

# PyramidFlow: High-Resolution Defect Contrastive Localization using Pyramid Normalizing Flow

Jiarui Lei<sup>1,2</sup> Xiaobo Hu<sup>1,2</sup> Yue Wang<sup>1,2</sup> Dong Liu<sup>1,2,3,4\*</sup>

<sup>1</sup>State Key Laboratory of Modern Optical Instrumentation, Zhejiang University

<sup>2</sup>ZJU-Hangzhou Global Scientific and Technological Innovation Center, China

<sup>3</sup>Jiaxing Key Laboratory of Photonic Sensing & Intelligent Imaging, China

<sup>4</sup>Intelligent Optics & Photonics Research Center, Jiaxing Research Institute Zhejiang University

{karrilett, huxiaobo, 426195, liudongopt}@zju.edu.cn

## Abstract

During industrial processing, unforeseen defects may arise in products due to uncontrollable factors. Although unsupervised methods have been successful in defect localization, the usual use of pre-trained models results in low-resolution outputs, which damages visual performance. To address this issue, we propose PyramidFlow, the first fully normalizing flow method without pre-trained models that enables high-resolution defect localization. Specifically, we propose a latent template-based defect contrastive localization paradigm to reduce intra-class variance, as the pre-trained models do. In addition, PyramidFlow utilizes pyramid-like normalizing flows for multi-scale fusing and volume normalization to help generalization. Our comprehensive studies on MVTecAD demonstrate the proposed method outperforms the comparable algorithms that do not use external priors, even achieving state-of-the-art performance in more challenging BTAD scenarios.

## 1. Introduction

Due to the uncontrollable factors in the complex industrial manufacturing process, unforeseen defects will be brought to products inevitably. As the human visual system has the inherent ability to perceive anomalies [25], quality control relies on manual inspection for a long time. However, large-scale images and tiny defects are challenging for manual inspection, so increasing research is focused on automated machine vision inspection. Among all the methods, supervised deep learning has achieved great success. It relies on annotated datasets to learn discriminative features, effectively overcoming the hand-crafted shortcomings. However, because of insufficient negative samples, the high demand for labels, and the absence of prior knowledge, those approaches based on supervised learning may

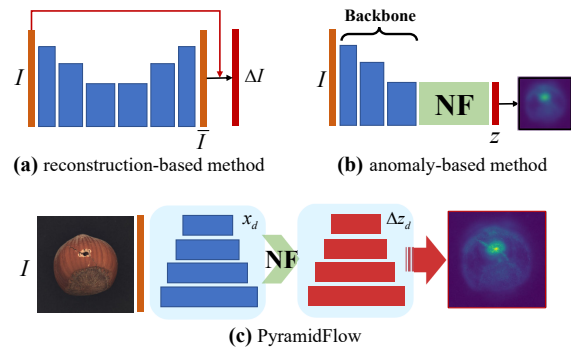


Figure 1. Illustration of various anomaly localization methods. (a) Reconstruction-based method. (b) Anomaly-based method, where NF denotes normalizing flow. (c) Our PyramidFlow, which combines latent templates and normalizing flow, enables high-resolution localization.

suffer in identifying unseen defects in practices,

Recently, unsupervised methods have been applied to defect detection, as shown in Fig. 1(a,b). Reconstruction-based methods [4, 15, 23, 29] are the most famous, which take reconstructed images as templates and then apply explicit contrast in image space to achieve high-resolution localization. However, reconstructing using decoders is an ill-posed inverse problem, it is hard to reconstruct complex details. To overcome the above limitations, anomaly-based methods [6, 7] utilizing texture-aware pre-trained models achieves high image-level performance, which also damages pixel-level visual performance. One of the most promising methods is convolutional normalizing flows [10, 22, 27], which models the probability distribution further from pre-trained features, earning higher performance.

In this paper, a Pyramid Normalizing Flow (PyramidFlow) is proposed. It develops the idea of templates from image space into latent space by normalizing flow, then performing contrast  $\Delta z_d$  for high-resolution anomaly localization, as shown in Fig. 1(c). Specifically, we propose the multi-scale Pyramid Coupling Block, which includes

invertible pyramid and volume normalization, as the critical module to construct volume-preserving PyramidFlow. To the best of our knowledge, PyramidFlow is the first UNet-like fully normalizing flow specifically designed for anomaly localization, analogous to UNet [19] for biomedical image segmentation. Our main contributions can be summarized as follows.

- We propose a latent template-based defect contrastive localization paradigm. Similar to the reconstruction-based methods, we perform contrast localization in latent space, which avoids the ill-posedness and reduces intra-classes variance efficiently.
- We propose PyramidFlow, which includes invertible pyramids and pyramid coupling blocks for multi-scale fusing and mapping, enabling high-resolution defect localization. Additionally, we propose volume normalization for improving generalization.
- We conduct comprehensive experiments to demonstrate that our advanced method outperforms comparable algorithms that do not use external priors, and even achieves state-of-the-art performance in complex scenarios.

## 2. Related Work

### 2.1. Deep learning-based Defect Localization

With the rise of deep learning, numerous works apply generalized computer vision methods for defect detection. Some works are based on object detection [13,14,28,30,31], which relies on annotated rectangular boxes, enabling locating and classifying defects end-to-end. The other is applying semantic segmentation [5,18,24], which enables pixel-level localization, suit for complex scenarios with difficult-to-locate boundaries. However, these works still rely on supervised learning, they attempt to collect sufficient defective samples to learn well-defined representations.

Recently, some promising work has considered the scarcity of defects in real-world scenarios, where defect-free samples are only obtained. These methods can be classified as reconstruction-based and anomaly-based. The reconstruction-based method relies on generative models such as VAE or GAN, which encode a defective image and reconstruct it to a defect-free image, then localize the defect with the contrast of these two images. The reconstruction-based method performs well on single textural images, but they cannot generalize to non-textural images for ill-posedness and degeneracy [25]. The anomaly-based method treats defects as anomalous, applying neural networks to discriminate between normality and anomalous. These methods extract pre-trained features, then estimate their probability density using Mahalanobis distances

or K-NearestNeighbor, while the lower probability indicates where the image patches are abnormal. Although anomaly-based methods had achieved great success in defect detection, it locates defects with low pixel-level resolution compared with reconstruction-based methods, usually 1/16th or even lower, which greatly limits practical industrial applications.

To overcome existing shortcomings, we propose a latent template-based defect contrastive localization paradigm, which breaks the limitation of low-frequency texture-aware-only models, enabling more accurate results.

### 2.2. Normalizing Flow

Normalizing flow is a kind of invertible neural network with bijective mappings and traceable Jacobi determinants. It was first proposed for nonlinear independent component estimation [8] and applied to anomaly detection [21] recently for its invertibility helps prevent mode collapse. The normalizing flow comprises coupling blocks, these basic modules for realizing nonlinear mappings and calculating Jacobi determinants. Originally, NICE [8] proposed the additive coupling layer with unitary Jacobi determinants, while RealNVP [9] further proposed the affine coupling layer that enables the generation of non-volume-preserving mappings. However, redundant volume degrees of freedom can lead to increased optimization complexity, creating a domain gap between maximum likelihood estimation and anomaly metrics, which may potentially compromise the generalization performance in anomaly detection.

Previous works [10,21,27] on anomaly localization usually follow the methods proposed in RealNVP, but some challenges remain. Some studies [22] have found that convolutional normalizing flow focuses on local rather than semantic correlations, which are usually addressed by image embeddings [12]. Hence, earlier studies [21] adopted pre-trained backbones, while recent trends used pre-trained encoders to extract image patches [10,22,27]. However, pre-trained-based methods rely on task-irrelevant external priors, which limit generalization in unforeseen scenarios.

To address the above challenges, we propose a pyramid-like normalizing flow called PyramidFlow, which utilizes volume normalization to preserve volume mappings that include task-relevant implicit priors. Additionally, our method offers the option of using pre-trained models, and we have observed that external priors from pre-trained models can improve generalization performance. We will discuss these contributions in Sec. 4.3 and Sec. 4.4.

## 3. Methodology

Our algorithm consists of two processes, training and evaluation, as shown in Fig. 2. The training process is similar to siamese networks, the model is optimized by minimizing the Frequency differences  $\|\mathcal{F}(\Delta z_d)\|$  within the im-

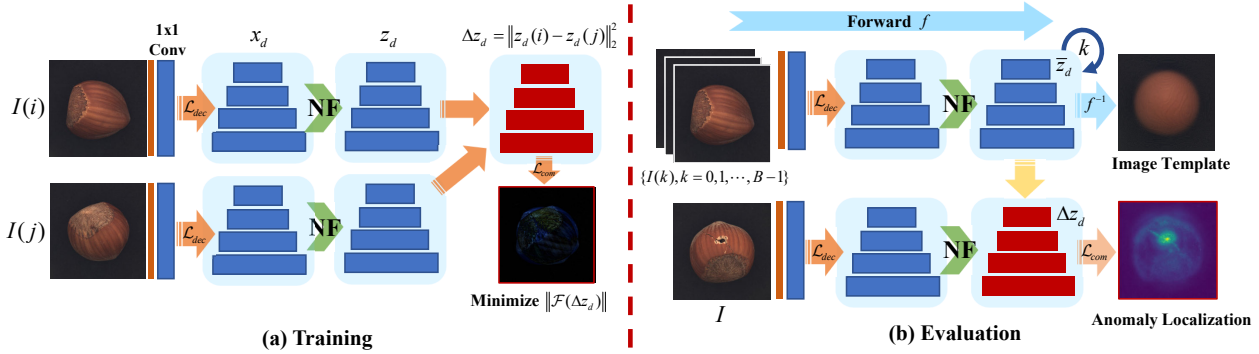


Figure 2. Schematic of training and evaluation for PyramidFlow. **(a)** Given any normality pair, minimize the distance of latent variables. **(b)** The means of the latent variables are contrasted to the examples, then apply pyramid composition to obtain an anomaly localization map.

age pair. For the evaluation process, latent templates are obtained through inference at the total training dataset, then latent contrast and pyramid composition are applied to obtain an anomaly localization map. The details are shown in the following sections.

### 3.1. Invertible Pyramid

Defect images contain various frequency components. Usually, the low-frequency components represent the slow gradient background, while the high-frequency components correspond to details or defects. To decouple the frequency components and identify each frequency component independently, we propose invertible pyramids, which enable multi-scale decomposition and composition for a single feature. To facilitate feature learning, previous work applies pre-trained encoders to extract features. Although pre-trained methods with external priors help performance improvement, to fully explore the advantages of our approach in our primary study, let's consider a baseline without any pre-trained model.

For a three-channel image  $I$ , apply orthogonally initialized  $1 \times 1$  convolution  $\mathbf{W} \in \mathbb{R}^{C \times 3}$  to the image for obtain features  $x = \mathbf{W}I$ . Given a feature  $x$  and a positive integer  $L$ , the pyramid decomposition is a mapping from features to feature sets  $\mathcal{L}_{dec} : x \rightarrow \{x_d | d \in \mathbb{Z}_{L-1}\}$ , where  $x_d$  is the  $d$ -level pyramid can be calculated as

$$x_d = D^d(x) - U(D^{d+1}(x)) \quad (1)$$

where  $D(\cdot)$  and  $U(\cdot)$  are arbitrary linear upsampling and downsampling operators, while  $D^d(\cdot)$  represents repeated downsampling  $d$  times. If Eq. (1) is further satisfied  $D^0(x) = x, D^L(x) = 0$ , then the inverse operation  $\mathcal{L}_{com} : \{x_d | d \in \mathbb{Z}_{L-1}\}$  of the pyramid decomposition can be described as

$$x = \sum_{d=0}^{L-1} U^d(x_d) \quad (2)$$

Differing from Gaussian pyramid, Eq. (2) indicates that

there is always an inverse operation for pyramid decomposition, which is called pyramid composition. The method based on Eqs. (1) and (2), enabling perform multi-scale feature decomposition and composition, is a critical invertible module for PyramidFlow.

### 3.2. Pyramid Coupling Block

**Invertible Modules.** Invertible modules are the essential elements to implementing invertible neural networks. The invertible modules introduced in this paper are invertible convolution, invertible pyramid, and affine coupling block. The affine coupling block is the basic module that constitutes the normalizing flow. It is based on feature splitting for invertible nonlinear mappings with easily traceable Jacobian determinants and inverse operations.

As shown in Fig. 3(a), the conventional affine coupling block splits a single feature along the channel dimension, where one sub-feature keeps its identity while another is performed affine transformation controlled by it. Denote the splitted features are  $x_0, x_1$  and its outputs are  $y_0, y_1$ , then the corresponding transformation can be described as

$$\begin{aligned} y_0 &= x_0 \\ y_1 &= \exp(s(x_0)) \odot x_1 + t(x_0) \end{aligned} \quad (3)$$

where  $s(\cdot), t(\cdot)$  are affine parameters, can be estimated by zero-initialized convolutional neural networks. For formula(3), there is an explicit inverse transformation:

$$\begin{aligned} x_0 &= y_0 \\ x_1 &= \exp(-s(y_0)) \odot (y_1 - t(y_0)) \end{aligned} \quad (4)$$

Denote the element at position  $i, j$  of  $s(\cdot)$  as  $s_{i,j}(\cdot)$ . As the Jacobian matrix of transformation(3) is a triangular matrix, its logarithmic determinant can be estimated as

$$\log \left| \frac{\partial(y_0, y_1)}{\partial(x_0, x_1)} \right| = \sum_{i,j} s_{i,j}(x_0) \quad (5)$$

Eqs. (3) to (5) are the basis of all affine coupling blocks. However, the coupling block shown in Fig. 3(a) remains

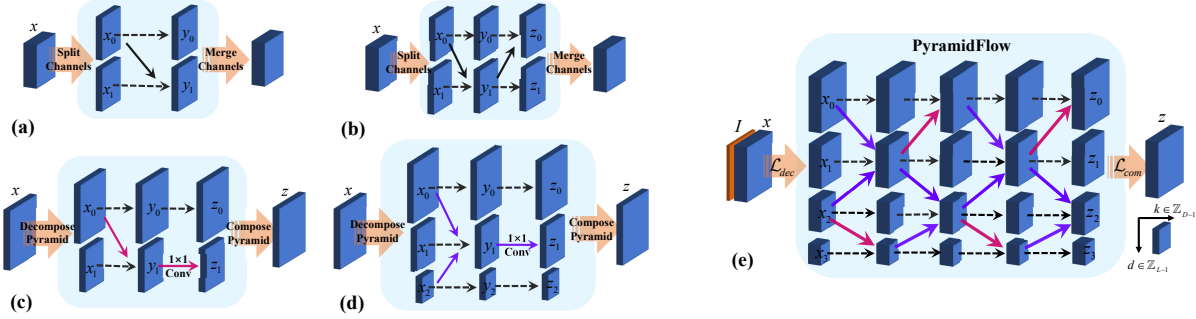


Figure 3. The proposed pyramid coupling block and PyramidFlow, where the solid line symbolizes the transformation while the dotted line refers to identity. (a) Channel-splitting affine coupling block. (b) The reverse cascade of (a)-architecture. (c) The proposed scale-wise pyramid coupling block. (d) The reverse parallel and reparameterized of (c)-architecture. (e) The proposed PyramidFlow, is a stacking of (c,d)-architecture both in depths and layers, where  $1 \times 1$  convolution is neglected to represent.

identical for one part. Therefore the reverse cascade architecture is proposed in NICE [8] such that both parts are transformed, as shown in Fig. 3(b). The previous works construct the holistic invertible normalizing flow by iterative applying the structure shown in Fig. 3(b).

**Implementation.** Our method decomposes a single feature along the scale and realizes multi-scale feature fusion based on Eqs. (3) to (5). In our implementation, the multi-scale affine parameters  $s(\cdot), t(\cdot)$  are estimated using a convolutional neural network with two linear layers, where bilinear interpolation is applied to match the target shape.

In addition, we employ invertible  $1 \times 1$  convolution [11] for feature fusion within features. Specifically, denoting the full rank matrix corresponding to the invertible  $1 \times 1$  convolution as  $\mathbf{A}$ , which can be decomposed by PLU as

$$\mathbf{A} = \mathbf{P}\mathbf{L}(\mathbf{U} + \text{diag}(\exp(s_i))) \quad (6)$$

where  $\mathbf{P}$  is a frozen permutation matrix,  $\mathbf{L}$  is a lower triangular matrix with unit diagonal elements,  $\mathbf{U}$  is an upper triangular matrix with zero diagonal elements, and  $\exp(s_i)$  is the  $i$ -th eigenvalue of the matrix  $\mathbf{A}$ , which always holds nonnegativity. The matrix  $\mathbf{A}$  is always invertible during optimization, then its logarithmic Jacobian determinant can be estimated as

$$\log |\mathbf{A}| = \sum_i s_i \quad (7)$$

In summary, Eqs. (3) to (7) describe proposed pyramidal coupling block mathematically, as shown in Fig. 3(c). First, multi-scale feature fusion(3-5) is performed, and then apply linear fusion(6-7) for shuffle channels. Furthermore, we propose a dual coupling block as shown in Fig. 3(d), which is equivalent to the reverse parallel of the coupling block in Fig. 3(c). The dual coupling block is reparameterized in our implementation, and its affine parameters  $s(\cdot), t(\cdot)$  are estimated from concatenated features.

**Volume Normalization.** Suppose that the invertible transformation  $f : x \rightarrow z$  maps the variable  $x$  to the latent variable  $z$ . Previous works have assumed that the latent variable

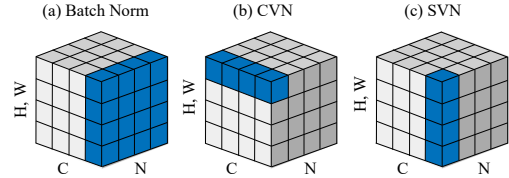


Figure 4. Illustration of volume normalization. (a) Batch Normalization. (b) The proposed Channel Volume Normalization (CVN). (c) The proposed Spatial Volume Normalization (SVN).

follows basic probability distribution (e.g. Gaussian distribution), then estimates sample probability density based on the following equation:

$$P(x) = P(z) \left| \frac{\partial f(x)}{\partial x} \right| \quad (8)$$

However, this approach relies on the basic distribution assumption and ignores the effect of the implicit prior in the probability density transform on generalization. When such approaches are applied to anomaly detection, the inconsistency between the training objectives and the anomaly evaluation results in domain gaps.

Similar to batch normalization or instance normalization in deep learning, the proposed volume normalization will be employed for volume-preserving mappings, as illustrated in Fig. 4. Particularly, for the affine coupling block, the parameter  $s(\cdot)$  is subtracted from its mean value before performing Eq. (3); for the invertible convolution, the parameter  $s_i$  is subtracted from its mean value before calculating the matrix  $\mathbf{A}$  based on Eq. (6). Depending on the statistical dimension, we propose Spatial Volume Normalization (SVN) and Channel Volume Normalization (CVN). SVN performs mean statistics along the spatial dimension, while CVN is along the channel dimension. Various volume normalization methods contain different priors, then we will explore their impact in Sec. 4.2.

### 3.3. Pyramid Normalizing Flow

**Architecture.** Our PyramidFlow can be obtained by stacking the pyramid coupled blocks of Fig. 3(c,d) along the depth  $D - 1$  times and along the layer  $L - 1$  times, as



shown in Fig. 3(e). Specifically, PyramidFlow boosts the image  $I$  to feature  $x$  using matrix  $\mathbf{W}$ , then performs the pyramid decomposition based on Eq. (1). The pyramid coupling blocks described in Eqs. (3) to (7) are calculated in the order described in Fig. 3(e) to obtain potential pyramid features  $z_d, d = 0, 1, \dots, L - 1$ , which are finally composed into latent variables according to Eq. (2).

**Loss Function.** In the cases with volume normalization, the loss function excludes the probability density coefficients. Moreover, the logarithmic Jacobian determinant of the semi-orthogonal matrix  $\mathbf{W}$  is sample-independent, so its effect could be ignored during training.

Suppose a training batch with 2 normal samples, its latent variables are  $z_d(i), z_d(j)$ . Previous studies train neural networks using spatial difference  $\Delta z_d = \|z_d(i) - z_d(j)\|^2$ . However, it ignored the impact of high-frequency defects. To address the above shortcoming, we propose the following Fourier loss function.

$$\mathcal{L}_{loss} = \|\mathcal{F}(\mathcal{L}_{com}(\{\Delta z_d | d \in \mathbb{Z}_{L-1}\}))\| \quad (9)$$

where  $\mathcal{F}$  is the fast Fourier transform of the image. Training the normalizing flow using Eq. (9) enables the model to focus on the high-frequency, allowing faster convergence. We will discuss this trick in Sec. 4.3.

**Defect Localization.** Previous studies [22, 27] usually localize defects with obvious differences based on category-independent zero templates. In our method, the defects are modeled as anomalous deviations with respect to the template. Then, the anomaly of the latent pyramid  $z_d$  is defined as  $\sigma(z_d) = \|z_d - \bar{z}_d\|$ , where  $\bar{z}_d$  is mean of the latent pyramid. Finally, the total anomaly can be estimated as

$$\sigma(z) = \mathcal{L}_{com}(\{\sigma(z_d) | d \in \mathbb{Z}_{L-1}\}) \quad (10)$$

The Eq. (10) shows that the total anomaly is a composition of anomalies at various scales, which is consistent with the empirical method proposed by Rudolph, *et al.* [22].

**Image Template Estimation.** The image template is a prototype of normal samples, a visualization of the latent template. Our fully normalizing flow is based on  $1 \times 1$  convolution instead of pre-trained encoders, maintaining end-to-end and near-invertibility, thus the flow’s input  $x_{temp}$  can be retrieved using Eq. (2) and Eq. (4) from latent mean  $\bar{z}_d$ , then solve the least square problem  $\mathbf{W}I_{temp} = x_{temp}$  for image template  $I_{temp}$ .

## 4. Experiment and Discussion

In this chapter, we perform unsupervised anomaly localization experiments on MVTec Anomaly Detection Dataset [2] (MVTecAD) and BeanTech Anomaly Detection Dataset [16] (BTAD). MVTecAD contains 15 categories of industrial defect images, five of which are textural images and the other ten are object images. The object images contain

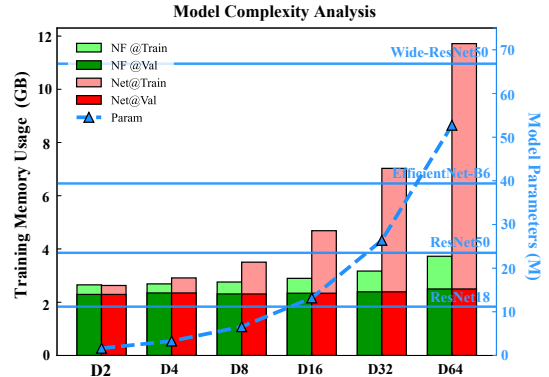


Figure 5. Analysis of model complexity for various depths. The bars correspond to the Training Memory Usage (GB) in the left vertical coordinate, while the line graph and horizontal lines relate to the Model Parameters (M) on the other side. For each depth  $D$ , the left bar presents the normalizing flow implemented based on *autoFlow* framework with memory-saving tricks, while the right is implemented by PyTorch with auto-differentiation.

three classes (grid, metal nut, screw) without rough registration and one class (hazelnut) without fine registration, and we will discuss these cases in Sec. 4.5. The BTAD contains three types of real-world and industry-oriented textural images, which is more challenging for pixel-level localization.

All experiments take Area Under the Receiver Operating characteristic Curve (AUROC) and Area Under the Per Region Overlap (AUPRO) as evaluation metrics. AUROC is the most widely used anomaly evaluation metric, and higher values indicate that various thresholds have less impact on performance. However, AUROC prefers larger anomalies and may fail in small proportions of anomalies. Thus, we further evaluate AUPRO for localization metric, similar to Intersection Over Union (IoU) commonly used in semantic segmentation. Detailed definitions can be found in [2].

### 4.1. Complexity Analysis

The normalizing flow based on Eqs. (3) and (4) is computationally invertible, which indicates that only one copy of the variables is necessary for all stages. This feature decreases the memory footprint during backpropagation from linear to constant complexity. We have analyzed the above characteristics based on a fixed number of pyramid layers  $L = 8$ , image resolution with  $256 \times 256$ , and channels  $C = 24$ , then changed the number of stacked layers  $D$  to explore the trends of memory usage and model parameters. The forward and memory-saving backpropagation is implemented based on the self-developed PyTorch-based [17] framework *autoFlow*. All indicators are recorded during steady-state training, then plotted as bar and line graphs, as shown in Fig. 5.

The memory usage based on auto-differentiation increases linearly with depth  $D$ , while the implementation based on normalizing flow achieves approximate depth-independent memory usage. The memory superiority enables the proposed method to be trained in memory-constrained devices below 4G without powerful hardware. The line graph shows the exponential trends between model parameters and depth, while the horizontal line represents the parameters of the usual pre-trained model. We mainly adopt methods with  $D < 8$  or even shallower, where the number of parameters is far smaller than popular pre-training-based methods. In summary, in scenarios of memory constraint, the proposed PyramidFlow enables dealing with larger images and requires fewer parameters than others.

### 4.2. Study on Volume Normalization

In this subsection, based on MVTecAD, we investigate the impact of volume normalization on generalization. The experiment without data augmentation, fixed pyramid layers  $L = 4$ , channels  $C = 16$ , and linear interpolated image to  $256 \times 256$ . During the training, the volume normalization applies sample mean normalization and updates the running mean with 0.1 momenta, while in the testing, the volume normalization is based on the running mean. We sufficiently explored the volume normalization methods proposed in Sec. 3.2. Some representative categories are shown as Tab. 1.

Table 1. Quantitative results of CVN and SVN on different categories. For each case in the table, the first column is Pixel-AUROC% and the second is AUPRO%, while the values within parentheses represent the relative improvement.

Classes	CVN		SVN	
	AUROC	AUPRO	AUROC	AUPRO
capsule	<b>96.1(+2.6)</b>	<b>93.1(+5.1)</b>	93.5(+0.0)	88.0(+0.0)
pill	<b>96.2(+1.8)</b>	<b>96.3(+1.4)</b>	94.4(+0.0)	94.9(+0.0)
toothbrush	<b>98.9(+2.5)</b>	<b>97.9(+4.3)</b>	96.4(+0.0)	93.6(+0.0)
carpet	88.9(+0.0)	88.3(+0.0)	<b>90.8(+1.9)</b>	<b>91.0(+2.7)</b>
grid	86.2(+0.0)	84.5(+0.0)	<b>94.2(+8.0)</b>	<b>92.7(+8.2)</b>
zipper	92.2(+0.0)	91.9(+0.0)	<b>95.4(+3.2)</b>	<b>95.1(+3.2)</b>

The result in Tab. 1 shows the performance differences between the various volume normalization methods: CVN outperforms SVN for the first three classes, while the latter behaves the opposite. We further visualize these defect distributions in Fig. 6, which shows that SVN-superior classes are commonly textural images with a larger range of defects, while CVN-superior classes are object images.

SVN with larger receptive fields achieves non-local localization by aggregating an extensive range of texture features, while CVN realizes accurate localization by shuffling channels. In a word, different volume normalization techniques implicitly embody distinct task-specific priors. Fur-

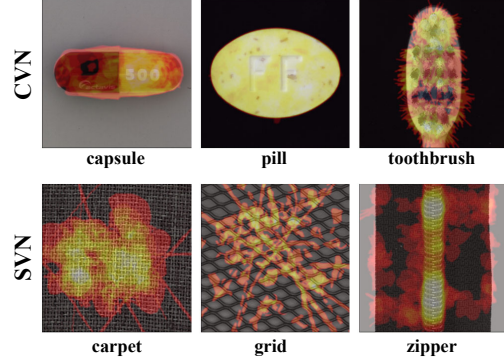


Figure 6. Defects in Tab. 1 are visualized as heat maps. The top row displays CVN-superior class object images, while the bottom row displays SVN-superior class texture images.

thermore, our ablation study in Sec. 4.3 shows that volume normalization does help to improve average performance.

### 4.3. Ablation Study

This subsection discusses the impact of some proposed methods on performance. The study is conducted on full MVTecAD, and other settings are the same as Sec. 4.2. We ablate four methods from baseline individually. In particular, experiment I is based on latent Gaussian assumption and Eq. (8) without volume normalization. For experiment II, the category-independent zero template  $\bar{z}_d = 0$  is applied. Then, Experiment III does not adopt the method of Eq. (10) but composes the pyramid first and localizes its difference later. Finally, experiment IV adopts a spatial version of the loss function instead of Eq. (9). The result of the above ablation experiments is shown in Tab. 2.

Table 2. The ablation study on full MVTecAD. For each cell in the table, the first row is Pixel-AUROC% and the second is AUPRO%. The number within parentheses means the change relative to baseline, the larger absolute value with larger importance.

Method	Classes		
	Texture	Object	MEAN
Ours(baseline)	<b>95.2(+0.0)</b>	<b>95.7(+0.0)</b>	<b>95.5(+0.0)</b>
I. w/o Volume Normalization	89.4(-5.8)	85.2(-10.5)	86.6(-8.9)
II. w/o Latent Template	93.1(-2.1)	90.7(-5.0)	91.5(-4.0)
III. w/o Pyramid Difference	87.8(-7.4)	93.1(-2.6)	91.3(-4.2)
IV. w/o Fourier Loss	92.0(-3.2)	93.3(-2.4)	92.9(-2.6)

Tab. 2 demonstrates that experiments I-IV present various performance degradation. Experiment I has the largest average degradation, with the object classes being more affected. Although the non-volume-preserving enables larger outputs and higher Image-AUROC performance, the im-

Table 3. Quantitative results of various challenging methods on MVTEcAD. In the table, the fully normalized flow method is labeled as FNF, while the abbreviations Res18, WRes50, EffiB5, and DTD are denoted as ResNet18, Wide-ResNet50-2, EfficientNet-B5, and Describable Textures Dataset, respectively. For each case in the table, the first row is Pixel-AUROC% and the second is AUPRO%, where the best results are marked in bold.

External Prior	Methods	carpet	leather	tile	wood	bottle	cable	capsule	hazelnut	pill	toothbrush	transistor	zipper	MEAN
×	AnoGAN [23]	54.2	64.1	49.7	62.1	85.8	78.0	84.1	87.1	86.8	90.0	79.9	78.1	75.0
		20.4	37.8	17.7	38.6	62.0	38.3	30.6	69.8	77.6	74.9	54.9	46.7	47.4
	Vanilla VAE [15]	62.0	83.5	52.0	69.9	89.4	81.6	90.7	95.1	87.9	95.3	85.1	77.5	80.8
		61.9	64.9	24.2	57.8	70.5	77.9	77.9	77.0	79.3	85.4	61.0	60.8	66.6
	AE-SSIM [4]	87.0	78.0	59.0	73.0	93.0	82.0	94.0	97.0	91.0	92.0	80.0	88.0	84.5
		64.7	56.1	17.5	60.5	83.4	47.8	86.0	91.6	83.0	78.4	72.4	66.5	67.3
	Ours (FNF)	<b>90.8</b>	<b>99.6</b>	<b>97.9</b>	<b>93.8</b>	<b>95.9</b>	<b>92.1</b>	<b>96.1</b>	<b>98.0</b>	<b>96.2</b>	<b>98.9</b>	<b>97.4</b>	<b>95.4</b>	<b>96.0</b>
		<b>91.0</b>	<b>99.7</b>	<b>95.8</b>	<b>96.2</b>	<b>94.0</b>	<b>86.4</b>	<b>93.1</b>	<b>97.3</b>	<b>96.3</b>	<b>97.7</b>	<b>91.4</b>	<b>95.1</b>	<b>94.5</b>
Res18	S-T [3]	93.5	97.8	92.5	92.1	97.8	91.9	96.8	98.2	<b>96.5</b>	97.9	73.7	95.6	93.7
		87.9	94.5	94.6	91.1	93.1	81.8	96.8	96.5	<b>96.1</b>	93.3	66.6	95.1	90.6
WRes50	SPADE [6]	97.5	97.6	87.4	88.5	<b>98.4</b>	<b>97.2</b>	<b>99.0</b>	<b>99.1</b>	<b>96.5</b>	97.9	94.1	96.5	95.8
		94.7	97.2	75.9	87.4	95.5	90.9	93.7	95.4	94.6	93.5	<b>97.4</b>	92.6	92.4
WRes50	PaDiM [7]	<b>99.1</b>	99.2	94.1	94.9	98.3	96.7	98.5	98.2	95.7	<b>98.8</b>	<b>97.5</b>	<b>98.5</b>	<b>97.5</b>
		96.2	97.8	86.0	91.1	94.8	88.8	93.5	92.6	92.7	93.1	84.5	<b>95.9</b>	92.3
EffiB5	CS-Flow [22]	98.0	98.4	93.9	88.6	90.9	95.3	97.9	96.3	95.7	96.3	95.5	96.4	95.3
		<b>98.0</b>	98.5	94.5	92.9	88.7	<b>94.0</b>	96.1	95.1	91.1	89.9	96.9	95.4	94.2
DTD	DRÆM [29]	94.9	<b>96.6</b>	<b>99.6</b>	<b>97.3</b>	97.6	95.4	94.0	99.2	95.0	98.1	90.0	94.4	96.0
		96.1	97.9	<b>99.7</b>	<b>97.9</b>	<b>97.2</b>	90.4	96.5	<b>98.7</b>	93.7	97.1	92.9	94.7	96.1
Res18	Ours	97.4	98.7	97.1	97.0	97.8	91.8	98.6	98.1	96.1	98.5	96.9	96.6	97.1
		97.2	<b>99.2</b>	97.2	<b>97.9</b>	95.5	90.3	<b>98.3</b>	98.1	<b>96.1</b>	<b>97.9</b>	94.7	95.4	<b>96.5</b>

plicit prior in volume normalization discussed in Sec. 4.2 is more helpful for generalization. For experiment II, it shows that the latent template benefits the performance, and object classes are improved greatly. It is because the category-specific latent template reduces intra-class variance, helping convergence during training. Then, experiment III suggests that multi-scale differences had a more pronounced impact on textural classes, as higher-level operators with larger receptive fields correspond to large defects. Finally, Experiment IV reveals that Fourier loss (9) is the icing on the cake that helps performance improvement. To summarize, methods I -III are critical for the proposed model, while tricks IV help further improvements.

#### 4.4. Anomaly Localization

**MVTecAD.** We performed defect localization for 12 registered classes in MVTEcAD. In our comparisons, the method based on pre-trained models or using external datasets is viewed as requiring external prior, corresponding to the first column of Tab. 3. In our implementation, we augment textural classes with flips and rotations, each with a probability of 0.5, while object categories do not undergo any augmentation operation. It is worth noting that those base on complex augmentation or weak supervision is not considered in our comparisons, as our approach is capable of incorporating these techniques to improve performance. The detailed results are shown in Tab. 3.

First, we take three methods based on image contrast, AnoGAN [23], Vanilla VAE [15], and AE-SSIM [4]. They

are not dependent on external datasets, so it is fair to compare them with our FNF model. Furthermore, we also compared our method to those that utilize external priors, such as S-T, SPADE, *etc.* All methods were reproduced based on the official implementation or AnomaLib [1]. For fair comparisons, we adapted the  $1 \times 1$  convolution  $\mathbf{W}$  to the pre-trained encoder, where the pre-trained encoder is the first two layers of ResNet18 for extracting the image into features of original 1/4 size and with 64 channels.

As shown in Tab. 3, our FNF method greatly outperforms the comparable methods without external priors, even exceeding S-T, SPADE, and CS-Flow that using external priors. Most of the reconstruction-based methods in Tab. 3 suffer from ill-posedness in complex scenarios (*e.g.*, tile and wood.), while our method achieves the best AUPRO score owing to high-resolution contrast in latent space. However, a larger resolution implies larger intra-class variance, which degrades the overall AUROC performance for hard-to-determine anomaly boundaries. Furthermore, Fig. 7(a) visualizes representative examples of MVTEcAD anomaly localization, which shows that our method achieves precise localization with reasonable scale.

**BTAD.** To fully illustrate our superiority, we experimented on the more challenging BTAD dataset without any data augmentation, and other settings are the same as MVTEcAD. The detailed result in Tab. 4 shows that our method also achieves state-of-the-art performance, and Fig. 7(b) visualizes representative examples of BTAD anomaly localization.

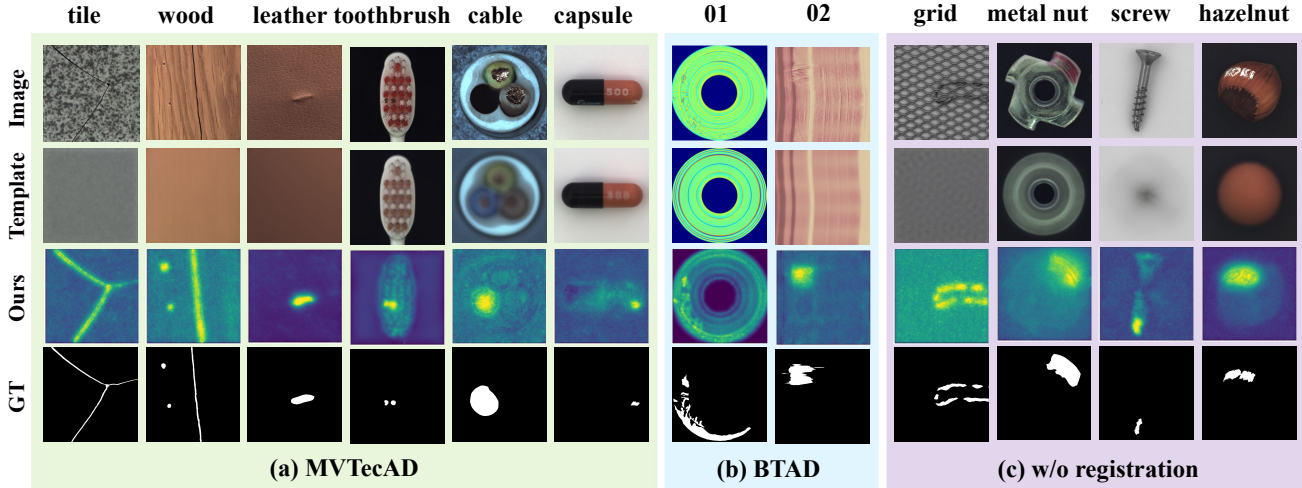


Figure 7. Visualization of our results on MVTecAD and BTAD. From top to bottom are original images, estimated image templates, our localization results, and ground truths. (a) The six challenging results on MVTecAD. (b) Two representative results on BTAD. (c) Results for the four unregistered categories on MVTecAD.

Table 4. Quantitative results of various challenging methods on BTAD. For each case in the table, the first row is Image-AUROC% and the second is Pixel-AUROC%, where the best results are marked in bold.

Methods	Classes			MEAM
	01	02	03	
VT-ADL [16]	97.6	71.0	82.6	83.7
	<b>99.0</b>	94.0	77.0	90.0
P-SVDD [26]	95.7	72.1	82.1	83.3
	91.6	93.6	91.0	92.1
SPADE [6]	91.4	71.4	<b>99.9</b>	87.6
	97.3	94.4	99.1	96.9
PatchCore [20]	90.9	79.3	99.8	90.0
	95.5	94.7	<b>99.3</b>	96.5
PaDiM [7]	99.8	82.0	99.4	93.7
	97.0	96.0	98.8	97.3
Ours (Res18)	<b>100.0</b>	<b>88.2</b>	99.3	<b>95.8</b>
	97.4	<b>97.6</b>	98.1	<b>97.7</b>

#### 4.5. Study on unregistered categories

In principle, template-based methods require pixel-level registration between images and templates, which is typically satisfied in most real-world scenarios, but may fail in some cases. This subsection explores the performance of the proposed method on unregistered (*e.g.* rotation, shift) categories (*e.g.* grid, metal nut, screw, and hazelnut), as shown in Tab. 5. The reconstruction-based AE-SSIM heavily relies on pixel-level contrast, which decreases the average localization accuracy (AUPRO%). In contrast, the anomaly-based SPADE avoids registration issues and achieves better performance. Fortunately, our ResNet18-based method, which utilizes normalizing flow to reduce patch variance, remains competitive in unregistered scenes, although it falls short of state-of-the-art performance. We

visualize these categories in Fig. 7(c).

Table 5. Quantitative results on unregistered categories without Rough Registration (RR) or Fine Registration (FR). For each case in the table, the first row is Pixel-AUROC% and the second is AUPRO%, where the best results are marked in bold.

Methods	Classes				MEAN
	w/o RR		w/o FR		
	grid	metal nut	screw	hazelnut	
AE-SSIM [4]	94.0	89.0	96.0	97.0	94.0
	84.9	60.3	88.7	91.6	81.4
SPADE [6]	93.7	<b>98.1</b>	<b>98.9</b>	<b>99.1</b>	<b>97.5</b>
	86.7	<b>94.4</b>	<b>96.0</b>	95.4	93.1
Ours (Res18)	<b>95.7</b>	97.2	94.6	98.1	96.4
	<b>94.3</b>	91.4	94.7	<b>98.1</b>	<b>94.6</b>

## 5. Conclusion

In this paper, we propose PyramidFlow, the first fully normalizing flow method based on the latent template-based contrastive paradigm, utilizing pyramid-like normalizing flows and volume normalization, enabling high-resolution defect contrastive localization. Our method can be trained end-to-end from scratch, similar to UNet, and our comprehensive experiments demonstrate that it outperforms comparable algorithms that do not use external priors, even achieving state-of-the-art performance in complex scenarios. While experiments on unregistered categories show that our method falls short of state-of-the-art, it still exhibits competitive performance. Future research will focus on improving performance in such scenarios.

**Acknowledgment.** This work was supported by the State Key Laboratory of Modern Optical Instrumentation (MOI2021ZD01); Fundamental Research Funds for the Central Universities (2021XZZX019).



## References

- [1] S. Akcay, D. Ameln, A. Vaidya, B. Lakshmanan, N. Ahuja, and U. Genc. Anomalib: A deep learning library for anomaly detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1706–1710. 7
- [2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600. 5
- [3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4183–4192. 7
- [4] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *VISIGRAPP (5: VISAPP)*, 2019. 1, 7, 8
- [5] Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 2021. 2
- [6] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *CoRR*, abs/2005.02357, 2020. 1, 7, 8
- [7] Thomas Defard, Aleksandr Setkov, Angélique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer. 1, 7, 8
- [8] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. 2, 4
- [9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 2
- [10] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107. 1, 2
- [11] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 4
- [12] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems*, 33:20578–20589, 2020. 2
- [13] Feng Li, Feng Li, and QingGang Xi. Defectnet: Toward fast and effective defect detection. *IEEE Transactions on Instrumentation and Measurement*, 70:1–9, 2021. 2
- [14] J. Luo, Z. Yang, S. Li, and Y. Wu. Fpcb surface defect detection: A decoupled two-stage object detection framework. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11, 2021. 2
- [15] Takashi Matsubara, Kazuki Sato, Kenta Hama, Ryosuke Tachibana, and Kuniaki Uehara. Deep generative model using unregularized score for anomaly detection with heterogeneous complexity. *IEEE Transactions on Cybernetics*, 2020. 1, 7
- [16] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. 5, 8
- [17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5
- [18] R. Ren, T. Hung, and K. C. Tan. A generic deep-learning-based approach for automated surface inspection. *IEEE Transactions on Cybernetics*, 48(3):929–940, 2018. 2
- [19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2
- [20] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328. 8
- [21] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but different: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1907–1916. 2
- [22] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1088–1097. 1, 2, 5, 7
- [23] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer. 1, 7
- [24] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020. 2
- [25] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 2022. 1, 2

- [26] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*. 8
- [27] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *CoRR*, abs/2111.07677, 2021. 1, 2, 5
- [28] Xuyi Yu, Wentao Lyu, Di Zhou, Chengqun Wang, and Weiqiang Xu. Es-net: Efficient scale-aware network for tiny defect detection. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022. 2
- [29] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 1, 7
- [30] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, and X. Liu. A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2022. 2
- [31] Jiabin Zhang, Hu Su, Wei Zou, Xinyi Gong, Zhengtao Zhang, and Fei Shen. Cadn: A weakly supervised learning-based category-aware object detection network for surface defect detection. *Pattern Recognition*, 109:107571, 2021. 2