Learning Diffusion Models for Multi-View Anomaly Detection

Chieh Liu[1], Yu-Min Chu[1], Ting-I Hsieh[1], and Hwann-Tzong Chen[1,2]
Tyng-Luh Liu[3]

[1] National Tsing Hua University, Taiwan
[2] Aeolus Robotics, Taiwan
[3] Institute of Information Science, Academia Sinica, Taiwan

Abstract. We are exploring an emerging formulation in anomaly detection (AD) where multiple instances of the same object are produced simultaneously and distinctly to address the limitation that using only a single instance may not effectively capture any underlying defects. More specifically, we concentrate on a specific scenario where each object of interest is linked to seven distinct data views/representations. The first six views involve capturing images with a stationary camera under six different lighting conditions, while the seventh view pertains to the 3D normal information. We refer to our intended task as *multi-view anomaly detection*. To tackle this problem, our approach involves training a view-invariant ControlNet that can produce consistent feature maps regardless of the data views. This training strategy enables us to mitigate the impact of varying lighting conditions and to fuse information from both the RGB color appearance and the 3D normal geometry effectively. Moreover, as the diffusion process is not deterministic, we utilize the denoising diffusion implicit model (DDIM) [38] scheme to improve the applicability of our established memory banks of diffusion-based features for anomaly detection inference. To demonstrate the efficacy of our approach, we present extensive ablation studies and state-of-the-art experimental results on the Eyecandies dataset.

Keywords: Anomaly detection · Diffusion model · ControlNet

1 Introduction

We explore a multi-view setting of anomaly detection (AD), where each *view* of the underlying data corresponds to a specific type of feature representation or a specific way of data acquisition. For example, consider a challenging but practical scenario of anomaly detection where the objects of interest have a shiny metal surface. In such a case, taking pictures from a fixed camera viewing angle may not consistently capture the defect spots due to light reflection. To address this issue, it is feasible to enrich the *views* of each sample by taking pictures either from several fixed camera positions or under different lighting conditions. In this study, we leverage the availability of the Eyecandies dataset [7] to assess our proposed approach to multi-view anomaly detection.

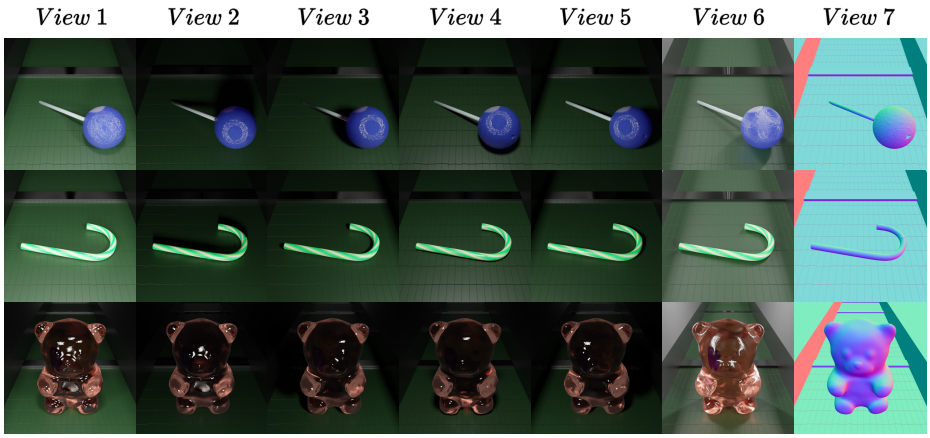

Fig. 1: Examples from Eyecandies [7]. Each row depicts that an object of interest is presented in seven different *views*, including lighting conditions and 3D normals.

Multiple Data Views. Fig. 1 illustrates our targeted multi-view scenario from the Eyecandies dataset [7]. Views 1 to 6 showcase the same object with different 2D-RGB representations, caused by varying the lighting conditions, while view 7 presents its corresponding 3D-Normal representation. This dataset is informative by revealing that anomalies are not consistently observable across all data views; rather, their recognition is contingent on specific views. Therefore, rather than relying solely on a single view, this dataset necessitates the simultaneous consideration of multiple views to detect anomalies. Furthermore, we observe that certain anomalies are particularly conspicuous in specific single-modality views, either 2D-RGB or 3D-Normal, as illustrated in Fig. 2. For example, the discoloration on the `PeppermintCandy` is prominently visible in 2D-RGB views, while the 3D-Normal view fails to capture this purely 2D anomaly. In contrast, the 3D structure anomaly in `HazelnutTruffle` is difficult to identify through the six RGB views, as the chopped hazelnuts on the surface camouflage the anomalies. This anomaly can be more easily discerned from the 3D-Normal map view. These findings highlight the importance of integrating cross-modality views for effective anomaly detection in the Eyecandies dataset.

Multi-view Anomaly Detection. The PAD dataset [52] focuses on multi-pose anomaly detection using LEGO toys as its theme. In contrast to casting the multi-view scenario with different lighting conditions and the 3D normals as in Eyecandies, this dataset offers visual cues from various pose angles for pose-sensitive anomaly detection. However, in industrial inspection, RGB information alone is insufficient to detect 3D structural defects. Another dataset, Real-3D [25], uses a 360-degree view of pure point clouds for anomaly detection. Yet, the cost of scanning 360-degree point clouds is high and thus impractical for applications. Comprehensive multi-view anomaly detection approaches are still lacking.

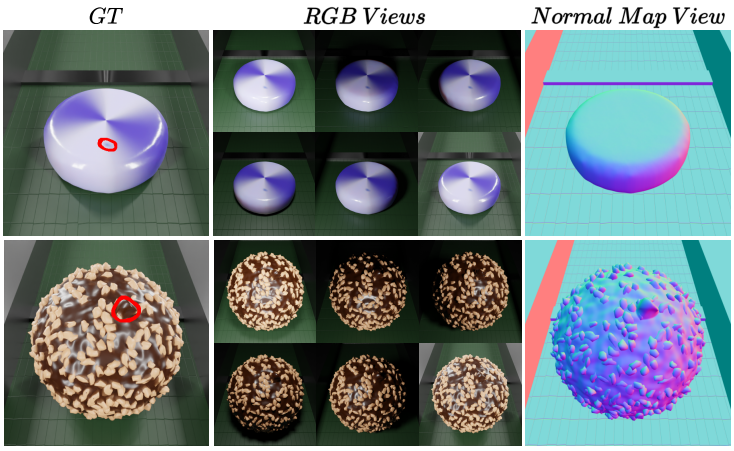

Fig. 2: Complementary AD cues between multi-modality views. Top: The defect ((circled in red) is discernible in 2D-RGB views but not in the 3D-Normal view. Bottom: The other defect is evident from the 3D-Normal view. Fusing the information from the 7 views of the two modalities could achieve complementary effects to the AD task.

Motivations and Contributions. Anomaly detection datasets are structured in a one-class format, where all samples in the training set are assumed to be *normal*. Given the impracticality of modeling all potential forms of defects in each specific application, it becomes valuable to enhance the diversity of data views of the normal samples and design techniques to integrate complementary information from their resulting feature representations. We focus on a popular multi-view dataset, Eyescandies [7], to tackle anomaly detection by exploring seven data views of distinct lighting conditions and the 3D normals. Our method for addressing the task of multi-view anomaly detection is founded on new learning strategies for establishing a view-invariant ControlNet as well as building memory banks of diffusion-based features in a determistic manner. Our experimental results demonstrate the effectiveness of the proposed method.

2 Related Work

Anomaly detection has been moving beyond traditional RGB datasets [3] to the rich detail of 3D point clouds [5, 25]. The Eyecandies dataset [7] simulates factory environments and multi-view scenarios to introduce a more challenging dimension to anomaly detection.

A common approach (*e.g.*, [46]) involves training an autoencoder on normal samples. The anomalies are then identified by analyzing the differences between the input and the reconstructed output of the autoencoder. A similar strategy [2, 4,50] uses a stronger Teacher Network and a weaker Student Network that differ in their reconstruction capabilities. This approach thus evaluates the reconstruction disparities between the Teacher and Student networks to identify the anomalies.

Furthermore, many 3D reconstruction methods [6, 35] favor the Student-Teacher (S-T) Network approach to avoid the difficulty of direct reconstruction of point clouds but face limitations when working with multi-modality datasets.

Synthetic-based methods are among the most intuitive approaches for anomaly detection. For instance, DRAEM [45] generates anomalies via noise sampling, and CutPaste [24] introduces anomalies by augmenting other data and overlaying it onto the original dataset. Subsequent methods [1, 27, 36] follow a similar idea to enhance anomaly detection. The latest methods [8, 44] involve generating pseudo-depth defects in the depth map. Despite these advancements, synthetic-based methods still fall short of fully capturing the distribution of anomalies.

Embedding-based methods [10,12,23,33,34] are also popular. They extract and store the anomaly-free features in a memory bank during training; while testing, the potentially defected features are compared with anomaly-free features in the memory bank to identify anomalies. In multi-modality datasets, the previously state-of-the-art methods, such as BTF [18], M3DM [41], and Shape-Guided [9], all use this technique to store and analyze 3D and 2D features.

2.1 Diffusion Models

Diffusion models generate high-quality photorealistic or stylish images through guided diffusion processes [13, 17, 32]. Current models allow for content control using specified conditions such as edge, depth [48], normal [22], or even manipulation through text prompts [21, 30]. Besides the generative capability, diffusion models can also be used to detect out-of-distribution instances [26, 42].

Diffusion-based anomaly detection methods such as [47, 49] first synthesize anomalous data and then add random Gaussian noise to obscure the distinction between anomalous and normal pixels, which allows the diffusion model to easily restore the image back to a normal state. Finally, a discriminator is trained to localize the anomalies. Wyatt *et al.* [43] employ simplex noise to corrupt data and use the diffusion model for data recovery. Anomalous regions can be identified by comparing the differences between recovered and original data. Mousakhan *et al.* [31] use the input image as a condition during the repair process for preserving normal regions. Lu *et al.* [28] incorporate different noise scales into anomaly samples and compute the anomaly score using the KL-divergence derived from the diffusion model. He *et al.* [15] design a semantic-guided network that preserves the semantics of the original input image while restoring anomalous regions. Hu *et al.* [19] use learnable Spatial Anomaly Embedding to guide the diffusion model to generate diverse anomalous images from limited training data.

Latent diffusion models [32] trained on large-scale datasets achieve impressive performance. Studies also confirm that these models yield meaningful internal representations. For example, the method of Tumanyan *et al.* [40] employs intermediate-layer features from a diffusion model to achieve fine-grained control over the image generation process. Hudson *et al.* [20] modulate diffusion decoder layers to enhance latent space disentanglement, which provides a compact representation for various downstream tasks. Tang *et al.* [39] show that pretrained stable diffusion models possess implicit knowledge within the latent

features that enable semantic correspondences. Luo *et al*. [29] use an aggregator network to integrate raw diffusion features across time and space into meaningful hyper-features. Building upon diffusion models' ability to learn meaningful representations, we propose a novel anomaly detection approach that leverages the rich internal representations developed by diffusion models as embeddings for anomaly detection.

3 Method

3.1 Problem Setting and Notations

Given a multi-view training set $D = \{X^1, X^2, \ldots, X^m\}$, where m is number of data views and $X^i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_N^i\}$ is the data matrix with N samples of the ith view. The corresponding label matrix is denoted $Y = \{y_1, y_2, \ldots, y_N\}$ where $y_j \in \{1, 2, \ldots, C\}$ and C is the total number of classes. Since all of our experiments on multi-view anomaly detection are done with the Eyecandies dataset [7], it is constructive to assume it as the default dataset, and thus we have $m = 7$ and $C = 10$. Specifically, each of the first six views corresponds to the RGB-based representation under a specific lighting condition, and the seventh view represents data as 3D normal maps. Similarly to solving the conventional task of anomaly detection, the training set D does not include any *abnormal* samples across all the m data views.

3.2 View-agnostic Latent Diffusion Model

Let U_θ be the denoising UNet of the pretrained latent diffusion model [32] and E be the VAE encoder used to yield the latent features. To adapt the pretrained model to tackle the task of multi-view anomaly detection, we fine-tune U_θ with respect to the multi-view training set D as follows. For any $\mathbf{x} \in D$, we have

$$\mathbf{x} \overset{E}{\longmapsto} \mathbf{z} \overset{\{t, \boldsymbol{\epsilon}\}}{\longmapsto} \mathbf{z}_t \overset{U_\epsilon}{\longmapsto} \boldsymbol{\epsilon}_\theta, \tag{1}$$

where $\mathbf{z} = E(\mathbf{x})$ is the latent representation of the input $\mathbf{x}$, and $\mathbf{z}_t$ is the noisy version after applying $\boldsymbol{\epsilon}$-forward for t steps. In our implementation, we fine-tune the parameters of U_θ for ten epochs by minimizing $\|\boldsymbol{\epsilon}_\theta - \epsilon\|$.

We purposely fine-tune the latent diffusion model, namely, U_θ, without differentiating between the various data views of input samples. This approach allows our technique to transform a pretrained diffusion model into a flexible one that can handle inputs from all the data views, whether they are RGB images under different lighting conditions or 3D normal maps.

3.3 Feature Representation via View-invariant ControlNet

Although the resulting model from (1) can already be used to achieve anomaly detection, its fine-tuning step does not take account of the fact that the $m = 7$ data

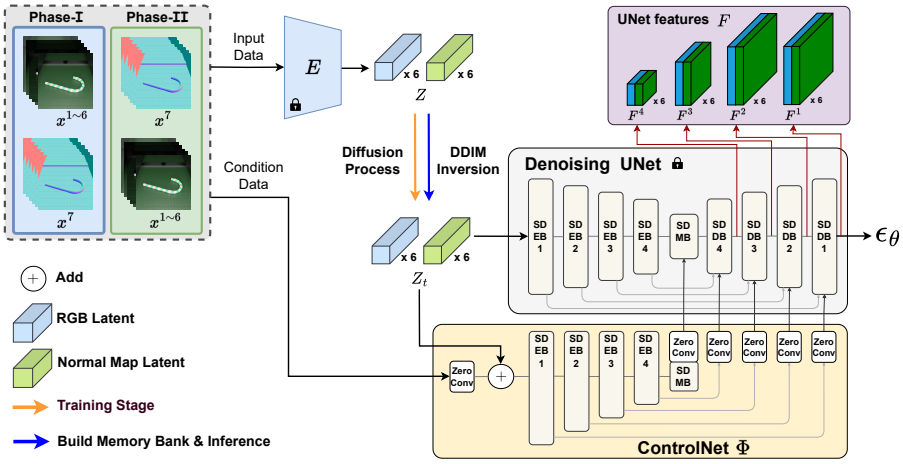

Fig. 3: Our proposed architecture leverages a *view-invariant* ControlNet and a pre-trained latent diffusion model to achieve the task of anomaly detection. In Phase-I and Phase-II, we utilize six RGB views, denoted as $\mathbf{x}^{1\sim6}$, and a 3D normal map $\mathbf{x}^7$ (replicated six times) as input data. These inputs are processed through the VAE encoder E to obtain latent representations Z, which include both RGB and normal map latents. Following the diffusion process, we generate the noise version of the latent representation. Subsequently, we employ complementary information ($\mathbf{x}^7$ for Phase-I and $\mathbf{x}^{1\sim6}$ for Phase-II) as condition data for the ControlNet. This allows the ControlNet to learn the view-invariant property based on the pretrained denoising UNet, enabling the extraction of UNet features F for both modalities, which are composed of F^1 to F^4 derived from each block of the stable diffusion decoder within the UNet.

views: $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^7\}$ are indeed associated with the same underlying sample $\mathbf{x}$. Furthermore, fusing information from different views is often advantageous for defect detection, as anomalies in this scenario may manifest as complementary irregularities in RGB color appearance or 3D surface structure.

To make the latent diffusion model behave similarly to the m different views of each sample $\mathbf{x}$ and consequently generate consistent blockwise feature maps of U_θ for the use of anomaly detection. We expand the fine-tuned latent diffusion model to a ControlNet version, denoted as Φ, and propose a *view-invariant* training strategy for learning our ControlNet, as illustrated in Fig. 3. Note that we freeze the parameters of the diffusion model during ControlNet training.

We now describe how to proceed with the model training leading to a view-independent ControlNet. To begin with, we comprise a training batch with the same sample of the $m = 7$ views, *i.e.*, $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^7\}$. We then carry out batch training in two subsequent phases. Phase-I takes in turn each of the six RGB views as input and uses $\mathbf{x}^7$ (the 3D normal map) as the *condition* to the ControlNet Φ. We express this training setting to Φ by

$$\boldsymbol{\epsilon}_\theta^i \leftarrow \Phi\{(\mathbf{x}^i, \mathbf{x}^7); t, \boldsymbol{\theta}\} \quad \text{for } i = 1, \ldots, 6, \tag{2}$$

where t is again the forward diffusion steps and $\boldsymbol{\theta}$ are the ControlNet parameters to be learned. To achieve the view-invariant property, we prefer Φ to produce consistent feature maps among the six 2-tuples $\{(\mathbf{x}^i, \mathbf{x}^7)\}_{i=1}^6$ for each of block-1 to block-4 of the stable diffusion decoder of the UNet. Taking all these into account, we learn the ControlNet by minimizing the following loss:

$$\mathcal{L}_{\text{total}}\big(\{(\mathbf{x}^i, \mathbf{x}^7)\}_{i=1}^6; \boldsymbol{\theta}\big) = \underbrace{\sum_{i=1}^{6} \|\epsilon_\theta^i - \epsilon\|}_{\mathcal{L}_\epsilon} + \lambda \cdot \underbrace{\sum_{i=1}^{6}\sum_{b=1}^{4} \|F^b(\mathbf{x}^i, \mathbf{x}^7) - \bar{F}^b\|}_{\mathcal{L}_F}, \quad (3)$$

where $\mathcal{L}_F$, the second term on the right, is a loss term ensuring feature consistency over the six views (*i.e.*, lighting conditions) for each decoder block b, and $\bar{F}^b$ is the average of the six feature maps $\{F^b(\mathbf{x}^i, \mathbf{x}^7)\}_{i=1}^6$ of block b. (See Fig. 4.)

In the Phase-II of a batch training, we instead use the 3D normal map $\mathbf{x}^7$ as the input and each of the other six views as the condition to the ControlNet. That is, we replace the six 2-tuples in (2) and (3) by $\{(\mathbf{x}^7, \mathbf{x}^i)\}_{i=1}^6$ and repeat the same optimization procedure to update the parameters of Φ. Note that we train the ControlNet for 50 epochs and set λ to 0.001 for all experiments.

We now justify the above training scheme for learning a view-invariant ControlNet. Phase-I batch training aims to ensure that the feature representation of decoder block-1 to block-4 of the UNet is consistent across the six RGB-related views. Phase II training also targets blockwise feature consistency, but for the same input of a 3D normal map, conditioned on different RGB-related views, *i.e.*, different lighting conditions. Finally, the switching roles of input and condition between the RGB-related views and the 3D-normal view further enforce the blockwise feature consistency yielded by assessing the same underlying object.

3.4 DDIM Memory Banks and Inference

Having completed the ControlNet training, we are now ready to build two memory banks of diffusion-based features for the online inference of anomaly detection.

DDIM Memory Banks. At this stage of building memory banks for inference, we design a methodology to obtain *UNet features* from a specific decoder block b, which are generated in a way similar to that in training the ControlNet Φ. However, it is crucial to highlight our deliberate choice to employ DDIM inversion [13] over the conventional diffusion process to generate the noise version of the latent representation $\mathbf{z}_t$ for an arbitrary input $\mathbf{x}$. This decision is driven by the aim to mitigate the instability associated with the standard diffusion process, where random noise is added to the latent space. DDIM inversion guarantees that when given the same latent input, the results remain fixed. This stability is essential for our subsequent inference tasks, particularly in the extraction of UNet features to construct the two memory banks: $\mathcal{M}_{\text{rgb}}$ and $\mathcal{M}_{\text{norm}}$. We establish the former using the same settings as in the Phase-I batch training, and the latter with the Phase-II setting. Specifically, to build $\mathcal{M}_{\text{rgb}}$, for each $\mathbf{x} \in D$, we

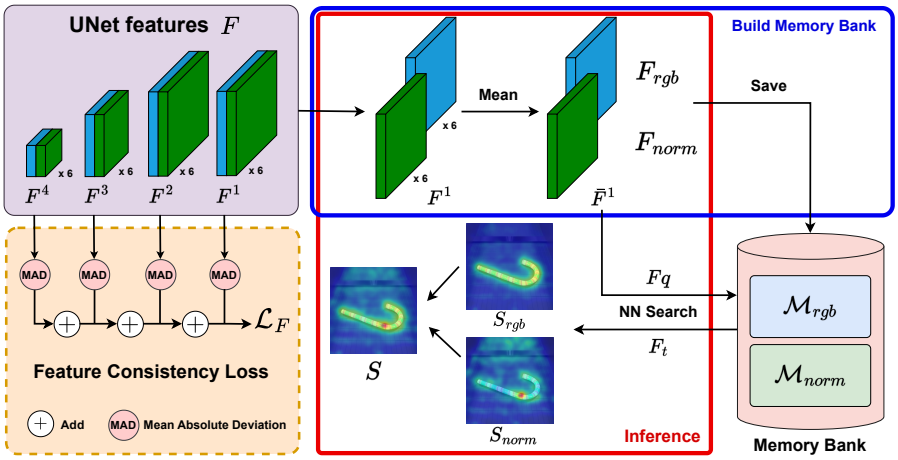

Fig. 4: UNet features utilization. (1) During the training phase, the feature consistency loss is computed by summing the mean absolute deviation (MAD) across F^1 to F^4 within the UNet features F. (2) In the Build Memory Bank phase, we calculate the mean of F^1 to obtain $\bar{F}^1$, which includes RGB features F_{rgb} and normal map features F_{norm}. Subsequently, these features are stored in memory banks respectively, $\mathcal{M}_{\text{rgb}}$ and $\mathcal{M}_{\text{norm}}$. (3) In the inference phase, the F_{rgb} and F_{norm} use the NN search (nearest neighbor search) to find the closest feature in memory banks respectively. The score maps for both modalities S_{rgb} and S_{norm} are the distance between the F_q (query feature) and the F_t (target feature), then we fuse them to get a final result S.

compute its mean feature representation of the UNet decoder block-b by averaging the six feature maps of block-b obtained from $\Phi\{(\mathbf{x}^i, \mathbf{x}^7); t\}$. Note that for each 2-tuple $(\mathbf{x}^i, \mathbf{x}^7)$, we perform t-step DDIM to obtain their noise versions of the latent vectors. Similarly, we can build the memory bank $\mathcal{M}_{\text{norm}}$. For convenience, hereafter, we name the representations from $\mathcal{M}_{\text{rgb}}$ the RGB features and those from $\mathcal{M}_{\text{norm}}$, the NMap features. In our implementation, we empirically extract features from the diffusion-based representation of the decoder block-1 and set the DDIM timestep $t = 80$, designating them as our target UNet features.

Inference. During the inference stage, each testing sample is also provided in multiple views, and our system can perform feature extractions following the same DDIM setting in constructing the two memory banks. Now, for each pixel on the resulting RGB feature map and the NMap feature map, we search for its nearest-neighbor feature in the corresponding memory bank: $\mathcal{M}_{\text{rgb}}$ or $\mathcal{M}_{\text{norm}}$, and calculate the distance between the query and the search results. Based on these distance values, we can obtain two score maps S_{rgb} and S_{norm}.

Since S_{rgb} and S_{norm} are derived from the same ControlNet model, there is no need to align the two score maps, unlike in other previous methods, *e.g.*, Shape-Guided [9], M3DM [41]. The final score map can then be obtained by

simply averaging the two pixel-level score maps. In addition, we obtain the image-level anomaly score by averaging the top-3 scores from S_{rgb} and S_{norm}.

4 Experiments

4.1 Experiment Setup

Dataset. We evaluate our approach using the recently published Eyecandies Dataset [7]. Each sample instance in the dataset comprises six images captured from different lighting angles and a corresponding 3D normal vector map. Specifically, each category has 1000 training instances and 50 test instances. The training set consists entirely of normal (anomaly-free) examples, while the test set contains an equal number of normal and abnormal instances. Our task with this dataset is to identify the most anomalous regions within the presented multi-views. This involves integrating information from six illuminating images and a 3D normal vector map to detect anomalies effectively. This follows the principle of multiple-instance learning, with the challenge that our model must learn to identify anomalies without any examples during training. We also evaluate our method on the MVTec 3D-AD dataset [5] to demonstrate the generalization of our method. Each instance sample contains a single view RGB image and the corresponding 3D point cloud data. The detailed settings are in Appendix A.

Diffusion Model Setting. Our work uses the Stable Diffusion Model [16, 32] version 1-4, which is pretrained on a large-scale dataset LAION-5B [37]. Its stability and robust feature representations make it ideal for our application. In fine-tuning the Diffusion UNet and training the ControlNet [48], we use the AdamW optimizer with a 5×10^{-6} learning rate. The values for β_1 and β_2 are set at 0.9 and 0.999. A weight decay of 0.01 is applied. We restrict the maximum gradient norm 1.0 before each optimization step to avoid overfitting. The fine-tuning process for the Diffusion UNet is completed in just 10 epochs, while the training of ControlNet is set to 50 epochs.

Inference Details. We experiment with various configurations to achieve optimal performance. We find the following settings optimal for both memory-bank construction and inference: noise intensity 80, step size 20, and features from the decoder block-1 of UNet. Furthermore, we observe that the PatchCore [18] re-weighting mechanism does not improve the scoring. Therefore, we adopt the top-k ($k = 3$) scoring method, which effectively enhances the overall score.

Metrics. We use the Area Under the Receiver Operator Curve (AUROC) [5], a common score in anomaly detection tasks, to evaluate our method at both image level (Img-AUROC) and pixel level (Pix-AUROC). We also employ the Area Under the PRO curve (AUPRO) [51] for pixel-level performance evaluation. Due to the possibility of overlooking the detection of small anomalous regions, AUPRO calculates scores for each anomalous component regardless of size.

Table 1: Assessing anomaly detection performance on the Eyecandies dataset using the Img-AUROC metric [5]. The top-performing outcomes are highlighted in red, while the second-best results are indicated in blue.

	Method	Candy Cane	Chocolate Cookie	Chocolate Praline	Confetto	Gummy Bear	Hazelnut Truffle	Licorice Sandwich	Lollipop	Marsh -mallow	Peppermint Candy	**Mean**
RGB only	PatchCore [18]	0.522	0.853	0.621	0.950	0.710	0.624	0.866	0.779	0.982	0.845	0.775
	PADiM [12]	0.531	0.816	0.821	0.856	0.826	0.727	0.784	0.665	0.987	0.924	0.794
	EasyNet [8]	0.723	0.925	0.849	0.966	0.705	0.815	0.806	0.851	0.975	0.960	0.858
	M3DM [41]	0.648	0.949	0.941	1.000	0.878	0.632	0.933	0.811	0.998	1.000	0.879
	DSR [44]	0.706	0.965	0.950	0.966	0.870	0.790	0.885	0.857	0.998	0.992	0.898
	Ours	0.710	0.998	0.955	0.966	0.843	0.592	0.947	0.875	1.000	0.995	0.888
3D only	EasyNet [8]	0.629	0.716	0.768	0.731	0.660	0.710	0.712	0.711	0.688	0.731	0.706
	M3DM [41]	0.482	0.589	0.805	0.845	0.780	0.538	0.766	0.827	0.800	0.822	0.725
	FPFH [18]	0.670	0.728	0.806	0.806	0.721	0.514	0.794	0.757	0.758	0.757	0.731
	3DSR [44]	0.600	0.768	0.742	0.770	0.761	0.749	0.811	0.831	0.811	0.917	0.776
	Ours	0.790	0.885	0.933	0.915	0.837	0.517	0.888	0.960	0.941	0.949	0.862
3D+RGB	BTF [18]	0.712	0.882	0.784	0.942	0.774	0.579	0.829	0.840	0.986	0.882	0.821
	EasyNet [8]	0.737	0.934	0.866	0.966	0.717	0.822	0.847	0.863	0.977	0.960	0.869
	CFM [11]	0.680	0.931	0.952	0.880	0.865	0.782	0.917	0.840	0.998	0.962	0.881
	M3DM [41]	0.624	0.958	0.958	1.000	0.886	0.758	0.949	0.836	1.000	1.000	0.897
	3DSR [44]	0.651	0.998	0.904	0.978	0.875	0.861	0.965	0.899	0.990	0.971	0.909
	MMRD [14]	0.854	1.000	0.946	0.998	0.908	0.747	0.966	0.984	1.000	1.000	**0.940**
	Ours	0.859	1.000	1.000	0.995	0.910	0.738	0.998	0.976	1.000	1.000	**0.948**

4.2 Experimental Results

In the experiments, we compare with SOTA multi-modality methods using the image-level metric Img-AUROC in Tab. 1 and the pixel-level metric AUPRO in Tab. 2. Our approach combines the multi-view instance diffusion learning technique, resulting in the highest scores at the image level, even when evaluating individually for the RGB or Nmap modality. Note that for both RGB-only and Nmap-only scores, we employ the UNet to compute the single-modality scores. Table 3 provides the Img-AUROC results on MVTec 3D-AD for our method and other approaches. Since there is only one data view available, we exclude the feature loss during training. Our approach consistently achieves high scores across all categories and sets a new standard among feature embedding-based methods. When compared to reconstruction-based techniques like MMRD [14] and CFM [11], our method demonstrates substantial competitiveness. It is important to note that MMRD and CFM are class-dependent models and CFM requires distinct backbones to process 2D RGB and 3D data separately. Additional experimental results on MVTec3D-AD can be found in Appendix A. We verify that our method is effective on multi-view and multi-modality datasets.

4.3 Qualitative Results

The qualitative results are shown in Fig. 5. Our RGB score map S_{rgb} distinguishes areas of discoloration on Chocolatecookie and Marshmallow (first and second rows). For 3D geometric anomalies that are challenging to detect visually, our method leverages the information from the Normal Map to identify these anomalies on the GummyBear and Chocolatepraline (the third and fourth rows) through the Nmap score map S_{nmap}. As depicted in the final score map S, our approach successfully integrates information from both 3D and RGB modalities.

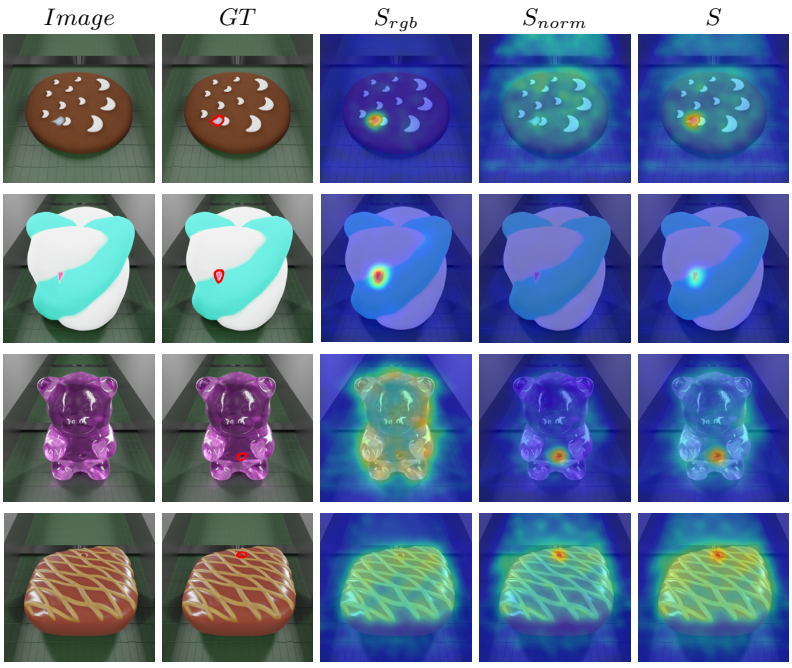

Fig. 5: Qualitative results on RGB, Normal, and RGB+Normal score maps.

Through the complementary capabilities of these two modalities, our method can detect anomalies in both appearance and shape. More visual results are provided in Appendixes B and C.

5 Ablation Study

5.1 The Effectiveness of ControlNet

In Tab. 4, we analyze the effectiveness of incorporating ControlNet in our approach. There is evidence that for both the single-modality UNet and the multi-modality UNet, including ControlNet during training leads to performance improvements. Specifically, our ControlNet demonstrates a 1.4% improvement in Img-ROC and a 0.5% improvement in AUPRO. This observation demonstrates that integrating ControlNet during training can further enhance the ability of internal representation within the diffusion model for anomaly detection tasks. The Fig. 7 exhibits the visual improvement for ControlNet. The heatmap score will be emphasized when the model includes ControlNet. On the contrary, the anomaly region is insignificant when the model lacks help from the ControlNet. By the phenomenon, the ControlNet improves the localization for anomalous regions.

Table 2: Assessing anomaly detection performance on the Eyecandies dataset using the AUPRO metric [51]. The top-performing outcomes are highlighted in boldface.

	Method	Candy Cane	Chocolate cookie	Chocolate praline	Confetto	Gummy bear	Hazelnut truffle	Licorice sandwich	Lollipop	Marsh mallow	Peppermint candy	**Mean**
RGB	PatchCore [18]	0.773	0.857	0.594	0.965	0.762	0.532	0.887	0.871	0.942	0.898	0.808
	M3DM [41]	0.867	0.904	0.805	0.982	0.871	0.662	0.882	0.895	0.970	0.962	0.880
	Ours	0.919	0.942	0.887	0.978	0.910	0.627	0.961	**0.946**	0.982	0.983	0.914
3D	FPFH [18]	0.944	0.725	0.687	0.601	0.651	0.471	0.636	0.885	0.598	0.594	0.679
	M3DM [41]	0.911	0.645	0.581	0.748	0.748	0.484	0.608	0.904	0.646	0.750	0.702
	Ours	0.842	0.841	0.870	0.868	0.814	0.591	0.838	0.865	0.776	0.774	0.808
Both	BTF [18]	0.871	0.900	0.698	0.966	0.823	0.567	0.884	0.905	0.953	0.897	0.846
	M3DM [41]	0.906	0.923	0.803	0.983	0.855	0.688	0.880	0.906	0.966	0.955	0.882
	CFM [11]	0.942	0.902	0.831	0.965	0.875	0.762	0.791	0.913	0.939	0.949	0.887
	MMRD [14]	**0.975**	**0.970**	0.942	**0.985**	0.917	0.680	**0.970**	0.941	**0.990**	**0.992**	0.936
	Ours	0.964	0.953	**0.951**	0.982	**0.931**	**0.765**	0.969	0.935	0.982	0.983	**0.941**

Table 3: Assessing anomaly detection performance on the MVTec 3D-AD dataset [5] using the Img-AUROC metric. "SG" means Shape-Guided [9]. The top-performing outcomes are highlighted in red, while the second-best results are indicated in blue.

	Method	Bagel	Cable -Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	**Mean**
RGB	BTF [18]	0.854	0.840	0.824	0.687	0.974	0.716	0.713	0.593	0.920	0.724	0.785
	SG [9]	0.911	0.936	0.883	0.662	0.974	0.772	0.785	0.641	0.884	0.706	0.815
	M3DM [41]	0.944	0.918	0.896	0.749	0.959	0.767	0.919	0.648	0.938	0.767	0.850
	Ours	0.910	0.871	0.854	0.687	0.908	0.859	0.885	0.556	0.914	0.720	0.816
3D	BTF [18]	0.820	0.533	0.877	0.769	0.718	0.574	0.774	0.895	0.990	0.582	0.753
	M3DM [41]	0.941	0.651	0.965	0.969	0.905	0.760	0.880	0.974	0.926	0.765	0.874
	SG [9]	0.983	0.682	0.978	0.998	0.960	0.737	0.993	0.979	0.966	0.871	0.916
	Ours	0.965	0.852	0.962	0.988	0.963	0.900	0.990	0.984	0.965	0.823	0.939
3D+RGB	BTF [18]	0.938	0.765	0.972	0.888	0.960	0.664	0.904	0.929	0.982	0.726	0.873
	M3DM [41]	0.994	0.909	0.972	0.976	0.960	0.942	0.973	0.899	0.972	0.850	0.945
	SG [9]	0.986	0.894	0.983	0.991	0.976	0.857	0.990	0.965	0.960	0.869	0.945
	MMRD [14]	0.999	0.943	0.964	0.943	0.992	0.912	0.949	0.901	0.994	0.901	0.950
	CFM [11]	0.988	0.875	0.984	0.992	0.997	0.924	0.964	0.949	0.979	0.950	0.960
	Ours	0.967	0.880	0.964	0.984	0.963	0.978	0.990	0.973	0.960	0.829	0.949

5.2 Different Noise Intensity

Our approach uses DDIM Inversion to invert the VAE latent to a noisy state. Although higher noise intensity leads to more powerful feature representations (as shown in Fig. 6), it also increases the number of iterations needed for DDIM Inversion, which means longer processing time. To balance performance and efficiency, we set the noise intensity to 80. Furthermore, the ControlNet consistently outperforms the UNet across all noise intensities.

5.3 The Effectiveness of Feature Loss

The Eyecandies images are captured in six different views. We use the feature loss to integrate the representations of features consistent between six views. Fig. 8 shows that the model with feature loss has better-separated distributions

Table 4: Comparison of UNet and ControlNet architectures. *U* denotes UNet, *C* represents ControlNet, and Both means RGB+3D.

Model	Img-ROC	Pix-ROC	AUPRO
RGB *U*	0.896	0.979	0.918
Nmap *U*	0.857	0.924	0.813
Both *U*	0.934	0.981	0.936
RGB *C*	0.900	0.980	0.919
Nmap *C*	0.865	0.828	0.823
Both *C*	**0.948**	**0.983**	**0.941**

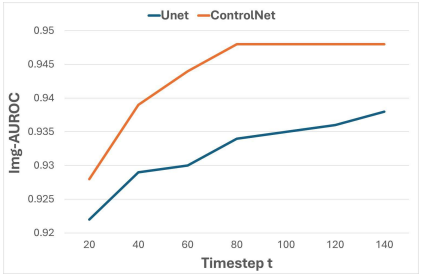

Fig. 6: The impact of different timesteps on the scores.

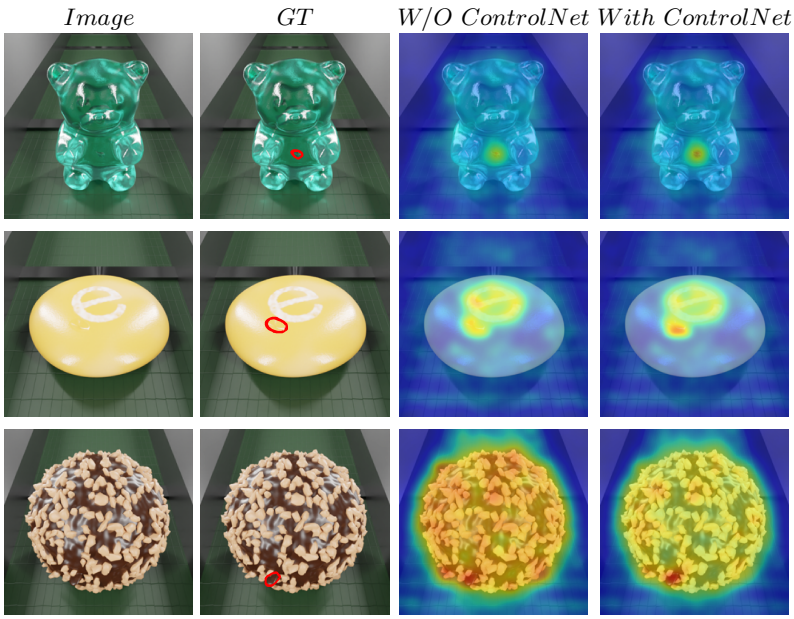

Fig. 7: ControlNet is useful for highlighting abnormal regions on the heatmap.

of normal and anomaly distances (distance of 0.458), and the overlapping area (orange box) between normal and anomaly is smaller compared to the model without feature loss. In Tab. 5, systematically comparing each modality, we find that models with feature loss unanimously outperform those without feature loss.

5.4 Different Feature Layers

For embedding-based anomaly detection methods, assessing features from different network layers is crucial. Our approach is the first to explore diffusion models with feature embedding techniques. We further evaluate the features from different

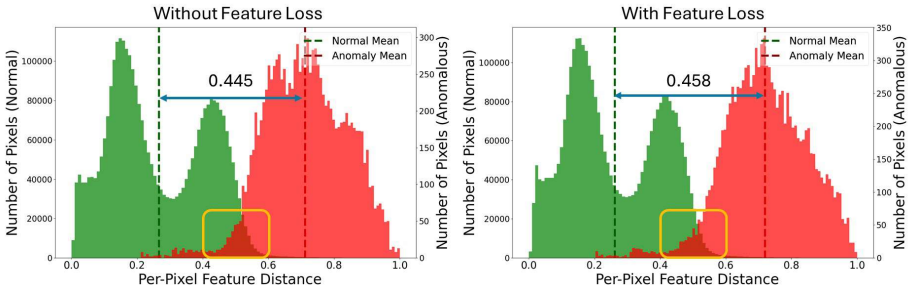

Fig. 8: Per-pixel feature distance distribution of the ControlNet without and with feature loss. The distributions of both images are results after min-max normalization.

Table 5: Ablation on feature loss of our ControlNet. "Both" refers to RGB+3D.

Modality	Feature Loss	Img ROC	Pix ROC	AU PRO
RGB		0.890	0.978	0.915
Nmap		0.858	0.924	0.810
Both		0.934	**0.983**	0.940
RGB	✓	0.897	0.980	0.921
Nmap	✓	0.869	0.934	0.827
Both	✓	**0.948**	**0.983**	**0.941**

Table 6: Comparison of using different decoder block features as embeddings.

SDDB	Img-ROC	Pix-ROC	AUPRO
1, 2, 3	0.868	0.972	0.876
2, 3	0.830	0.967	0.848
1, 2	0.920	**0.983**	0.928
3	0.815	0.959	0.825
2	0.875	0.979	0.901
1	**0.948**	**0.983**	**0.941**

decoder blocks of the diffusion model in Tab. 6, and find that the features from the decoder block-1 of UNet achieve the optimal performance.

6 Conclusion

We have developed a diffusion-based technique to address the task of multi-view anomaly detection. The efficacy of our method is based on three key designs. First, we propose to fine-tune a pretrained latent diffusion model in a way that is independent of the various data views. Second, we introduce a feature consistency loss and mutual-conditioning between the 2D-RGB and 3D-Normal modalities to learn a view-invariant ControlNet. Third, we leverage the DDIM scheme to build memory banks of deterministic diffusion-based features, ensuring consistently good performance in inference. For future work, we plan to further explore our new finding in treating the ControlNet model as a versatile diffusion-related information fuser for multi-modal computer vision applications.

Acknowledgements. This work was supported in part by the NSTC grants 111-2221-E-001-015-MY3 and 112-2634-F-007-002 of Taiwan. We are grateful to the National Center for High-performance Computing for providing computational resources and facilities.

References

1. Bae, J., Lee, J., Kim, S.: PNI: industrial anomaly detection using position and neighborhood information. In: ICCV (2023)
2. Batzner, K., Heckler, L., König, R.: Efficientad: Accurate visual anomaly detection at millisecond-level latencies. CoRR (2023)
3. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C.: The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. Int. J. Comput. Vis. (2021)
4. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: CVPR (2020)
5. Bergmann, P., Jin, X., Sattlegger, D., Steger, C.: The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. In: Farinella, G.M., Radeva, P., Bouatouch, K. (eds.) VISIGRAPP (2022)
6. Bergmann, P., Sattlegger, D.: Anomaly detection in 3d point clouds using deep geometric descriptors. In: WACV (2023)
7. Bonfiglioli, L., Toschi, M., Silvestri, D., Fioraio, N., Gregorio, D.D.: The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In: Wang, L., Gall, J., Chin, T., Sato, I., Chellappa, R. (eds.) ACCV (2022)
8. Chen, R., Xie, G., Liu, J., Wang, J., Luo, Z., Wang, J., Zheng, F.: Easynet: An easy network for 3d industrial anomaly detection. In: El-Saddik, A., Mei, T., Cucchiara, R., Bertini, M., Vallejo, D.P.T., Atrey, P.K., Hossain, M.S. (eds.) International Conference on Multimedia (2023)
9. Chu, Y., Liu, C., Hsieh, T., Chen, H., Liu, T.: Shape-guided dual-memory learning for 3d anomaly detection. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) ICML (2023)
10. Cohen, N., Hoshen, Y.: Sub-image anomaly detection with deep pyramid correspondences. CoRR (2020)
11. Costanzino, A., Zama Ramirez, P., Lisanti, G., Di Stefano, L.: Multimodal industrial anomaly detection by crossmodal feature mapping. In: CVPR (2024)
12. Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: A patch distribution modeling framework for anomaly detection and localization. In: Bimbo, A.D., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R. (eds.) Pattern Recognition. ICPR International Workshops and Challenges (2020)
13. Dhariwal, P., Nichol, A.Q.: Diffusion models beat gans on image synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) NeurIPS (2021)
14. Gu, Z., Zhang, J., Liu, L., Chen, X., Peng, J., Gan, Z., Jiang, G., Shu, A., Wang, Y., Ma, L.: Rethinking reverse distillation for multi-modal anomaly detection. Proceedings of the AAAI Conference on Artificial Intelligence (2024)
15. He, H., Zhang, J., Chen, H., Chen, X., Li, Z., Chen, X., Wang, Y., Wang, C., Xie, L.: Diad: A diffusion-based framework for multi-class anomaly detection. CoRR (2023)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) NeurIPS (2020)
17. Ho, J., Salimans, T.: Classifier-free diffusion guidance. CoRR (2022)
18. Horwitz, E., Hoshen, Y.: Back to the feature: Classical 3d features are (almost) all you need for 3d anomaly detection. In: CVPR (2023)
19. Hu, T., Zhang, J., Yi, R., Du, Y., Chen, X., Liu, L., Wang, Y., Wang, C.: Anomalydiffusion: Few-shot anomaly image generation with diffusion model. CoRR (2023)

20. Hudson, D.A., Zoran, D., Malinowski, M., Lampinen, A.K., Jaegle, A., McClelland, J.L., Matthey, L., Hill, F., Lerchner, A.: SODA: bottleneck diffusion models for representation learning (2023)
21. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: Direct inversion: Boosting diffusion-based editing with 3 lines of code (2023)
22. Karras, J., Holynski, A., Wang, T., Kemelmacher-Shlizerman, I.: Dreampose: Fashion image-to-video synthesis via stable diffusion. In: ICCV (2023)
23. Lee, S., Lee, S., Song, B.C.: CFA: coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. IEEE Access (2022)
24. Li, C., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: Self-supervised learning for anomaly detection and localization. In: CVPR (2021)
25. Liu, J., Xie, G., Chen, R., Li, X., Wang, J., Liu, Y., Wang, C., Zheng, F.: Real3d-ad: A dataset of point cloud anomaly detection (2023)
26. Liu, Z., Zhou, J.P., Wang, Y., Weinberger, K.Q.: Unsupervised out-of-distribution detection with diffusion inpainting. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) ICML (2023)
27. Liu, Z., Zhou, Y., Xu, Y., Wang, Z.: Simplenet: A simple network for image anomaly detection and localization. In: CVPR (2023)
28. Lu, F., Yao, X., Fu, C., Jia, J.: Removing anomalies as noises for industrial defect localization. In: ICCV (2023)
29. Luo, G., Dunlap, L., Park, D.H., Holynski, A., Darrell, T.: Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) NeurIPS (2023)
30. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: CVPR (2023)
31. Mousakhan, A., Brox, T., Tayyub, J.: Anomaly detection with conditioned denoising diffusion models. CoRR (2023)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
33. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.V.: Towards total recall in industrial anomaly detection. In: CVPR (2022)
34. Rudolph, M., Wehrbein, T., Rosenhahn, B., Wandt, B.: Fully convolutional cross-scale-flows for image-based defect detection. In: Winter Conference on Applications of Computer Vision (WACV) (2022)
35. Rudolph, M., Wehrbein, T., Rosenhahn, B., Wandt, B.: Asymmetric student-teacher networks for industrial anomaly detection. In: WACV (2023)
36. Schlüter, H.M., Tan, J., Hou, B., Kainz, B.: Natural synthetic anomalies for self-supervised anomaly detection and localization. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV (2022)
37. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022)
38. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
39. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) NeurIPS (2023)
40. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to- image translation. In: CVPR (2023)

41. Wang, Y., Peng, J., Zhang, J., Yi, R., Wang, Y., Wang, C.: Multimodal industrial anomaly detection via hybrid fusion. In: CVPR (2023)
42. Wu, A., Chen, D., Deng, C.: Deep feature deblurring diffusion for detecting out-of-distribution objects. In: ICCV (2023)
43. Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In: CVPR (2022)
44. Zavrtanik, V., Kristan, M., Skocaj, D.: Cheating depth: Enhancing 3d surface anomaly detection via depth simulation. CoRR (2023)
45. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem – a discriminatively trained reconstruction embedding for surface anomaly detection (2021)
46. Zavrtanik, V., Kristan, M., Skočaj, D.: Reconstruction by inpainting for visual anomaly detection. Pattern Recognition (2021)
47. Zhang, H., Wang, Z., Wu, Z., Jiang, Y.G.: Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection (2023)
48. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
49. Zhang, X., Li, N., Li, J., Dai, T., Jiang, Y., Xia, S.: Unsupervised surface anomaly detection with diffusion probabilistic model. In: ICCV (2023)
50. Zhang, X., Li, S., Li, X., Huang, P., Shan, J., Chen, T.: Destseg: Segmentation guided denoising student-teacher for anomaly detection. In: CVPR (2023)
51. Zheng, Y., Wang, X., Qi, Y., Li, W., Wu, L.: Benchmarking unsupervised anomaly detection and localization. CoRR (2022)
52. Zhou, Q., Li, W., Jiang, L., Wang, G., Zhou, G., Zhang, S., Zhao, H.: Pad: A dataset and benchmark for pose-agnostic anomaly detection (2023)