Unveiling Multi-View Anomaly Detection: Intra-view Decoupling and Inter-view Fusion

Kai Mao*, **Yiyang Lian***, **Yangyang Wang**, **Meiqin Liu**, **Nanning Zheng**, **Ping Wei**†

National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China

Abstract

Anomaly detection has garnered significant attention for its extensive industrial application value. Most existing methods focus on single-view scenarios and fail to detect anomalies hidden in blind spots, leaving a gap in addressing the demands of multi-view detection in practical applications. Ensemble of multiple single-view models is a typical way to tackle the multi-view situation, but it overlooks the correlations between different views. In this paper, we propose a novel multi-view anomaly detection framework, Intra-view Decoupling and Inter-view Fusion (IDIF), to explore correlations among views. Our method contains three key components: 1) a proposed *Consistency Bottleneck* module extracting the common features of different views through information compression and mutual information maximization; 2) an *Implicit Voxel Construction* module fusing features of different views with prior knowledge represented in the form of voxels; and 3) a *View-wise Dropout* training strategy enabling the model to learn how to cope with missing views during test. The proposed IDIF achieves state-of-the-art performance on three datasets. Extensive ablation studies also demonstrate the superiority of our methods.

Code — https://github.com/Kerio99/IDIF

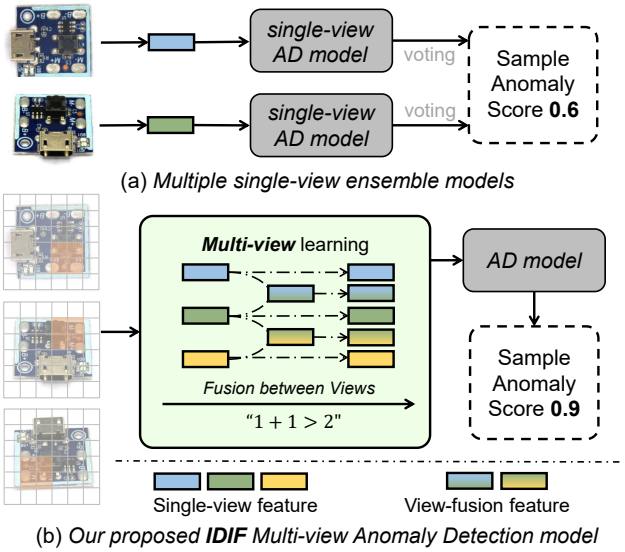

Figure 1: Illustration of ensemble of single-view anomaly detection (AD) models (*Top*) and our IDIF multi-view Anomaly Detection model (*Bottom*).

Introduction

Anomaly detection aims to distinguish different patterns (i.e., abnormal samples) from known normal patterns. It has various applications, such as industrial defect detection (Roth et al. 2022), medical image analysis (Raza and Singh 2021; Liu, Aviles-Rivero, and Schönlieb 2023), and autonomous driving (Zhang et al. 2023a).

While existing methods have made remarkable progress in anomaly detection recently (Bergmann et al. 2019; Roth et al. 2022; Liu et al. 2023), most methods focus on single-view anomaly detection, which would lead to missing anomalies hidden in blind spots. It is necessary to detect anomalies from multiple views. Training the model for each view individually and voting on the results might be a straightforward option, as shown in Fig. 1(a). However, such protocol is computationally expensive and neglects the potential correlations and complementary information between different views, which could be critical for identifying certain types of anomalies.

On actual production lines, we find that human inspectors check areas that are potentially abnormal from multiple angles. Observation from different views can provide information about how the same area behaves under different lighting conditions or depth flatness. Representations of anomalous regions of different views are complementary. We believe a single model of multi-view anomaly detection can take advantage of the complementary property to improve the confidence of anomaly detection, as shown in Fig. 1 (b).

In this paper, we propose an Intra-view Decoupling and Inter-view Fusion (IDIF) model for multi-view anomaly detection. Different from ensemble of models for each view, IDIF pays more attention to utilizing the complementary features of different views. Through intra-view decoupling, IDIF is able to decompose the single view image features

*These authors contributed equally.

†Ping Wei is the corresponding author (pingwei@xjtu.edu.cn).

into the view-common and view-specific features. Afterwards, the process of inter-view fusion is able to fuse the view-specific features of different views to obtain richer representations for subsequent anomaly detection.

A *Consistency Bottleneck (CB)* module is proposed for implementing the intra-view decoupling. In the *CB* module, information from all views is compressed into a lower dimensional space. The compressed features are driven to capture the view-common features by maximizing mutual information. The view-common features are used as the reference to extract the view-specific features from the original features. These view-specific features contain complementary information about the same region. Therefore, integrating and exchanging information between these features can provide a more comprehensive description of the region.

An *Implicit Voxel Construction (IVC)* module is proposed for inter-view fusion. Concretely, we design a learnable sample-independent 3D voxel prototype as the basis of the fusion. With successive interactions between the 3D voxel and different view-specific features, the prior knowledge in the prototype and information from different views are integrated. Finally, the fused 3D feature is mapped back from different angles to the 2D features. The model is supervised in the form of denoising knowledge distillation with a pretrained teacher network.

Multi-view detection models usually require a strictly fixed view number of images as the input. However, in real deployment situations, images from some views may be unavailable. Robust multi-view methods must retain the detection ability with incomplete views. The proposed *View-wise Dropout* is a simple but effective strategy to handle the case of missing views. By randomly dropping some views during training and incorporating a designed masked-view cross attention mechanism, the model can make predictions on various combination of missing views.

We evaluate our method on the large-scale, multi-view anomaly detection dataset Real-IAD (Wang et al. 2024). We additionally process two 3D datasets MvTec 3D-AD (Bergmann et al. 2019) and Eyecandies (Bonfiglioli et al. 2022) into the multi-view 2D form. Our IDIF achieves the state-of-the-art performance on all three datasets.

Our contributions are summarized below:

1. We propose a multi-view anomaly detection framework IDIF, which leverages the *Consistency Bottleneck* module for Intra-view Decoupling and the *Implicit Voxel Construction* module for Inter-view Fusion.
2. We propose a *View-wise Dropout* strategy that enables the multi-view model to maintain detection performance with missing-view data.
3. We conduct extensive experiments on three datasets and our method achieves the state-of-the-art performance.

Related Work

Anomaly Detection

Many studies have been proposed to distinguish unknown anomalies based on normal images. Memory-based methods (Cohen and Hoshen 2020; Roth et al. 2022) store normal patterns in a memory bank, then predict anomaly scores by searching for the most similar patterns during testing. Reconstruction-based methods (Zavrtanik, Kristan, and Skočaj 2022; Zhang et al. 2023b; Zhang, Xu, and Zhou 2024) restore anomalous images to normal ones and then compare their differences to determine the anomaly scores. Some other methods (Deng and Li 2022; Zhang et al. 2023c; Gu et al. 2024) introduce knowledge distillation into anomaly detection. DeSTSeg (Zhang et al. 2023c) emphasizes training the student network to match feautres of the teacher network by using synthetically corrupted normal images. The proposed denoising procedure enables the student network to learn more robust representations.

These studies have greatly advanced anomaly detection. However, most studies focus on the single-view scenarios. In real industrial applications, detecting anomalies from multiple views is essential for achieving more accurate sample-level predictions.

Multi-View Tasks

Multi-view representation learning aims to utilize the characteristics of multiple views to extract better information. Canonical Correlation Analysis (CCA) (Hotelling 1992) and DCCA (Andrew et al. 2013) are classical approaches for unsupervised cross-view representation learning. They project two different views into one common latent space where the two views are maximally correlated. DMF-MVC (Zhao, Ding, and Fu 2017) proposes to utilize deep matrix factorization to extract common representations. TC (Hwang et al. 2021) measures the amount of information shared between different viewpoint representations through KL divergence and designs the objective optimization function accordingly.

3D reconstruction is another task related to our work. It aims at recovering 3D models of objects from images of different views. Pix2Vox (Xie et al. 2019) and Attsets (Yang et al. 2020) use convolutional layers to extract global features and fuse them to restore 3D voxels. EVolT (Wang et al. 2021) and Legoformer (Yagubbayli et al. 2021) perform 3D reconstruction based on Transformer (Vaswani 2017) architecture. Recently, Nerf (Mildenhall et al. 2021), Gaussian Splatting (Kerbl et al. 2023) and other related methods are proposed and have achieved impressive reconstruction results. However, the demand for intrinsic and extrinsic parameters of cameras makes them unusable in some scenes.

Method

Problem Formulation

Let $\mathcal{X}_{train}$ be the training set consisting exclusively of normal samples and $\mathcal{X}_{test}$ be the test set including both normal and abnormal samples. Each sample $\boldsymbol{x}_i \in \mathcal{X}_{train} \cup \mathcal{X}_{test}$ is represented as $\boldsymbol{x}_i = \{x_i^{(1)}, x_i^{(2)}, ..., x_i^{(V)}\}$, which consists of images from V distinct views. $x_i^{(v)} \in \mathbb{R}^{C \times H \times W}$ denotes the v-th view image of the i-th sample, where C, H and W represent the number of channels, height and width of the image respectively. For the sake of neatness of the formula, the subscript i will be omitted afterwards. The model is designed to learn from normal samples in $\mathcal{X}_{train}$ and generalize to detect both normal and abnormal samples in $\mathcal{X}_{test}$.

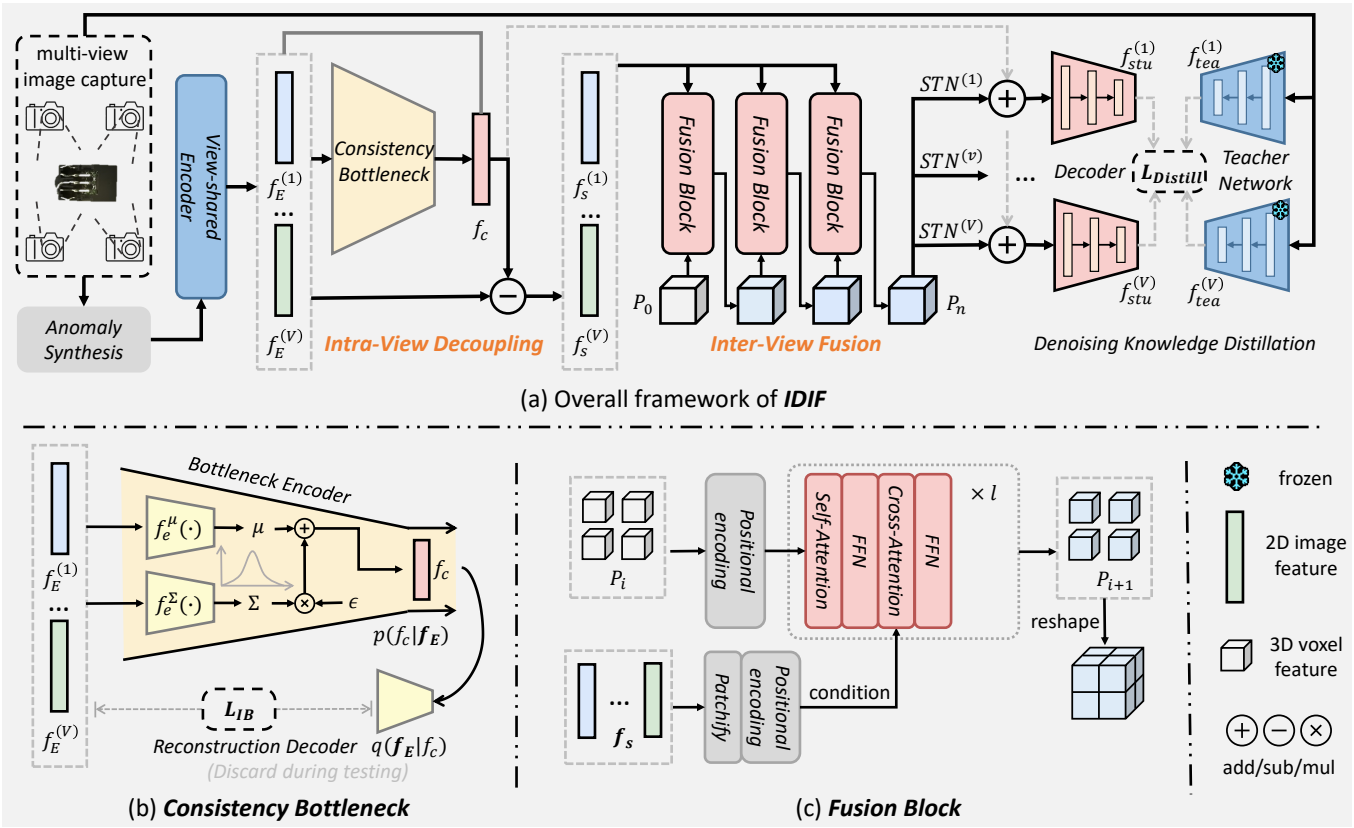

Figure 2: Overview of the proposed IDIF. (a) The overall framework of our model. The multi-view image features are separated into the view-common feature f_c and view-specific features $\boldsymbol{f}_s$ by *Consistency Bottleneck* (yellow). *Implicit Voxel Construction* (red) is utilized for tnter-view fusion. The denoising knowledge distillation is then performed to train the student network. (b) *Consistency Bottleneck*. (c) Fusion Block in *Implicit Voxel Construction*.

Distillation Architecture

Knowledge distillation is a mainstay framework in the field of anomaly detection. Following DeSTSeg (Zhang et al. 2023c), our model consists of a pretrained teacher network T, a denoising student network S, and a Segmentation Head generating the final segmentation results. The teacher network T is a frozen resnet (He et al. 2016) model pretrained on ImageNet (Deng et al. 2009). During training, the teacher network T is input with normal images from V views individually to get the corresponding normal feature $f_{tea} = \{f_{tea}^{(v)} | v = 1, ..., V]\}$, where $f_{tea}^{(v)} = T(x^{(v)})$. The student network S is input with multi-view pseudo anomalous images, and the output is denoted as $f_{stu} = \{f_{stu}^{(v)} | v = 1, ..., V]\}$, where $f_{stu}^{(v)} = S(pseudo(x^{(v)}))$. $pseudo(x^{(v)})$ represents the pseudo anomaly synthesis of $x^{(v)}$ following DeSTSeg (Zhang et al. 2023c). The student network learns to reconstruct normal features from $pseudo(x^{(v)})$ as Eq. (1).

$$L_{Distill} = \sum_{v=1}^{V} (1 - cossim(f_{tea}^{(v)}, f_{stu}^{(v)})), \qquad (1)$$

where $cossim(\cdot)$ represents the cosine similarity between the two vectors. The above process is called "denoising".

The denoising result $S(pseudo(x^{(v)}))$ plays an important role in the whole process. By exploring the correlation between images from multiple views, the denoising performance could be better. In student network S, we utilize the proposed *Consistency Bottleneck* for **Intra-view Decoupling**, extracting the view-common and view-specific features. The view-specific features are used to perform **Inter-view Fusion** by our proposed *Implicit Voxel Construction* module. Through filling a sample-aware 3D voxel and then mapping it back to 2D features with learnable angles, information between different views is fused together to obtain more comprehensive representations. The view-common feature is excluded from the fusion process and added as residuals to the fused result. Fig. 2(a) illustrates the overall framework of our model.

Intra-view Decoupling

Consistency Bottleneck is designed to separate the view-common feature and view-specific feature in the student network S explicitly, i.e. Intra-view Decoupling. First, all images are mapped to the feature space. For example, the v-th view image $x^{(v)}$ is input to a shared feature extractor F_E, acquiring the corresponding feature $f_E^{(v)} = F_E(x^{(v)})$.

$\boldsymbol{f}_E = \{f_E^{(v)} | v = 1, ..., V\}$ is the feature of all th

Our goal is to get the view-common feature f To this end, we design to maximize the mutual in between the view-common feature f_c and $\boldsymbol{f}_E$. f_c to be low-dimensional and therefore limited in i capacity. With this bottleneck design, a trivial sol to maximize the mutual information with $\boldsymbol{f}_E$ is t information of each single view. The optimizatic is shown by Eq. (2).

$$f_c = \arg\max_{f_c} \sum_{v=1}^{V} I(f_c, f_E^{(v)}) \Rightarrow \arg\max_{f_c} I(f_$$

$$\Rightarrow \arg\max_{f_c} \int df_c d\boldsymbol{f}_E \, p(f_c, \boldsymbol{f}_E) \log \frac{p(f_c,}{p(f_c)p($$

Optimizing the mutual information directly tationally intractable, we can turn to optimize its variational lower bounds to find an approximate solution (Alemi et al. 2017). Let $q(\boldsymbol{f}_E|f_c)$ be a variational approximation to $p(\boldsymbol{f}_E|f_c)$. From the nonnegativity of the Kullback Leibler divergence $KL[p(\boldsymbol{f}_E|f_c), q(\boldsymbol{f}_E|f_c)] \geq 0$, we have

$$\begin{aligned} I(f_c, \boldsymbol{f}_E) &\geq \int df_c d\boldsymbol{f}_E \, p(f_c, \boldsymbol{f}_E) \log \frac{q(\boldsymbol{f}_E|f_c)}{p(\boldsymbol{f}_E)} \\ &= \int df_c d\boldsymbol{f}_E \, p(f_c, \boldsymbol{f}_E) \log q(\boldsymbol{f}_E|f_c) + H(\boldsymbol{f}_E) \\ &\geq \int d\boldsymbol{f}_E \, p(\boldsymbol{f}_E) \int df_c \, p(f_c|\boldsymbol{f}_E) \log q(\boldsymbol{f}_E|f_c). \end{aligned} \tag{3}$$

$H(\boldsymbol{f}_E)$ refers to the entropy of $\boldsymbol{f}_E$. It is independent of the optimization procedure of f_c and can therefore be ignored. $p(\boldsymbol{f}_E)$ can be approximated using the empirical data distribution. Thus we have

$$I(f_c, \boldsymbol{f}_E) \geq \frac{1}{N} \sum_{n=1}^{N} \int df_c \, p(f_c|\boldsymbol{f}_E) \log q(\boldsymbol{f}_E|f_c). \tag{4}$$

Suppose we have a variational bottleneck encoder in VAE (Kingma and Welling 2013) with the form of $p(f_c|\boldsymbol{f}_E) = \mathcal{N}(f_c|f_e^{\mu}(\boldsymbol{f}_E), f_e^{\Sigma}(\boldsymbol{f}_E))$, where f_e is a network which outputs the mean $f_e^{\mu}(\boldsymbol{f}_E)$ and covariance matrix $f_e^{\Sigma}(\boldsymbol{f}_E)$. We can write $p(f_c|\boldsymbol{f}_E)df_c = p(\epsilon)d\epsilon$ using the reparameterization trick, where $f_c = F(\boldsymbol{f}_E, \epsilon)$ is a deterministic function of $\boldsymbol{f}_E$ and the Gaussian random variable ϵ. The $q(\boldsymbol{f}_E|f_c)$ can be seen as a reconstruction decoder. In this way, the optimization objective can be further written in the following loss form.

$$L_{IB} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{\epsilon \sim p(\epsilon)} [-\log q(\boldsymbol{f}_E|F(\boldsymbol{f}_E, \epsilon))] \tag{5}$$

Through back propagation, the bottleneck encoder learns how to extract view-common features from the input, as in Fig. 2(b). Each element $f_s^{(v)}$ in view-specific features $\boldsymbol{f}_s = \{f_s^{(v)} | v = 1, ..., V\}$ is considered to be the difference between the original feature $f_E^{(v)}$ and the view-common feature f_c:

$$f_s^{(v)} = f_E^{(v)} - f_c, \; v = 1, ..., V \tag{6}$$

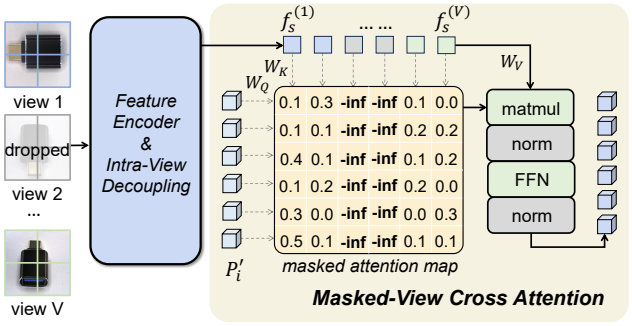

Figure 3: Illustration of our View-wise Dropout strategy and Masked-View Cross Attention. The masked attention area will be equal to 0 after the softmax operation.

Inter-view Fusion

In order to perform inter-view fusion, we design to construct 3D voxels implicitly in the feature space. The form of 3D voxels is the most natural way to fuse features from different views. It can better utilize the spatial characteristics of features from different views without losing information.

To do so, we design a learnable sample-level 3D voxel prototype P_0 and several fusion blocks. The prototype learns to provide sample-independent prior information for voxel construction. The fusion blocks consist of self-attention layers (SA), cross-attention layers (CA), and feed-forward networks (FFN). During the construction process, view-specific features $\boldsymbol{f}_s$ participate as conditions on the construction of sample-relative 3D voxels P_n from prototype P_0 through fusion blocks, as illustrated in Fig. 2(c). This process is described as:

$$\begin{aligned} P_i^{'} &= FFN(SA(P_i)), \\ P_{i+1} &= FFN(CA(P_i^{'}, \boldsymbol{f}_s)), \; i = 0, ..., n-1. \end{aligned} \tag{7}$$

To align with the output of the teacher network T and perform 2D anomaly segmentation in the v-th view, we instantiate a $STN^{(v)}$ module (Jaderberg et al. 2015) to map P_n to the 2D plane with a learnable angle $\alpha^{(v)}$,

$$f_p^{(v)} = STN^{(v)}(P_n, \alpha^{(v)}), \; v = 1, ..., V. \tag{8}$$

The feature $f_p^{(v)}$, along with the view-common feature f_c, is fed into a view-specific decoder $Decoder^{(v)}$,

$$f_{stu}^{(v)} = Decoder^{(v)}(f_p^{(v)} + f_c), \; v = 1, ..., V. \tag{9}$$

The output $f_{stu}^{(v)}$ is supervised by the output of the teacher network $f_{tea}^{(v)}$ as in Eq.(1).

The total loss to train the model is the weighted sum of $L_{Distill}$ and L_{IB},

$$L_{Total} = \alpha \cdot L_{Distill} + \beta \cdot L_{IB} \tag{10}$$

where α and β in Eq.(10) are scale factors.

By optimizing L_{Total}, the student network S can learn to denoise multi-view images through our designed two explicit steps: intra-view decoupling and inter-view fusion. The well-designed denoising reconstruction would ensure the unsupervised anomaly detection performance.

Category	PaDim	PatchCore	RD	UniAD	DeSTSeg	SimpleNet	**Ours**
Audiojack	92.2 / 94.2	89.3 / **98.4**	81.9 / 97.9	91.2 / 97.2	95.3 / 97.3	91.2 / 98.2	**96.1 / 98.4**
Bottle Cap	98.1 / 97.7	**99.4** / 99.3	93.7 / 99.0	97.3 / 99.2	92.4 / 99.4	**99.4** / 98.5	98.2 / **99.7**
Button Battery	88.7 / 90.4	90.6 / **98.8**	83.3 / 98.3	87.5 / 93.7	93.3 / 98.7	95.8 / 98.0	**97.7 / 98.8**
End Cap	76.1 / 95.2	91.9 / **98.0**	68.1 / 97.3	89.4 / 96.7	82.3 / 95.4	**94.2** / 94.2	94.1 / 97.7
Eraser	96.5 / 95.7	95.6 / 98.7	82.9 / 98.6	91.2 / 99.0	91.9 / 99.4	94.7 / 98.3	**96.6 / 99.5**
Fire Hood	**96.9** / 94.5	89.3 / 98.7	81.4 / 98.3	83.0 / 98.5	**96.9** / 98.6	95.6 / 97.5	92.6 / **99.2**
Mint	69.1 / 90.3	85.7 / **98.8**	67.7 / 97.7	73.0 / 94.3	77.7 / 93.9	86.8 / 94.1	**87.8** / 97.9
Mounts	98.4 / 97.5	**99.7** / 98.3	92.5 / 98.3	97.0 / 99.4	99.1 / 99.3	99.4 / 98.0	99.3 / **99.5**
PCB	88.4 / 91.0	93.0 / 99.1	79.3 / 98.8	83.2 / 96.6	83.6 / 98.5	90.7 / 98.4	**94.2 / 99.3**
Phone Battery	91.7 / 89.6	95.1 / 98.9	89.4 / 98.9	93.6 / 97.9	98.2 / 96.6	94.7 / 96.5	**98.5 / 99.3**
Plastic Nut	**98.2** / 95.1	97.8 / 98.8	72.8 / 98.9	87.1 / 98.6	94.4 / 98.2	95.7 / 98.1	96.9 / **99.5**
Plastic Plug	87.4 / 93.5	95.7 / 98.5	89.3 / 98.3	78.0 / 97.9	95.6 / 95.1	94.4 / 96.1	**97.6 / 99.1**
Porcelain Doll	93.8 / 91.7	96.1 / 98.0	89.6 / 98.2	92.8 / 97.3	94.6 / 97.5	96.2 / 96.6	**98.6 / 98.4**
Regulator	96.5 / 91.9	86.0 / 99.2	92.5 / **98.7**	55.5 / 93.7	93.0 / **98.7**	92.0 / 97.0	**99.8 / 98.7**
Rolled Strip Base	98.6 / 92.3	**99.7** / 99.1	80.3 / 99.0	99.3 / 98.9	98.9 / 99.4	99.6 / 98.8	**99.7 / 99.8**
SIM Card Set	94.2 / 85.4	99.3 / 99.0	89.9 / 97.7	94.0 / 96.7	98.3 / 97.6	99.2 / 97.3	**99.6 / 99.3**
Switch	82.1 / 97.3	94.6 / 98.5	87.3 / 98.6	95.3 / 99.4	96.6 / 99.5	**98.8** / 99.1	98.3 / **99.7**
Tape	99.8 / 97.9	99.9 / 99.1	89.5 / 99.0	99.1 / 99.5	99.1 / 99.6	**100.0** / 99.2	98.7 / **99.8**
Terminal Block	96.9 / 96.7	97.5 / 99.2	89.8 / 99.0	93.8 / 98.9	96.1 / 99.7	97.7 / 99.3	**97.9 / 99.8**
Toothbrush	91.7 / 87.2	94.7 / 96.2	86.7 / 96.3	95.0 / 96.8	97.9 / 92.1	95.3 / 94.3	**99.0 / 97.3**
Toy	91.4 / 83.3	92.8 / **98.3**	75.0 / 95.2	77.2 / 96.4	96.5 / 91.4	92.9 / 91.9	**97.8** / 96.6
Toy Brick	84.3 / 94.1	82.6 / 97.5	72.5 / 96.3	78.3 / 97.9	87.0 / 96.2	85.7 / 94.3	**92.8 / 98.7**
Transistor1	90.3 / 95.4	**99.8** / 98.9	94.7 / 98.8	99.3 / 98.8	99.0 / 98.5	99.7 / 99.1	**99.8 / 99.4**
U Block	98.3 / 96.3	**98.8** / 98.9	86.9 / 98.4	96.3 / 99.0	98.5 / 99.5	98.5 / 98.6	98.5 / **99.6**
USB	77.0 / 93.6	93.9 / 99.1	89.4 / 98.9	83.1 / 98.5	93.3 / 97.3	93.9 / 98.9	**98.4 / 99.6**
USB Adaptor	93.2 / 93.0	90.6 / 98.2	65.3 / 96.5	85.1 / 97.0	93.6 / 96.8	93.0 / 95.7	**95.1 / 98.6**
Vcpill	94.7 / 93.4	96.5 / 98.3	87.2 / 97.7	89.4 / **99.1**	96.4 / 98.1	97.5 / 98.6	**99.6** / 98.9
Wooden Beads	91.1 / 90.5	91.4 / 98.1	85.0 / 97.4	82.5 / 97.5	91.9 / **98.6**	92.9 / 96.7	**94.5 / 98.6**
Woodstick	81.8 / 93.7	74.5 / 97.3	71.9 / 98.0	76.0 / 96.6	90.2 / 98.1	81.5 / 93.5	**91.1 / 98.6**
Zipper	99.3 / 90.5	**100.0** / 98.3	96.1 / 98.6	98.8 / 97.5	99.7 / 91.2	99.7 / 98.6	**100.0 / 98.7**
Average All	91.2 / 93.0	93.7 / 98.5	83.7 / 98.1	88.1 / 97.6	94.0 / 97.3	94.9 / 96.8	**97.0 / 98.9**

Table 1: Comparison of IDIF with previous methods on Real-IAD. The results are presented in the form of S-AUROC(%)/P-AUROC(%), where S-AUROC is used as an evaluation metric for anomaly detection and P-AUROC for localization.

View-wise Dropout

In real-world deployments with missing views, a multi-view model's detection capability is likely to fail due to inconsistent inputs. We propose a *View-wise Dropout* training strategy to address the issue of missing views, as illustrated in Fig. 3. During training, some views of input images in $\boldsymbol{x}$ are randomly dropped. The dropped views are then excluded from the intra-view decoupling process. To mitigate the impact of dropped views on the fusion process, we further design a Masked-View Cross Attention mechanism (MVCA) in the *IVC* module. Let $\widehat{Q} = P_i^{'} W_Q$, $\widehat{K} = Concat(f_s^{(1)}, ..., f_s^{(V)})W_K$, and $\widehat{V} = Concat(f_s^{(1)}, ..., f_s^{(V)})W_V$ denote the querys of voxels, keys and values of image patches at each layer in the fusion blocks. W_Q, W_K and W_V represent the projection matrices in attention mechanism (Vaswani 2017). The portions of the attention map corresponding to those dropped views are masked with negative infinity, as described in Eq.(11),

$$MVCA(\widehat{Q}, \widehat{K}, \widehat{V}) = softmax(M + \frac{\widehat{Q}\widehat{K}^T}{\sqrt{d}})\widehat{V}, \quad (11)$$

where M denotes the attention mask. d is the feature dimension. The value of attention mask M at location (i, j) is defined as:

$$M(i,j) = \begin{cases} \text{-inf} & \text{, if view of patch } j \text{ is dropped} \\ 0 & \text{, otherwise} \end{cases} \quad (12)$$

When view-wise dropout is performed, the original cross attention in the fusion block is replaced with the proposed masked-view cross attention.

Experiments

Experimental Details

Datasets. Real-IAD is a large-scale, multi-view anomaly detection dataset (Wang et al. 2024). For every sample in the dataset, images from five different views are captured. It contains 150K images spanning 30 categories. For its multi-view characteristics, we use this dataset as the primary benchmark to evaluate our method.

MvTec 3D-AD dataset (Bergmann et al. 2019) and Eyecandies dataset (Bonfiglioli et al. 2022) only provide 3D format data. Thus we process them to obtain images in multi-view 2D format.

MvTec 3D-AD dataset consists of 4147 scans from 10 real-world object categories. Each scan provides the point clouds with RGB. We capture RGB images from five different views for each scan, based on the original 3D point

Method	MvTec 3D-AD Dataset										
	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
PatchCore	91.8	74.8	96.7	88.3	93.2	58.2	89.6	91.2	92.1	88.6	86.5
AST	98.3	87.3	97.6	97.1	93.2	88.5	97.4	98.1	**100.0**	79.7	93.7
M3DM	99.4	90.9	97.2	97.6	96.0	**94.2**	97.3	89.9	97.2	85.0	94.5
MMRD	**99.9**	94.3	96.4	94.3	**99.2**	91.2	94.9	90.1	99.4	90.1	95.0
MCFM	99.4	88.8	**98.4**	99.3	98.0	88.8	94.1	94.3	98.0	**95.3**	95.4
Ours	98.2	**96.1**	89.0	**99.9**	97.1	90.8	**99.5**	**98.8**	99.7	86.5	**95.6**

Method	Eyecandies Dataset										
	Candy Cane	Chocolate Cookie	Chocolate Praline	Confetto	Gummy Bear	Hazelnut Truffle	Licorice Sandwich	Lollipop	Marsh	Pepperming Candy	Mean
PatchCore	44.8	95.0	77.9	92.8	88.8	41.6	91.2	83.1	**100.0**	96.3	81.1
AST	58.7	84.6	80.7	83.3	83.3	54.3	74.4	87.0	94.6	83.5	78.4
M3DM	62.4	95.8	**95.8**	**100.0**	88.6	**78.5**	94.6	83.6	**100.0**	**100.0**	89.7
MCFM	68.0	93.1	95.2	88.0	86.5	78.2	91.7	84.0	99.8	96.2	88.1
MMRD	85.4	**100.0**	94.6	99.8	90.8	74.7	**96.6**	**98.4**	**100.0**	**100.0**	94.0
Ours	**88.6**	99.4	94.1	99.2	**94.2**	77.8	91.2	98.0	**100.0**	99.2	**94.2**

Table 2: Comparison of IDIF with previous methods on MVTec 3D-AD and Eyecandies using S-AUROC(%).

DKD	*3DConv*	*IVC*	*CB*	Sample AUROC
✓	-	-	-	94.0
✓	✓	-	-	95.8
✓	-	✓	-	96.1
✓	-	✓	✓	**97.0**

Table 3: Results of ablation experiments. *DKD* refers to ensemble of DeSTSeg models. *3DConv* refers to fusion based on 3D-conv. *IVC* and *CB* refer to the usage of *Implicit Voxel Construction* and *Consistency Bottleneck*.

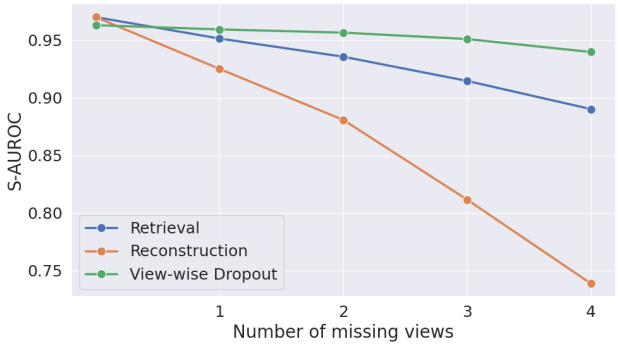

Figure 4: Comparison of different methods for handling missing views. Comparisons are conducted with the test set under the same missing cases for fairness.

cloud. Besides the RGB images, we also capture depth maps from the same five views.

Eyecandies dataset consists 1,500 scans of 10 object categories. Each scan provides an RGB image, a depth map, and a surface normal. We also capture images from five different views. The surface normals, instead of the depth maps, are obtained from the same five views.

Evaluation Metrics. We use the Area Under the Receiver Operating Characteristic Curve of Sample-level (S-AUROC) to evaluate the performance of anomaly detection. A sample is considered normal only when images from all views are normal. The Pixel AUROC (P-AUROC) is used to evaluate the performance of anomaly localization.

Implementation details. We utilize ResNet18 (He et al. 2016) pretrained on ImageNet (Deng et al. 2009) as the teacher network. The outputs from block1 (64×64), block2 (32×32), and block3 (16×16) of the teacher network are used to guide the training of the student network. Images are resized to 256×256 for training and testing. The batch size is 8. We use the Describable Textures Dataset (DTD) (Cimpoi et al. 2014) as the source of pseudo anomalies.

Comparison with Other Methods

Results on Real-IAD. Table 1 shows the results for anomaly detection and localization on the Real-IAD dataset. We compare our method with previous SOTA anomaly detection methods, including PaDim (Defard et al. 2021), PatchCore (Roth et al. 2022), RD (Deng and Li 2022), UniAD (You et al. 2022), DeSTSeg (Zhang et al. 2023c), and SimpleNet (Liu et al. 2023).

Our method achieves state-of-the-art performance for anomaly detection with **97.0%** S-AUROC, outperforming the second-best approach by **2.1%**. IDIF achieves the best performance in 22 categories out of all 30 categories. The performance for anomaly localization is also competitive, with **98.9%** for P-AUROC. The model is able to better reconstruct the anomalous regions with fused features, thus showing a larger gap with the teacher network output and achieving better detection results. We also visualize several results from various categories, as shown in Fig. 5. As can be seen, our method accurately locates anomalous regions, even those with very small anomalies.

Results on MvTec 3D-AD and Eyecandies. We also evaluate our method on the MvTec 3D-AD and Eyecandies datasets, as shown in Table 2. We compare our method with PatchCore (Roth et al. 2022), AST (Rudolph et al. 2023), M3DM (Wang et al. 2023), MMRD (Gu et al. 2024), and

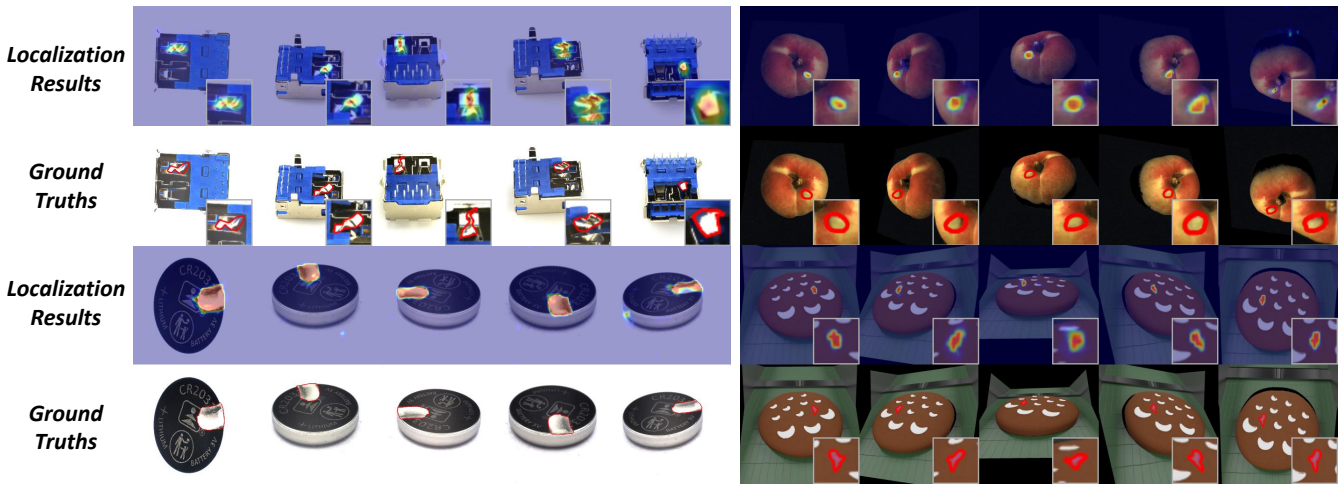

Figure 5: Visualization of anomaly localization results on Real-IAD (*Left*), MvTec 3D-AD and Eyecandies (*Right*).

MCFM (Costanzino et al. 2024). Since the input modalities are different, we only evaluate the detection performance (S-AUROC) for fair comparison. Our method outperforms the best 3D-based and multi-modal methods, achieving a S-AUROC of **95.6%** on MvTec 3D-AD and **94.2%** on Eyecandies. Fig. 5 shows the visualized results of our method. IDIF can achieve accurate detection of anomalies across different views.

Ablation Studies

Ablation experiments are performed mainly on important components of the model. In Table 3, *DKD* (denoising knowledge distillation) refers to training the DeSTSeg (Zhang et al. 2023c) model for each view individually and performing ensemble model. Sample-level predictions are obtained by taking the maximum value among the image-level results predicted by each model. *3DConv* refers to concatenating the $\boldsymbol{f}_E$ in a new dimension and then fusing the information using a 3D-conv-based fusion module (Xie et al. 2019). *IVC* refers to the information fusion using the proposed *Implicit Voxel Construction* module. Through our experiments, we found that using a single model for multi-view anomaly detection significantly improves the detection performance (from 94.0% to at least 95.8%) compared to ensemble model. And *IVC* is able to improve the metric to 96.1%, which we attribute to its explicit spatial structure construction.

In addition, adding *CB* block by first performing Intra-view Decoupling before Inter-view Fusing, further improves the metric to 97.0%. We believe that explicitly separating out and excluding information shared between views from fusion can make the fusion process easier to perform.

Missing Views During Test

We evaluate the performance of our *View-wise Dropout* strategy in addressing the challenge of missing views on the Real-IAD dataset. It achieves 95.9%, 95.7%, 95.1%, 94.0% S-AUROC with one, two, three and four views randomly missing respectively. The average performance degradation per missing view is only 0.57% S-AUROC.

We compare our *View-wise Dropout* strategy with two other methods of addressing the missing views problem. One way is to retrieve the similar samples from the train set to replace the missing views. Another way is to reconstruct the missing views based on the remaining captured images. Results of different ways to deal with missing views are shown in Fig. 4. View-wise Dropout exhibits the best ability to maintain performance as the missing number increases. The average performance degradation per missing view is 2.00% for Retrieve and 5.77% for Reconstruction, much higher than our proposed strategy.

Since dropout affects the whole model weights, our View-wise Dropout strategy introduces a little full-view performance degradation (from 97.0% to 96.4%). However, it is able to achieve good performance preservation with only a few training adjustments. With the dropout probability of each view set to 0.2 during training, the expected number of times for each combination of four missing views is less than twice in 1000 steps of training, while the strategy helps preserving 94.0% S-AUROC performance.

Conclusion

In this work, we propose IDIF, a novel paradigm for multi-view anomaly detection. IDIF consists of *Implicit Voxel Construction* and *Consistency Bottleneck* which perform Intra-view Decoupling and Inter-view Fusion, respectively. This approach allows for more effective reconstruction of anomalous features, and thus a better detection performance. Additionally, we highlight the issue of missing views in multi-view scenarios and provide the *View-wise Dropout* strategy to address this challenge. Experiments on three multi-view datasets demonstrate the effectiveness of our method by achieving state-of-the-art performance.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. U23B2060, No.62088102), and the Youth Innovation Team of Shaanxi Universities.

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *International Conference on Learning Representations*.

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*, 1247–1255.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Conference on Computer Vision and Pattern Recognition*, 9592–9600.

Bonfiglioli, L.; Toschi, M.; Silvestri, D.; Fioraio, N.; and De Gregorio, D. 2022. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *the Asian Conference on Computer Vision*, 3586–3602.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Conference on Computer Vision and Pattern Recognition*, 3606–3613.

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

Costanzino, A.; Ramirez, P. Z.; Lisanti, G.; and Di Stefano, L. 2024. Multimodal Industrial Anomaly Detection by Crossmodal Feature Mapping. In *Conference on Computer Vision and Pattern Recognition*, 17234–17243.

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

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

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 248–255.

Gu, Z.; Zhang, J.; Liu, L.; Chen, X.; Peng, J.; Gan, Z.; Jiang, G.; Shu, A.; Wang, Y.; and Ma, L. 2024. Rethinking Reverse Distillation for Multi-Modal Anomaly Detection. In *the AAAI Conference on Artificial Intelligence*, volume 38, 8445–8453.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 770–778.

Hotelling, H. 1992. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, 162–190.

Hwang, H.; Kim, G.-H.; Hong, S.; and Kim, K.-E. 2021. Multi-view representation learning via total correlation objective. *Advances in Neural Information Processing Systems*, 34: 12194–12207.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4): 139–1.

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

Liu, L.; Aviles-Rivero, A. I.; and Schönlieb, C.-B. 2023. Contrastive registration for unsupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

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

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Raza, K.; and Singh, N. K. 2021. A tour of unsupervised deep learning for medical image analysis. *Current Medical Imaging*, 17(9): 1059–1077.

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

Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; and Wandt, B. 2023. Asymmetric student-teacher networks for industrial anomaly detection. In *Winter Conference on Applications of Computer Vision*, 2592–2602.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, C.; Zhu, W.; Gao, B.-B.; Gan, Z.; Zhang, J.; Gu, Z.; Qian, S.; Chen, M.; and Ma, L. 2024. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Conference on Computer Vision and Pattern Recognition*, 22883–22892.

Wang, D.; Cui, X.; Chen, X.; Zou, Z.; Shi, T.; Salcudean, S.; Wang, Z. J.; and Ward, R. 2021. Multi-view 3d reconstruction with transformers. In *International Conference on Computer Vision*, 5722–5731.

Wang, Y.; Peng, J.; Zhang, J.; Yi, R.; Wang, Y.; and Wang, C. 2023. Multimodal industrial anomaly detection via hybrid fusion. In *Conference on Computer Vision and Pattern Recognition*, 8032–8041.

Xie, H.; Yao, H.; Sun, X.; Zhou, S.; and Zhang, S. 2019. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *International Conference on Computer Vision*, 2690–2698.

Yagubbayli, F.; Wang, Y.; Tonioni, A.; and Tombari, F. 2021. Legoformer: Transformers for block-by-block multi-view 3d reconstruction. *arXiv preprint arXiv:2106.12102*.

Yang, B.; Wang, S.; Markham, A.; and Trigoni, N. 2020. Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction. *International Journal of Computer Vision*, 128(1): 53–73.

You, Z.; Cui, L.; Shen, Y.; Yang, K.; Lu, X.; Zheng, Y.; and Le, X. 2022. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35: 4571–4584.

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

Zhang, J.; Wu, X.; Cheng, Z.-Q.; He, Q.; and Li, W. 2023a. Improving anomaly segmentation with multi-granularity cross-domain alignment. In *the 31st ACM International Conference on Multimedia*, 8515–8524.

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

Zhang, X.; Li, S.; Li, X.; Huang, P.; Shan, J.; and Chen, T. 2023c. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Conference on Computer Vision and Pattern Recognition*, 3914–3923.

Zhang, X.; Xu, M.; and Zhou, X. 2024. RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection. In *Conference on Computer Vision and Pattern Recognition*, 16699–16708.

Zhao, H.; Ding, Z.; and Fu, Y. 2017. Multi-view clustering via deep matrix factorization. In *the AAAI conference on artificial intelligence*, volume 31.