

RESEARCH ARTICLE

Factorized Industrial Anomaly Detection and Localization

YUHAO ZHU¹, LINLIN DAI², XINGZHI DONG², AND PING LI²¹Postgraduate Department, China Academy of Railway Sciences, Beijing 100081, China²China Academy of Railway Sciences Corporation Ltd., Beijing 100081, China

Corresponding author: Ping Li (liping@rails.cn)

This work was supported by the Scientific Research Project of Beijing Jingwei Information Technologies Company Ltd., under Grant DZYF22-06.

ABSTRACT Convolutional neural networks trained on large datasets can generalize various down-streaming tasks, including industrial anomaly detection and localization, which is critical in modern large-scale industrial manufacturing. Whereas previous methods have demonstrated that the feature fusion strategy across multiple layers is effective for better performance on industrial anomaly detection and localization, they lack flexibility in intervening and manipulating the local and global information composition process. Through experiments, we demonstrate that the brute-force feature fusion strategy used in previous methods leads to sub-optimal performance in most industrial anomaly detection scenarios. To this end, we propose a novel feature factorization and reversion framework based on invertible neural networks, enabling the selective emphasis or suppression of distinct information in a continuous space by request to fit various preferences for detecting different abnormalities. The preferred local and global info-combination for detecting different defects on 15 objects is studied by experiments exhaustively on the popular benchmark MVTec-AD. Based on feature factorization and reversion, our method is able to outperform previous state-of-the-art methods by a noticeable margin, achieving an image-level anomaly detection AUROC score of up to 99.67% (previously 99.4%), pixel-level anomaly localization AUROC score of 98.61% (previously 98.5%), and AUPRO of 95.15% (previously 94.6%), which validates the effectiveness of the proposed method for industrial anomaly detection and localization.


INDEX TERMS Industrial anomaly detection and localization, invertible neural network, feature factorization and reversion.

I. INTRODUCTION

For many years, industrial anomaly detection and localization aims at pursuing the human ability to differentiate between expected variance in the data and outliers given only a small number of normal samples. As the types of abnormality in industrial images could be unexpected and the defect variations are costly to specify in full, it is better to fit an anomaly detection and localization model using only non-defective examples. Previous works have explored Auto-Encoder [1], Variational Auto-Encoder [2], or Generative Adversarial Net [3] to restore the input samples to normal samples for comparison based on certain distance-based metrics, where the farther distance indicates a higher possibility

for abnormalities to happen. However, these methods can be limited by the ability of powerful neural networks to learn an “identical shortcut” [4], where both normal and anomalous samples can be well reconstructed with similar errors, and hence fail to spot abnormalities.

Recently, representation-based methods [5], [6] have been applied to industrial anomaly detection and localization, leveraging deep representations obtained from ImageNet classification models [7], [8]. Inspired by the multi-level feature fusion strategy used in other computer vision tasks [9], [10], [11], the feature fusion operation that integrates local information encoded in low-level features and the global information in high-level features could improve the model performance for anomaly detection and localization [12], [13], [14]. The success of these methods relies on the heuristic insight that abnormalities can be spotted if unusual local

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson .

semantic patterns are present or the global relations are topologically anomalous, *i.e.*, desired patterns are not shown in the area as they ought to locate.

Whereas multi-level feature fusion strategies have shown promising results for improving the accuracy of industrial anomaly detection and localization, the proportion of local against global information encoded in multi-level features is fixed after training under the supervision of proxy tasks, which would be sub-optimal for unsupervised industrial anomaly detection and localization using the feature fused across multiple feature pyramids: the inherited local and global info-combination determined by pre-trained models is NOT ideal for industrial anomaly detection and localization. Unfortunately, previous works primarily focused on training objectives of proxy tasks or the distance metric for feature matching based on the conventional feature fusion operations such as concatenation and neighborhood pooling, failing to study the information composition process for better performance in industrial anomaly detection and localization.

In this paper, we propose a novel method for industrial anomaly detection and localization, namely Feature Factorization and Reversion (F^2R), based on invertible neural networks [15]. Unlike previous methods that rely on the inherited info-combination determined by pre-trained models, F^2R intervenes in the feature fusion process through three steps: (1) factorizing the pre-trained multi-level features into separated local semantic and global positional representations, (2) emphasizing or suppressing distinct information by scaling the representations individually, and (3) reversing the scaled representations back to the original feature space with the preferred proportion of local versus global information. Notably, F^2R can be trained under an unsupervised setting and can serve as a plug-in module with arbitrary deep feature extractors. The experimental results on the popular benchmark MVTec-AD [16] demonstrate the superiority of the proposed F^2R over previous state-of-the-art methods, achieving an image-level anomaly detection AUROC score of up to 99.67%, which almost halving the error compared to the previous best score of 99.4%. Additionally, F^2R achieves a pixel-level anomaly localization AUROC score of 98.61% and AUPRO score of 95.15%, improved from the previous best score of 98.5% and 94.6%, respectively. Our main contributions are as follows:

- The conventional feature fusion strategy used in industrial anomaly detection and localization is verified to be sub-optimal, where the inherited local and global info-combination determined by pre-trained models needs to be delicately controlled to accommodate various preferences for detecting different defects.
- To this end, a novel method namely Feature Factorization and Reversion (F^2R) is proposed to intervene in the feature fusion process based on invertible neural networks, where the local and global information can be emphasized or suppressed in a continuous space. With the customized feature, abnormalities can be detected and localized using existing distance-based metrics.

As the result, F^2R can be plugged into any deep feature extractors to boost the performance of industrial anomaly detection and localization.

- The proposed method achieves a new state-of-the-art performance in industrial anomaly detection and localization, validating our insights on the optimality of feature fusion strategy and showing the effectiveness of the proposed method.

In the following of this paper, related works are summarized in section II. The proposed method is presented in section III, including the overall pipeline and the detailed components. Experiments are conducted in section IV with ablation studies and discussions. Finally, the conclusion is drawn in Section V.

II. RELATED WORKS

A. INDUSTRIAL ANOMALY DETECTION

Usually, industrial anomaly detection methods can be categorized into reconstruction-based and representation-based methods. Reconstruction-based methods assume that the anomalous regions cannot be reconstructed if the model is trained only on normal samples, where the reconstruction model could be Auto-Encoders [17], [18], [19], [20], Variational Auto-Encoders [21], [22], [23] or Generative Adversarial Nets [24], [25], [26], [27], [28]. However, reconstruction-based methods are strong enough to reconstruct anomalies, and hence fail to spot these defects. Accordingly, memory mechanism [29], [30] and iteration mechanism [31] are introduced to relieve this issue. Other works propose to generate pseudo-anomaly samples [32], [33], [34] for training as the proxy task. However, these methods partly rely on how well pseudo-anomalies match the real anomalies that are not known [35].

On the other hand, representation-based methods argue that large-scale natural image datasets such as ImageNet [36] can be more representative for pretraining compared to small application-specific datasets. Based on the multi-level feature extracted from ImageNet classification models, representation-based methods are able to fit a normal distribution using only features from normal instances, where the anomaly score could be formalized as Euclidean distance [12] or Mahalanobis distance [13] between the test and gallery samples. Our method is most related to the unsupervised representation learning methods via normalizing flow [37], a typical invertible neural network. Both DifferNet [38] and FastFlow [39] transform the feature into a tractable multidimensional Gaussian distribution and assign the likelihood to recognize anomalies, whereas our method differs from them in representation factorization to encode local and global information individually, which allows for flexible information integration. While previous methods have discussed the combination of local and global information in time series classification [40], [41], [42], our method is specifically designed for industrial anomaly detection and localization, where CFLOW-AD [43] is a related method that

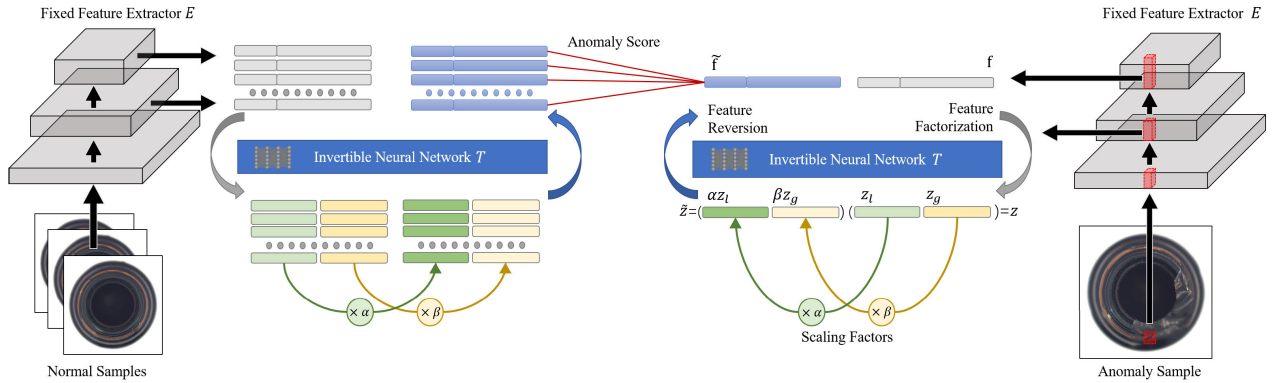


FIGURE 1. The overall pipeline of the proposed method consists of two neural networks, the pre-trained encoder E for feature extraction and an invertible neural network T for information intervention. Specifically, the multi-level feature \mathbf{f} is formalized by fusing feature maps from two layers of E for both normal and anomaly samples, and there are three steps to intervene in the information integration: (1) the flattened multi-level feature \mathbf{f} is projected to the latent representation \mathbf{z} using T , where \mathbf{z} follows a spherical multivariate Gaussian distribution $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}; \mathbf{0}, I)$; (2) the latent representation \mathbf{z} is factorized into two statistically independent variables \mathbf{z}_l and \mathbf{z}_g that encode local and global information respectively, and then both \mathbf{z}_l and \mathbf{z}_g are scaled by factors α and β to emphasize or suppress distinct information; and (3) the intervened latent representation $\tilde{\mathbf{z}} = [\alpha\mathbf{z}_l, \beta\mathbf{z}_g]$ is reversed back to $\tilde{\mathbf{f}} = T^{-1}(\tilde{\mathbf{z}})$ in the original feature space for feature matching, where the minimum Euclidean distance serves as the anomaly score for detection and localization.

explicitly adds positional information into transformed representations, assuming that the feature vectors lack a global view to perceive their spatial location. On the contrary, our method only intervenes in the inherent information already encoded in the pre-trained multi-level features and achieves better performance, showing the existence of both local and global information in multi-level features.

B. FEATURE FACTORIZATION

Principal component analysis [44] is a well-known technique for feature factorization, being used for reducing the dimensionality and enabling the visualization of multidimensional data. However, the principal component may not be interpretable in terms of human semantics. Therefore, deep feature factorization [45] is proposed to discover human semantic concepts in deep features, the most works of which focus on the context of image generation, where the semantic factors are manipulated to generate novel images with desired attributes [46], [47], [48], [49].

Besides the image generation, feature factorization has also been studied in domain generalization that fits a model to perform well on unseen target domains given multiple observed source domains during the training [50], [51], [52], [53]. To this end, deep features are decomposed into domain-invariant features and domain-specific features. During the testing, only domain-invariant features are retained when applied to target domains. Recently, feature factorization is adapted for image retrieval [54] and network factorization [55]. To the best of our knowledge, the proposed method is the first work that adopts feature factorization in industrial anomaly detection and localization and discusses the information preferences for spotting various defects.

III. METHODOLOGY

In this section, we first introduce the pipeline of our method, as shown in Fig.1, followed by the details of proposed

objectives for training the invertible neural network, including the losses for invertible transformation and feature factorization.

A. OVERVIEW OF THE PROPOSED METHOD

Given a pre-trained feature extractor E , the multi-level feature \mathbf{f} can be formalized by fusing multiple feature maps at different layers. In order to intervene in the connatural local and global information encoded in \mathbf{f} to fit the preferences for detecting different defects, the very naive way is to factorize \mathbf{f} into local semantic feature \mathbf{f}_l and global position feature \mathbf{f}_g in the original feature space for manipulation. However, it could be challenging to conduct such feature factorization as the true distribution $\mathbf{f} \sim p_\theta(\mathbf{f})$ with parameters θ is unknown. On the other hand, factorizing features that follow a tractable density, e.g. a spherical multivariate Gaussian distribution $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}; \mathbf{0}, I)$, is more viable. To this end, an invertible transformation T that satisfies $\mathbf{z} = T(\mathbf{f})$ and $\mathbf{f} = T^{-1}(\mathbf{z})$ is learnt, where the latent representation \mathbf{z} is supervised to encode local semantic representation \mathbf{z}_l and global position representation \mathbf{z}_g channel-wisely, i.e., $\mathbf{z} = [\mathbf{z}_l, \mathbf{z}_g]$, where $[\cdot, \cdot]$ denotes the concatenated tensor. Then, the local and global information can be intervened by $\tilde{\mathbf{z}} = [\alpha\mathbf{z}_l, \beta\mathbf{z}_g]$, where α and β are scaling coefficients as two hyper-parameters. When $\alpha > 1$ ($\beta > 1$), the local (global) information is emphasized, otherwise the corresponding information is suppressed. Especially, when $\alpha = \beta = 1$, the reversed feature $\tilde{\mathbf{f}} = T^{-1}(\tilde{\mathbf{z}})$ should keep intact and the information is not intervened to affect the original performance for anomaly detection and localization.

B. INVERTIBLE TRANSFORMATION OF FEATURES

Optionally, the invertible neural network T can be implemented by flow-based generative models [56], [57], [58]. Given a multi-level feature $\mathbf{f} \in \mathcal{R}^{h \times w \times c}$, where h , w , and c denote the height, width and the hidden dimension, the

multi-level feature can be flattened into the shape of $\mathcal{R}^{hw \times c}$ and $\mathbf{z} = T(\mathbf{f})$ is as the same shape of \mathbf{f} . In this way, each input image can generate hw high-dimensional vectors following unknown true distribution $\mathbf{f} \sim p_{\theta}(\mathbf{f})$ with parameters θ and the log-likelihood of \mathbf{f} can be written as:

$$\log p_{\theta}(\mathbf{f}) = \log p_{\theta}(\mathbf{z}) + \log |\det(d\mathbf{z}/d\mathbf{f})| \quad (1)$$

where the *log-determinant* $\log |\det(d\mathbf{z}/d\mathbf{f})|$ is the logarithm of the absolute value of the determinant of the Jacobian matrix $(d\mathbf{z}/d\mathbf{f})$ of the transformation function T [37]. Besides, when the latent representation \mathbf{z} follows spherical multivariate Gaussian distribution $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}; 0, I)$, minimizing the negative log-likelihood of \mathbf{f} is equivalent to minimize Gaussian negative log-likelihood of the latent variable \mathbf{z} and the negative log-determinant of the transformation function T . As the result, the loss function to minimize the negative log-likelihood can be written as:

$$\begin{aligned} \mathcal{L}_{nll} &= ||T(\mathbf{f})||^2 - \log |d\mathbf{z}/d\mathbf{f}| \\ &= ||\mathbf{z}||^2 - \log |d\mathbf{z}/d\mathbf{f}| \end{aligned} \quad (2)$$

where the log-determinant can be efficiently computed when the Jacobian of the transformation function is a triangular matrix [58]. Based on the invertible neural network T , the representation \mathbf{f} can be mapped to a latent representation \mathbf{z} and vice versa, which works as the basic tool for predicting factorized features and re-integrate the disentangled information during the feature reversion.

C. FEATURE FACTORIZATION

In order to factorize \mathbf{z} into \mathbf{z}_l and \mathbf{z}_g that encode local semantic information and global position information respectively, the mutual information between \mathbf{z}_l and \mathbf{z}_g , denoted as $\mathcal{I}(Z_l, Z_g)$, should be minimized. In addition, the local representation \mathbf{z}_l should mirror the semantic patterns in the original feature embedding \mathbf{f} , which requires a maximization of the mutual information $\mathcal{I}(Z_l, F)$. Similarly, the global representation \mathbf{z}_g should be able to preserve the absolute position information P . Thus, the mutual information $\mathcal{I}(Z_g, P)$ should be maximized as well. In summary, we propose the loss function for feature factorization as:

$$\mathcal{L}_f = \mathcal{I}(Z_l, Z_g) - \mathcal{I}(Z_l, F) - \mathcal{I}(Z_g, P) \quad (3)$$

Due to the difficulty to estimate mutual information for continuous variables, the variational lower and upper bounds [55] are adopted as the proxy for optimization:

$$\begin{aligned} \mathcal{I}(X, Y) &\geq \mathbb{E}_{p(\mathbf{y}, \mathbf{x})} [\log q(\mathbf{y}|\mathbf{x})] \\ \mathcal{I}(X, Y) &\leq \mathbb{E}_{p(\mathbf{y})} [D_{KL}[p(\mathbf{y}|\mathbf{x})||q(\mathbf{y})]] \end{aligned} \quad (4)$$

where $q(\mathbf{y}|\mathbf{x})$ and $q(\mathbf{y})$ are the approximations of $p(\mathbf{y}|\mathbf{x})$ and the marginal distribution of $p(\mathbf{y})$, both of which can be estimated by Multi-Layer Perceptrons (MLP) [59]. Based on (4),

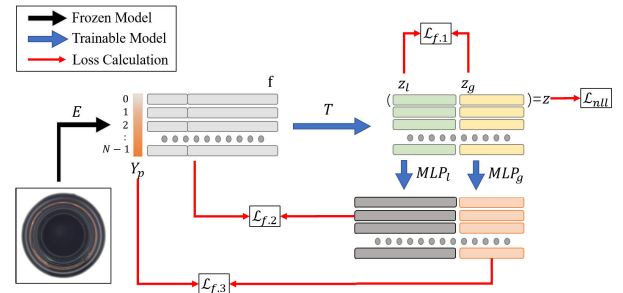


FIGURE 2. Objectives for training the invertible neural network T assisted by two auxiliary MLPs, including (1) \mathcal{L}_{nll} that trains T to predict \mathbf{z} following the distribution of Gaussian, (2) the first term of \mathcal{L}_f ($\mathcal{L}_{f,1}$) that factorizes \mathbf{z}_l and \mathbf{z}_g , (3) the second term of \mathcal{L}_f ($\mathcal{L}_{f,2}$) that trains \mathbf{z}_l to encode local semantic information by approximating the original features, and (4) the third term of \mathcal{L}_f ($\mathcal{L}_{f,3}$) that trains \mathbf{z}_g to encode global positional information by predicting the positional label $Y_p \in [0, N - 1]$, i.e., the index of its corresponding feature. Both \mathcal{L}_{nll} and \mathcal{L}_f can be calculated in an unsupervised manner.

\mathcal{L}_f can be upper bounded as:

$$\begin{aligned} \mathcal{L}_f &= \mathcal{I}(Z_l, Z_g) - \mathcal{I}(Z_l, F) - \mathcal{I}(Z_g, P) \\ &\leq \mathbb{E}_{p(\mathbf{z}_l)} [D_{KL}[p(\mathbf{z}_l|\mathbf{z}_g)||q(\mathbf{z}_l)]] \\ &\quad - \mathbb{E}_{p(\mathbf{f}, \mathbf{z}_l)} [\log q(\mathbf{f}|\mathbf{z}_l)] \\ &\quad - \mathbb{E}_{p(\mathbf{p}, \mathbf{z}_g)} [\log q(\mathbf{p}|\mathbf{z}_g)] \end{aligned} \quad (5)$$

where these three terms can be further simplified as the latent variable \mathbf{z} is supervised by \mathcal{L}_{nll} to follow spherical multivariate Gaussian distribution $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}; 0, I)$.

Term 1: the KL divergence between $p(\mathbf{z}_l|\mathbf{z}_g)$ and $q(\mathbf{z}_l)$ can be minimized when $p(\mathbf{z}_l|\mathbf{z}_g) = q(\mathbf{z}_l) \sim \mathcal{N}(\mathbf{z}; 0, I)$, i.e., \mathbf{z}_l is independent to \mathbf{z}_g . For two variables following zero-mean unit-variance Gaussian distribution, this term can be simplified as optimizing the covariance between \mathbf{z}_l and \mathbf{z}_g to zero, which could be represented as $|\Sigma(\mathbf{z}_l)(\mathbf{z}_g)^T| \rightarrow \mathbf{0}$. During the training, the covariance is approximately estimated batch-wisely.

Term 2: the negative log-likelihood of \mathbf{f} with respect to \mathbf{z}_l . The maximization of $\mathbb{E}_{p(\mathbf{f}, \mathbf{z}_l)} [\log q(\mathbf{f}|\mathbf{z}_l)]$ can be achieved when $||\mathbf{f} - MLP_l(\mathbf{z}_l)||^2$ is minimized for regression task or $\cos(\mathbf{f}, MLP_l(\mathbf{z}_l))$ is maximized for binary classification task, where MLP_l denotes a MLP model with \mathbf{z}_l as input and \cos denotes the cosine similarity between the two vectors.

Term 3: the negative log-likelihood of position information P with respect to Z_g . Similar to term 2, MLP_g is adopted to predict the position label $Y_p \in [0, 1, 2, \dots, N - 1]$ with input \mathbf{z}_g as the proxy task, where N equals to hw , the spatial size of the feature map. By default, the cross-entropy is adopted as the supervision loss to optimize this term.

To sum up, the invertible neural network T can be optimized through the combination of \mathcal{L}_{nll} and \mathcal{L}_f in an unsupervised manner:

$$\mathcal{L} = \gamma \mathcal{L}_{nll} + \mathcal{L}_f \quad (6)$$

where $\gamma = 0.001$ is the default weight coefficient in all experiments if not specially specified.

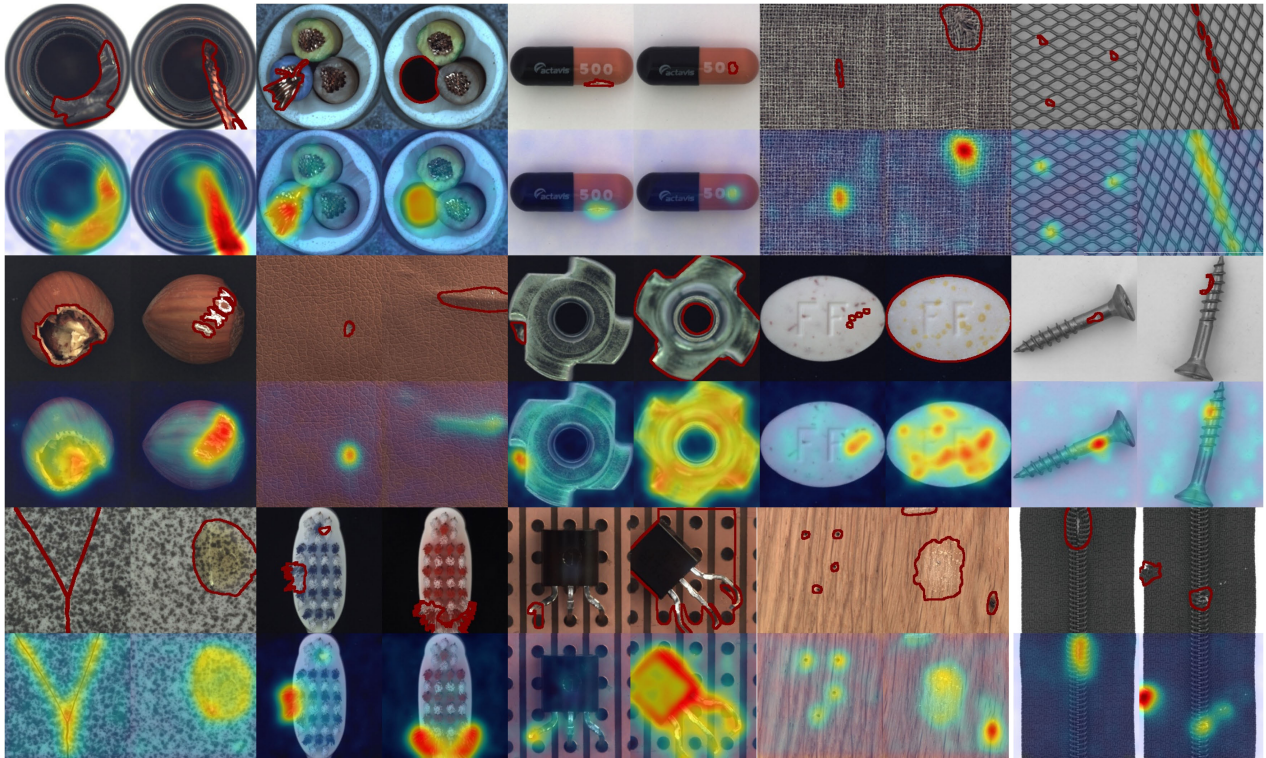


FIGURE 3. Examples of anomaly localization results on MVTec-AD, where odd rows display the input images with ground truth labeled in red edges, and even rows display the corresponding anomaly localization heatmaps.

Please Note that the proposed objective depends only on the outputs of the feature extractor and the invertible neural network, except the position label Y_p that can be inferred by the indexes of the feature vectors on the corresponding multi-level feature map. Consequently, the model can be trained in an unsupervised manner as shown in Fig.2.

IV. EXPERIMENTAL RESULTS

A. IMPLEMENT DETAILS

1) GENERAL SETTINGS

In all experiments, a pre-trained WideResNet-50 [8] model is used as the feature extractor. As shown in previous representation-based methods [12], [13], [14], [60], the feature maps from stages 2 and 3 are verified to be effective for anomaly detection and localization. Therefore, the multi-level feature is also formed in our experiments for a fair comparison. In addition, the invertible neural network T is implemented following [58] with 6 flow steps, while both MLP_l and MLP_g have 4 linear layers with BatchNorm [61] and LeakyReLU [62] layers in between.

For training, AdamW optimizer [63] is used with a learning rate of $2e-4$ and the coefficient betas are (0.9, 0.999). The batch size is 32 and the weight decay is set to 0.05. Besides, a simple image augmentation of random rotation in 30 degrees is adopted. The training process lasts 300 epochs with a cosine learning rate annealing strategy. After the training, two auxiliary MLPs are discarded.

For testing, the reversed features $\tilde{\mathbf{f}}_{normal}$ of normal samples are collected as the gallery. In order to reduce the gallery set

for fast feature matching, a subset of the whole gallery is selected following [14]. Specifically, multiple feature vectors that locate closely are considered redundant, therefore only a few representative vectors are selected as the centroids while others are discarded. Consequently, the minimum Euclidean distances between $\tilde{\mathbf{f}}_{test}$ and $\tilde{\mathbf{f}}_{normal}$ on each position are used as the anomaly heatmap for anomaly localization and the maximum value on the heatmap serves as the anomaly score for image-level anomaly detection.

2) DATASET

To study the information preferences for industrial anomaly detection and localization, the MVTec Anomaly Detection benchmark [16] is used, which contains 15 sub-datasets with 3629 images for training and 1725 images for testing. Each sub-dataset is divided into a nominal-only training set and a testing set containing both normal and anomalous samples with various defects as well as respective ground truth anomaly masks. All training and test images are resized to 256×256 and center-cropped to 224×224 for feature extraction, factorization, reversion, and anomaly score calculation.

3) EVALUATION METRICS

As the metrics of True Positive Rate and False Positive Rate are threshold-specific and the optimal threshold may vary to satisfy different scenarios, the commonly used threshold-agnostic metric, the area under the receiver operator curve (AUROC), is reported for the evaluation of image-level

TABLE 1. Anomaly detection performance on MVTEC-AD in the term of image-level AUROC, where bold values denote the best results and underlined values are the results just following. Our method achieves superior performance on 12 out of 15 objects and the best anomaly detection result on average.

Method→ ↓Object	DRAEM	UniAD	SPADE	CutPaste	PaDiM	PatchCore	DifferNet	FastFlow	CFLOW-AD	F ² R (Ours)
bottle	99.2	100.0	97.2	98.2	99.9	100.0	99.0	100.0	100.0	100.0
cable	91.8	97.6	84.8	81.2	92.7	99.0	95.9	100.0	97.59	<u>99.89</u>
capsule	98.5	85.3	89.7	98.2	91.3	97.8	86.9	100.0	97.68	<u>99.72</u>
carpet	97.0	<u>99.9</u>	92.8	93.9	99.8	99.2	92.9	100.0	98.73	<u>99.32</u>
grid	<u>99.9</u>	98.5	47.3	100.0	96.7	98.0	84.0	99.7	99.60	99.41
hazelnut	100.0	99.9	88.1	98.3	92.0	100.0	99.3	100.0	<u>99.98</u>	100.0
leather	100.0	100.0	95.4	100.0	100.0	100.0	<u>97.1</u>	100.0	100.0	100.0
metal nut	98.7	99.0	71.0	<u>99.9</u>	98.7	99.5	96.1	100.0	99.26	100.0
pill	<u>98.9</u>	88.3	80.1	94.9	93.3	95.6	88.8	99.4	96.82	98.34
screw	93.9	91.9	66.7	88.7	85.8	96.4	96.3	<u>97.8</u>	91.89	98.71
tile	99.6	99.0	96.5	94.6	98.1	99.4	99.4	100.0	99.88	<u>99.89</u>
toothbrush	100.0	95.0	88.9	99.4	96.1	100.0	98.6	94.4	<u>99.65</u>	100.0
transistor	93.1	100.0	90.3	96.1	97.4	98.9	91.1	<u>99.8</u>	95.21	100.0
wood	99.1	97.9	95.8	99.1	99.2	99.2	99.8	100.0	99.12	<u>99.91</u>
zipper	100.0	96.7	96.6	99.9	90.3	98.8	95.1	99.5	98.48	<u>99.92</u>
average	98.0	96.6	85.4	96.1	95.5	98.8	94.9	<u>99.4</u>	98.26	99.67

TABLE 2. Anomaly localization performance on MVTEC-AD in the term of pixel-level AUROC, where bold values denote the best results and underlined values are the results just following. Our method achieves superior performance on 8 out of 15 objects and the best pixel-level AUROC on average.

Method→ ↓Object	DRAEM	UniAD	SPADE	CutPaste	PaDiM	PatchCore	FastFlow	CFLOW-AD	F ² R (Ours)
bottle	99.1	98.1	98.4	97.6	98.3	98.7	97.7	98.76	<u>98.80</u>
cable	94.7	96.8	97.2	90.0	96.7	98.3	<u>98.4</u>	97.64	98.45
capsule	94.3	97.9	99.0	97.4	98.5	99.0	99.1	98.98	<u>99.05</u>
carpet	95.5	98.0	97.5	98.3	99.1	98.7	99.4	<u>99.23</u>	99.13
grid	99.7	94.6	93.7	97.5	97.3	98.1	98.3	96.89	<u>98.92</u>
hazelnut	99.7	98.8	99.1	97.3	98.2	98.5	<u>99.1</u>	98.82	98.83
leather	98.6	98.3	97.6	<u>99.5</u>	99.2	99.3	<u>99.5</u>	99.61	99.43
metal nut	99.5	95.7	98.1	93.1	97.2	98.6	98.5	98.56	<u>98.96</u>
pill	97.6	95.1	96.5	95.7	95.7	96.9	99.2	<u>98.95</u>	98.81
screw	97.6	97.4	98.9	96.7	98.5	99.3	99.4	98.10	99.51
tile	99.2	91.8	87.4	90.5	94.1	97.5	96.3	<u>97.71</u>	97.49
toothbrush	98.1	97.8	97.9	98.1	98.8	98.8	98.9	98.56	<u>98.84</u>
transistor	90.9	98.7	94.1	93.0	<u>98.5</u>	95.7	97.3	93.28	<u>97.52</u>
wood	<u>96.4</u>	93.4	88.5	95.5	94.9	95.7	97.0	94.49	96.27
zipper	98.8	96.0	96.5	99.3	98.5	98.3	98.7	98.41	<u>99.12</u>
average	97.3	96.6	96.0	96.0	97.5	98.1	<u>98.5</u>	97.87	98.61

anomaly detection performance. As to the evaluation of the pixel-wise anomaly localization performance, the pixel-wise AUROC and the area under the per-region-overlap curve (AUPRO) are reported, where the AUPRO score accounts better for varying anomaly sizes. Please refer to [64] for more evaluation details.

B. COMPARISON TO STATE-OF-THE-ARTS

1) ANOMALY DETECTION ON MVTEC-AD

The results of image-level anomaly detection on MVTEC-AD benchmark are reported in Table 1, where recent state-of-the-art methods [4], [12], [13], [14], [33], [34], [38], [39], [43] are listed for comparison. DRAEM [34] and UniAD [4] are two reconstruction-based methods. SPADE [12], CutPaste [33], PaDiM [13], and PatchCore [14] are representation-based methods. In addition, DifferNet [38], FastFlow [39], and CFLOW-AD [43] are listed as they adopt normalizing flow

models to project features to latent representation before anomaly score calculation. On the contrary, our method reversed the latent representation back to the original feature space for anomaly score calculation. The best setting of hyperparameter α and β is found by grid search from 0.0 to 1.2 with a step size of 0.1 and the best results of each object are reported for comparison. With the ability to intervene in the process of information integration, our method achieves an averaged image-level anomaly detection AUROC score of up to 99.67%, almost halving the error compared to the next best competitor, which is 99.4%. Besides, our method achieves superior performance in detecting defects on 12 out of 15 objects.

2) ANOMALY LOCALIZATION ON MVTEC-AD

Examples of visualized qualitative results are shown in Fig. 3. In addition, the detailed quantitative results are reported in the

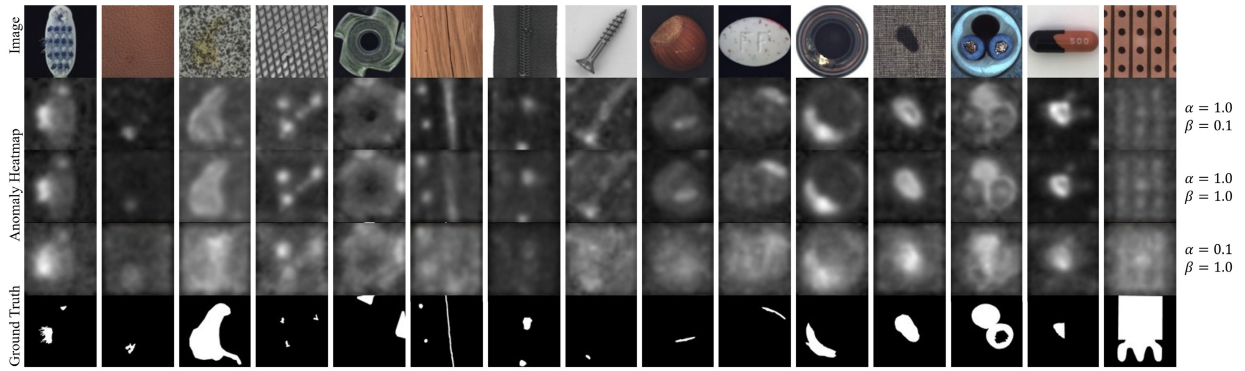


FIGURE 4. Anomaly heatmaps of localization examples with different α and β . The information preferences from left to right objects are roughly ordered from local to global, whereas the samples in the middle prefer balanced local and global information to spot abnormalities.

TABLE 3. Anomaly localization performance on MVTec-AD in the term of AUPRO, where bold values denote the best results and underlined values are the results just following. Our method achieves superior performance on 11 out of 15 objects and the best AUPRO on average.

Method → ↓Object	SPADE	PaDiM	PatchCore	CFLOW-AD	F ² R (Ours)
bottle	95.5	94.8	96.5	96.80	<u>96.63</u>
cable	90.9	88.8	<u>97.9</u>	93.53	97.94
capsule	93.7	93.5	<u>94.3</u>	93.40	94.64
carpet	94.7	<u>96.2</u>	95.5	97.70	95.73
grid	86.7	94.6	93.3	<u>96.08</u>	96.84
hazelnut	95.4	92.6	89.9	96.68	91.57
leather	97.2	97.8	97.0	99.35	<u>98.14</u>
metal nut	94.4	85.6	93.9	91.65	<u>93.95</u>
pill	94.6	92.7	93.9	<u>95.39</u>	95.37
screw	<u>96.0</u>	94.4	96.1	<u>95.30</u>	96.57
tile	75.9	86.0	89.9	<u>94.34</u>	95.03
toothbrush	93.5	93.1	<u>95.5</u>	95.06	95.63
transistor	87.4	84.5	<u>89.0</u>	81.40	91.59
wood	97.4	91.1	85.5	<u>95.79</u>	90.44
zipper	92.6	95.9	94.2	<u>96.60</u>	97.24
average	91.7	92.1	93.5	<u>94.60</u>	95.15

term of pixel-wise AUROC in Table.2, and AUPRO in Table.3 if available. Consistently, our method achieves the best averaged pixel-level AUROC and AUPRO of 98.61% and 95.15% respectively. For previous methods that fail to intervene in the inherited local and global information determined by pre-trained models, they may perform better when spotting defects on certain objects but compromises on others. For example, DRAEM outperforms other methods on the bottle, grid, hazelnut, and metal nut data. However, it performs unsatisfactorily on other objects, resulting in a lower average pixel-wise AUROC. For the second best method, FastFlow performs closely to ours but misses the importance and effectiveness of the intervention in local and global information, which empowers our method to achieve better performance in industrial anomaly detection and localization.

In particular, CFLOW-AD explicitly adds positional embedding into features using a conditional flow-based model, implying that the feature embedding lacks global information to perceive its position. In contrast, our method assumes that the feature map extracted by pre-trained models contains both local semantic and global position information

TABLE 4. Anomaly detection and localization results on MVTec-AD of the two implementations of the proposed method. The metrics are reported in a triplet of image-level AUROC, pixel-wise AUROC, and AUPRO, where bold values denote superior results.

Implementation → ↓Object	Ours(Reg.)			Ours(Cls.)		
bottle	100.0	98.70	96.63	100.0	98.80	96.63
cable	99.93	98.47	97.94	99.89	98.45	97.94
capsule	99.20	99.02	94.64	99.72	99.05	94.64
carpet	99.32	99.15	95.87	99.32	99.13	95.73
grid	99.25	98.75	96.37	99.41	98.92	96.84
hazelnut	100.0	98.75	91.38	100.0	98.83	91.57
leather	100.0	99.34	98.21	100.0	99.43	98.14
metal nut	99.90	98.73	94.25	100.0	98.96	93.95
pill	98.80	98.75	95.04	98.34	98.81	95.37
screw	98.93	99.50	96.58	98.71	99.51	96.57
tile	99.64	97.61	94.79	99.89	97.49	95.03
toothbrush	100.0	98.80	95.63	100.0	98.84	95.63
transistor	100.0	97.45	91.29	100.0	97.52	91.59
wood	99.74	96.15	89.40	99.91	96.27	90.44
zipper	99.76	99.09	97.10	99.92	99.12	97.24
average	99.63	98.55	95.01	99.67	98.61	95.15

as an entirety, of which the insight guides us to conduct feature factorization and achieves better results in general.

C. ABLATION STUDY

1) ON THE IMPLEMENTATION OF FEATURE FACTORIZATION LOSS

Our method has two implementations depending on the realization of the second term in representation factorization loss \mathcal{L}_f . Specifically, when $q(\mathbf{f}|\mathbf{z}_1)$ is assumed to follow Gaussian, the second term of \mathcal{L}_f could be supervised as a regression task, denoted as *Ours (Reg.)*. Besides, when $q(\mathbf{f}|\mathbf{z}_1)$ is assumed to follow the binomial distribution, this term could be supervised as a binary classification task, denoted as *Ours (Cls.)*. Accordingly, the performance of anomaly detection and localization on MVTec-AD of these two variants is reported in terms of image-level AUROC, pixel-wise AUROC, and AUPRO in Table.4, which shows that *Ours (Cls.)* performs a little better on most objects and $q(\mathbf{f}|\mathbf{z}_1)$ may be more realistic to be a binomial distribution. Consequently, *Ours (Cls.)* is used as the default implementation in other experiments.

TABLE 5. Spearman (Partial) Correlations between the variable and pixel-wise AUROC, where variable includes α , β , and $-|\alpha - \beta|$, showing information preferences and the necessity of information equilibrium.

Variable (Partial) Corr. \rightarrow \downarrow Object	α		β		$- \alpha - \beta $	
	ρ	p-val	ρ	p-val	ρ	p-val
bottle	0.68	1e-23	-0.16	0.04	0.66	1e-22
cable	-0.01	0.22	0.28	2e-4	0.72	6e-28
capsule	-0.15	0.05	0.55	2e-14	0.36	2e-6
carpet	0.31	3e-5	0.50	8e-12	0.62	2e-19
grid	0.70	2e-25	-0.57	1e-15	0.48	7e-11
hazelnut	0.74	3e-30	0.26	6e-4	0.51	1e-12
leather	0.84	2e-46	-0.70	7e-26	0.09	0.24
metal nut	0.61	5e-18	-0.60	5e-18	0.40	7e-8
pill	0.47	2e-10	-0.68	9e-24	0.43	7e-9
screw	0.75	1e-31	-0.25	1e-3	0.35	3e-6
tile	0.77	2e-33	-0.55	1e-14	0.40	9e-8
toothbrush	0.94	2e-76	0.51	1e-12	0.34	9e-6
transistor	-0.61	4e-18	0.54	8e-14	0.51	2e-12
wood	0.73	9e-29	-0.47	1e-10	0.47	1e-10
zipper	0.72	3e-28	-0.47	2e-10	0.54	3e-14

2) ON THE INFORMATION PREFERENCES TO SPOT ABNORMALITIES

The performance of the proposed method depends on the values of scaling factors α and β , reflecting the information preferences for spotting different types of abnormalities. Fig. 4 illustrates examples of anomaly heatmaps for all objects in the MVTec-AD dataset, where objects on the left demand local information to detect defects, while objects on the right benefit from global information. In most cases of texture data and unaligned objects, global information would introduce more background noise to the heatmaps, as seen in the samples from leather to hazelnut. In contrast, aligned objects require global information to refine the results of anomaly localization, as shown by the clearer and more salient heatmaps from pill to cable samples. Additionally, heatmaps that prioritize local information ($\beta = 0.1$) are more similar to the unaltered anomaly heatmaps ($\alpha = \beta = 1.0$) than those prioritizing global information ($\alpha = 0.1$), indicating that most defects can be detected using only local semantic information. However, global information could be more useful in detecting missing parts, as in the cases of cable, capsule, and transistor samples.

In addition, Table 5 reports the Spearman partial correlations [65] between α (β) and pixel-wise AUROC, as well as the Spearman correlations between negative difference $-|\alpha - \beta|$ and pixel-wise AUROC as the quantitative result for the stated discovery, which shows that most texture data and unaligned objects prefer local semantic information for anomaly detection and localization, as evidenced by highly positive correlations between α and pixel-wise AUROC. On the other hand, capsule and transistor samples prefer only global information, as evidenced by relatively high positive correlations between β and pixel-wise AUROC with negative correlations between α and pixel-wise AUROC. For other objects, both local and global information are required in equilibrium, as shown by relatively high positive correlations between $-|\alpha - \beta|$ and pixel-wise AUROC. Therefore, these quantitative results confirm the discovery that different types of defects require distinct information to spot.

TABLE 6. Computational complexity versus the performance of industrial anomaly detection and localization in terms of averaged image-level AUROC (iAUROC) and pixel-wise AUROC (pAUROC). Our method achieves the best performance on average with decent computational complexity, inferred by the number of multiply-accumulate operations (MACs) and parameters (Params).

	Model	Computational Complexity		Performance	
		MACs(G)	Params(M)	iAUROC	pAUROC
DRAEM	UNet	151.93	97.42	98.00	97.30
SPADE	WR50	9.23	24.86	85.40	96.00
PaDiM		9.23	24.86	95.50	97.50
PatchCore		9.23	24.86	98.80	98.10
DifferNet	WR50+INN	48.61	196.95	94.90	-
CFLOW-AD		25.83	106.51	98.26	97.87
FastFlow		46.25	82.63	99.40	98.50
F ² R (Ours)		43.27	80.72	99.67	98.61

3) ON THE COMPUTATIONAL COMPLEXITY VERSUS THE ANOMALY DETECTION AND LOCALIZATION PERFORMANCE

Table 6 compares the performance and computational complexity of several methods based on different neural networks, including UNet [66], WideResNet-50 (WR50) [8], and different invertible neural networks (INN) [15], [57], [58], [67]. DRAEM, as a reconstruction-based method, requires the most multiply-accumulate operations (MACs) for image reconstruction. On the contrary, representation-based methods, such as SPADE, PaDiM, and PatchCore, require fewer computational operations and have a smaller number of network parameters (Params). However, SPADE performs k-nearest-neighbor clustering between the test and gallery samples, which is typically slower than convolutional neural networks, while PaDiM stores training-time statistics that demand large memory. To this end, PatchCore proposes a feature selection process to reduce the required memory.

Regarding the methods based on both WR50 and INN, DifferNet has the highest computational operations and the number of network parameters. For the remaining methods, CFLOW-AD adopts a relatively large model, and FastFlow still requires a large amount of computational operations. In comparison, our method achieves the best performance with decent computational complexity. Furthermore, when compared to the methods based only on WR50, F²R achieves the most significant performance gain among the methods based on both WR50 and INN, demonstrating the effectiveness of the information intervention process in the proposed method.

V. CONCLUSION

In this paper, we studied the information preferences for spotting different defects in industrial anomaly detection and localization. To this end, a framework termed F²R is proposed to factorize multi-level features into two latent representations that encode local semantic information and global positional information respectively, based on which distinct information can be emphasized or suppressed by request to recognize different defects. As the result, our method not only outperforms previous state-of-the-art methods but also provides more insights into information preferences for spotting different defects by ablation studies.

As to the limitation of the proposed method, there are two hyper-parameters α and β that need manual adjustment in this work, which is expected to be automatic via reinforcement learning. Besides, the ablation study shows that it could be reasonable when $q(\mathbf{f}|\mathbf{z}_1)$ is trained following a binomial distribution, whereas a more comprehensive and preferable distribution needs investigation in the future work.

REFERENCES

- [1] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AICHE J.*, vol. 37, no. 2, pp. 233–243, Feb. 1991.
- [2] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Dec. 2014, pp. 1–14.
- [3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [4] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le, "A unified model for multi-class anomaly detection," in *Proc. Adv. Neural Inf. Process. Syst.*, Nov. 2022, pp. 4571–4584.
- [5] L. Bergman, N. Cohen, and Y. Hoshen, "Deep nearest neighbor anomaly detection," 2020, *arXiv:2002.10445*.
- [6] N. Li, K. Jiang, Z. Ma, X. Wei, X. Hong, and Y. Gong, "Anomaly detection via self-organizing map," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 974–978.
- [7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 770–778.
- [8] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2016, pp. 87.1–87.12.
- [9] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 2117–2125.
- [10] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, Jul. 2017, pp. 2961–2969.
- [11] G. Ghiasi, T. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. CVPR*, Apr. 2019, pp. 7036–7045.
- [12] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," 2020, *arXiv:2005.02357*.
- [13] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in *Proc. Int. Conf. Pattern Recognit. Workshops Challenges*, vol. 12664, Jan. 2021, pp. 475–489.
- [14] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14318–14328.
- [15] A. N. Gomez, M. Ren, R. Urtasun, and R. B. Grosse, "The reversible residual network: Backpropagation without storing activations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 1–11.
- [16] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9592–9600.
- [17] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proc. MLSDA*, Dec. 2014, pp. 4–11.
- [18] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," in *Proc. 14th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2019, pp. 1–12.
- [19] D. T. Nguyen, Z. Lou, M. Klar, and T. Brox, "Anomaly detection with multiple-hypotheses predictions," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2019, pp. 4800–4809.
- [20] A.-S. Collin and C. De Vleeschouwer, "Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7915–7922.
- [21] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps, "Towards visually explaining variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8642–8651.
- [22] T. Matsubara, K. Sato, K. Hama, R. Tachibana, and K. Uehara, "Deep generative model using unregularized score for anomaly detection with heterogeneous complexity," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5161–5173, Jun. 2022.
- [23] N. Kozamernik and D. Bračun, "Visual inspection system for anomaly detection on KTL coatings using variational autoencoders," *Proc. CIRP*, vol. 93, pp. 1558–1563, Jan. 2020.
- [24] S. Pidhorskyi, R. Almoheisen, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 1–12.
- [25] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3379–3388.
- [26] S. Akcay, A. Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, May 2019, pp. 622–637.
- [27] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: One-class novelty detection using GANs with constrained latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2898–2906.
- [28] M. Z. Zaheer, J.-H. Lee, M. Astrid, and S.-I. Lee, "Old is gold: Redefining the adversarially learned one-class classifier training paradigm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14183–14193.
- [29] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. V. D. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.
- [30] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 14372–14381.
- [31] D. Dehaene, O. Frigo, S. Combretelle, and P. Eline, "Iterative energy-based projection on a normal data manifold for anomaly localization," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2020, pp. 1–17.
- [32] M. Pourreza, B. Mohammadi, M. Khaki, S. Bouindour, H. Snoussi, and M. Sabokrou, "G2D: Generate to detect anomaly," in *Proc. IEEE Winter Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2003–2012.
- [33] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 9664–9674.
- [34] V. Zavrtanik, M. Kristan, and D. Skocaj, "DRAEM—A discriminatively trained reconstruction embedding for surface anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8330–8339.
- [35] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9737–9746.
- [36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.
- [37] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2015, pp. 1530–1538.
- [38] M. Rudolph, B. Wandt, and B. Rosenhahn, "Same same but DifferNet: Semi-supervised defect detection with normalizing flows," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1907–1916.
- [39] J. Yu, Y. Zheng, X. Wang, W. Li, Y. Wu, R. Zhao, and L. Wu, "FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows," 2021, *arXiv:2111.07677*.
- [40] Z. Xiao, X. Xu, H. Xing, S. Luo, P. Dai, and D. Zhan, "RTFN: A robust temporal feature network for time series classification," *Inf. Sci.*, vol. 571, pp. 65–86, Sep. 2021.
- [41] Z. Xiao, H. Zhang, H. Tong, and X. Xu, "An efficient temporal network with dual self-distillation for electroencephalography signal classification," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2022, pp. 1759–1762.
- [42] H. Xing, Z. Xiao, R. Qu, Z. Zhu, and B. Zhao, "An efficient federated distillation learning system for multitask time series classification," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

- [43] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 98–107.
- [44] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philos. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 374, Apr. 2016, Art. no. 20150202.
- [45] E. Collins, R. Achanta, and S. Susstrunk, "Deep feature factorization for concept discovery," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 336–352.
- [46] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.
- [47] X. Chen, Y. Duan, R. Houthoof, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, vol. 29, Dec. 2016, pp. 2180–2188.
- [48] H. Kim and A. Mnih, "Disentangling by factorising," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2018, pp. 2649–2658.
- [49] P. Esser, R. Rombach, and B. Ommer, "A disentangling invertible interpretation network for explaining latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9223–9232.
- [50] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, Dec. 2011, pp. 2178–2186.
- [51] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5542–5550.
- [52] V. Piratla, P. Netrapalli, and S. Sarawagi, "Efficient domain generalization via common-specific low-rank decomposition," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2020, pp. 7728–7738.
- [53] M.-H. Bui, T. Tran, A. Tran, and D. Phung, "Exploiting domain-specific features to enhance domain generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Dec. 2021, pp. 21189–21201.
- [54] E. Collins and S. Susstrunk, "Deep feature factorization for content-based image retrieval and localization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 874–878.
- [55] X. W. X. Yang and J. Ye, "Factorizing knowledge in neural networks," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 73–91.
- [56] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," 2014, *arXiv:1410.8516*.
- [57] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2017, pp. 1–32.
- [58] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 10236–10245.
- [59] F. Rosenblatt, "Principles of neurodynamics. Perceptrons and the theory of brain mechanisms," Cornell Aeronaut. Lab., Buffalo, NY, USA, Tech. Rep., VG-1196-G-8, Jan. 1961.
- [60] J. Yi and S. Yoon, "Patch SVDD: Patch-level SVDD for anomaly detection and segmentation," in *Proc. Asian Conf. Comput. Vis.*, Nov. 2020, pp. 375–390.
- [61] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2015, pp. 448–456.
- [62] A. L. Maas, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2013, pp. 3–8.
- [63] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
- [64] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4183–4192.
- [65] K. Baba, R. Shibata, and M. Sibuya, "Partial correlation and conditional correlation as measures of conditional independence," *Austral. New Zealand J. Statist.*, vol. 46, no. 4, pp. 657–664, 2004.
- [66] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2015, pp. 234–241.
- [67] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 4743–4751.



YUHAO ZHU received the B.S. degree in automation from Central South University, Changsha, Hunan, China, in 2016, and the M.S. degree in computer engineering from New York University, New York City, NY, USA, in 2018. He is currently pursuing the Ph.D. degree in traffic information engineering and control with the China Academy of Railway Sciences, Beijing, China.

From 2018 to 2021, he was an Engineer with the Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, pattern recognition, biometrics, and related fields.



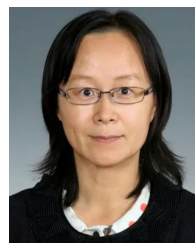
LINLIN DAI received the B.S. degree in computer science and technology from Jimei University, Xiamen, Fujian, China, in 2004, and the M.S. degree in network and switching technology from the Beijing University of Posts and Telecommunications, Beijing, China, in 2007.

From 2007 to 2010, she was a Software Research and Development Engineer and the Project Manager of the IBM China System and Technology Development Center. Since 2010, she has been a Software Research and Development Engineer. She is currently the Technical Director of the Fundamental Platform Department, Institute of Computing Technology, China Academy of Railway Sciences. She has been working on system formulation, development promotion, and scientific research in the fields of infrastructure platform construction, intelligent service, software, and hardware development for passenger transport.



XINGZHI DONG received the B.S. degree in software and traffic engineering from Dalian Jiaotong University, Dalian, Liaoning, China, in 2016, and the M.S. and Ph.D. degrees in traffic information engineering and control from the China Academy of Railway Sciences, Beijing, China, in 2022.

Since 2022, he has been a Research Assistant with the Institute of Computing Technology, China Academy of Railway Sciences. His research interests include big data analysis and mining digital and wireless communications.



PING LI received the B.S. degree in mechatronics from Shenyang Ligong University, Shenyang, Liaoning, China, in 1993, the M.S. degree in computer control systems from the Beijing Institute of Technology, Beijing, China, in 1996, and the Ph.D. degree in traffic information engineering and control from the China Academy of Railway Sciences, Beijing, in 2001.

She was an Associate Research Fellow and a Research Fellow with the Institute of Computing Technology, China Academy of Railway Sciences, from 1996 to 2009 and from 2009 to 2021. Since 2021, she has been a Chief Researcher with the Chinese Academy of Sciences. She is the author of seven books and more than 160 articles. She holds two patents. Her research interests include intelligent railway, big data, information planning, security emergency management, and related fields.

Dr. Li is a member of the Digital Platform Committee of International Union of Railways (IUC) and Intelligent Service Professional Committee of China Artificial Intelligence Society. She is the Deputy Secretary-General of the Rail Intelligent Transportation Sub-Committee of China Intelligent Transportation Association and Intelligent Railway Committee of China Railway Society. She is the Associate Editor-in-Chief of the *Smart and Resilient Transportation*. She serves on the editorial board for the *China Railway Science* and the *China Railway*.