Li, W.; Wu, Y.; Zhao, R.; and Wu, L. 2021. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2022. Dsr–a dual subspace re-projection network for surface anomaly detection. In *European conference on computer vision*, 539–554. Springer.

Zhang, X.; Li, N.; Li, J.; Dai, T.; Jiang, Y.; and Xia, S.-T. 2023. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6782–6791.

Zhang, X.; Xu, M.; and Zhou, X. 2024. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16699–16708.

Zhang, Y.; Li, M.; Li, R.; Jia, K.; and Zhang, L. 2022. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8035–8045.

Zhao, Y.; Zhong, Z.; Zhao, N.; Sebe, N.; and Lee, G. H. 2022. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *European Conference on Computer Vision*, 535–552. Springer.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*.

Zhu, J.; and Pang, G. 2024. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17826–17836.