electronics

MDPI

Article

Memoryless Multimodal Anomaly Detection via Student–Teacher Network and Signed Distance Learning

Zhongbin Sun [1,2,*], Xiaolong Li [2], Yiran Li [3] and Yue Ma [2]

[1] Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou 221116, China
[2] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China; lixiaolong@cumt.edu.cn (X.L.); 08222214@cumt.edu.cn (Y.M.)
[3] Sun YueQi Honors College, China University of Mining and Technology, Xuzhou 221116, China; 07220333@cumt.edu.cn
* Correspondence: zhongbin@cumt.edu.cn

Abstract: Unsupervised anomaly detection is a challenging computer vision task, in which 2D-based anomaly detection methods have been extensively studied. However, multimodal anomaly detection based on RGB images and 3D point clouds requires further investigation. The existing methods are mainly inspired by memory bank-based methods commonly used in 2D-based anomaly detection, which may cost extra memory for storing multimodal features. In the present study, a novel memoryless method MDSS is proposed for multimodal anomaly detection, which employs a lightweight student–teacher network and a signed distance function to learn from RGB images and 3D point clouds, respectively, and complements the anomaly information from the two modalities. Specifically, a student–teacher network is trained with normal RGB images and masks generated from point clouds by a dynamic loss, and the anomaly score map could be obtained from the discrepancy between the output of student and teacher. Furthermore, the signed distance function learns from normal point clouds to predict the signed distances between points and surfaces, and the obtained signed distances are used to generate an anomaly score map. Subsequently, the anomaly score maps are aligned to generate the final anomaly score map for detection. The experimental results indicate that MDSS is comparable but more stable than SOTA methods and, furthermore, performs better than other baseline methods.

Keywords: multimodal; anomaly detection; memory bank; student–teacher network; signed distance function

check for updates

Citation: Sun, Z.; Li, X.; Li, Y.; Ma, Y. Memoryless Multimodal Anomaly Detection via Student–Teacher Network and Signed Distance Learning. *Electronics* **2024**, *13*, 3914. https://doi.org/10.3390/electronics13193914

Academic Editors: Silvia Liberata Ullo

Received: 23 September 2024
Revised: 1 October 2024
Accepted: 1 October 2024
Published: 3 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Visual anomaly detection aims to detect abnormal objects from visual information, which is widely used in industrial and medical imaging fields [1–3]. In practical application scenarios, due to the low proportion of abnormal areas, unknown abnormal patterns and expensive annotation costs, it is difficult to obtain high-quality labeled datasets. Therefore, unsupervised anomaly detection has been the subject of researchers' interest, in which previous research has mainly focused on 2D anomaly detection with RGB images [4].

With the proposal of the MVTec 3D-AD dataset [5] in 2022, researchers have begun to study the feasibility of combining 3D point clouds with RGB images for multimodal anomaly detection [6–15]. The key to unsupervised multimodal anomaly detection lies in how to integrate information from two modalities to distinguish normal and abnormal samples. We categorize the existing methods in multimodal anomaly detection into two classes: (1) student–teacher network-based methods, (2) memory bank-based methods.

In student–teacher network-based methods, the RGB features and the depth from point clouds are concatenated and input into the student–teacher network for anomaly detection [10], as shown in Figure 1a. Particularly, the output difference between the teacher and the student can be regarded as anomaly scores for indicating the possibility of a sample

to be abnormal. However, only using depth discards the 3D information in the x and y coordinates and the direct concatenation may cause a disturbance between the features of different modalities, which harms the detection performance.

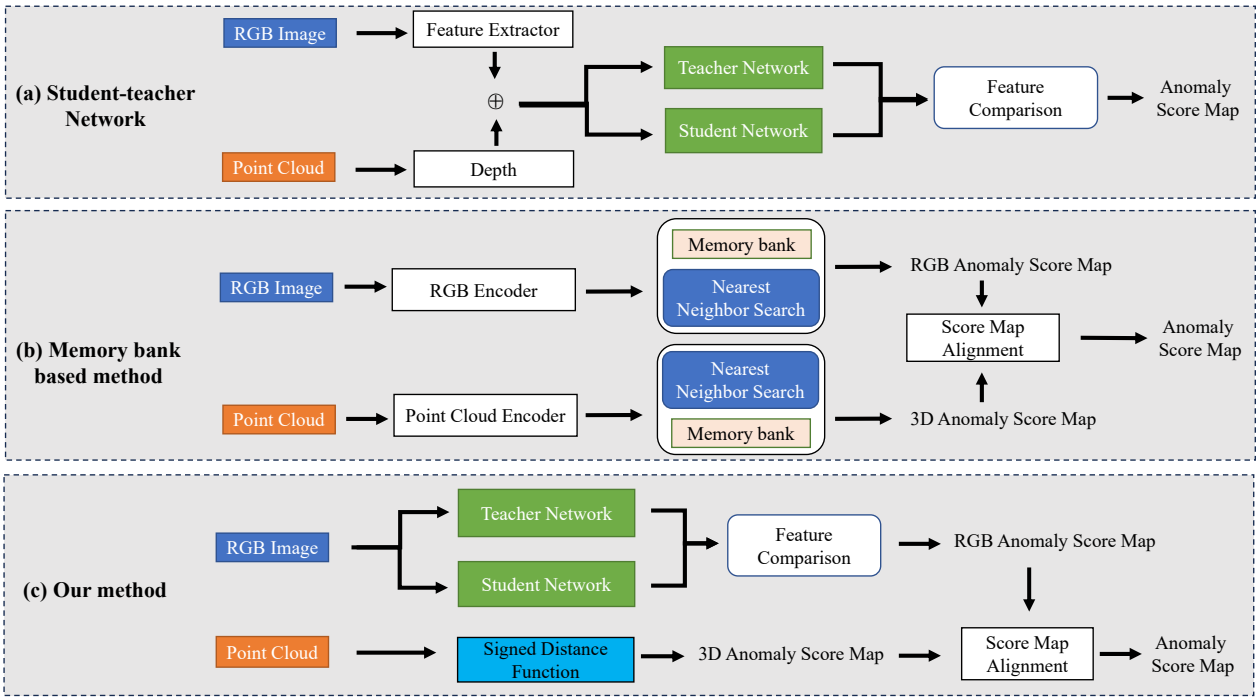


Figure 1. Comparison of our method with other methods in multimodal anomaly detection. (**a**) The process of student-teacher network based method. (**b**) The process of memory bank based method. (**c**) The process of our method.

In recent years, memory bank-based methods have drawn wide attention, in which the features of normal samples from different modalities are stored and the training process is eliminated with a pretrained backbone [7,9,11,13,15], as shown in Figure 1b. For the inference stage, the new samples are input into the same feature extractor, and corresponding output features are compared with the stored normal sample features to generate anomaly scores. However, considering the fact that the backbones of these methods are typically complex networks that have a lot of layers such as ResNet [16] or Vision Transformer [17], the feature extraction may consume a lot of time during testing, which affects the real-time performance of detection. Moreover, these memory bank-based methods consume extra memory and have high hardware requirements, which may be difficult to apply in some real application scenarios.

To address the aforementioned issues, a **M** emoryless multimodal anomaly **D**etection method by combining **S**tudent–teacher network and **S**igned distance learning (**MDSS**) is proposed, as shown in Figure 1c. In MDSS, for the purpose of reducing disturbance between modalities, the student–teacher network is used to process RGB images, and the signed distance function is employed to process 3D point clouds, respectively. Our method abandons the usage of the memory bank and combines a student–teacher network and signed distance function (SDF) for multimodal anomaly detection, improving detection accuracy and reducing hardware deployment costs. Both of them are lightweight in our model.

To be specific, in a student–teacher network, normal RGB images with masks generated from 3D point clouds are employed to obtain the RGB anomaly score map. Moreover, existing research [18] shows that excessive normal images for training may lead to homogenization of the student and teacher, causing the student to mimic the output of the teacher beyond normal samples, thereby hindering the detection accuracy, while insufficient train-

ing images may hinder the student from learning the features of normal samples. Therefore, MDSS employs a dynamic learning factor in loss function to train the student–teacher network to improve the anomaly detection performance. In addition, to the best of our knowledge, we are the first to propose the direct utilization of a signed distance function for unsupervised 3D anomaly detection. For signed distance learning in MDSS, the signed distance function is employed for surface reconstruction from normal point clouds and outputs the distances from the points to the surface. We assume that the distance between abnormal points and the surface is larger than that of normal points [19]; thus, the distance can be used to measure the possibility of a point being abnormal, and the corresponding 3D anomaly score map is obtained. Finally, a statistical approach is employed to align the RGB anomaly score map and the 3D anomaly score map. Then, the aligned score map, which combines the anomaly information from both RGB and 3D is used for image-level and pixel-level anomaly detection.

In the experimental study, the popular MVTec 3D-AD dataset [5] is used and several multimodal anomaly detection methods are selected as the baseline methods. The experimental results show that MDSS is comparable but more stable than SOTA methods and, furthermore, performs better than other baseline methods. In addition, we also conducted an ablation study to demonstrate the effectiveness of combining the student–teacher network and signed distance learning. Moreover, the influence analysis of hyper-parameters in MDSS is provided and different alignment strategies are compared for validating the effectiveness of the employed alignment strategy.

Figure 2 illustrates the complementary advantage of our method for precisely localizing the anomaly from different modalities. From Figure 2, it could be observed that some anomalies in objects are caused by irregularities in color such as in Carrot, which can mostly be seen in the S-T score map. Meanwhile, others are due to their 3D structural shape just like Potato, which can only be seen in an SDF score map. When the two models are combined, the score map highlights the anomaly from both modalities. We can draw the conclusion that the two modules in our method MDSS are effective and complementary.

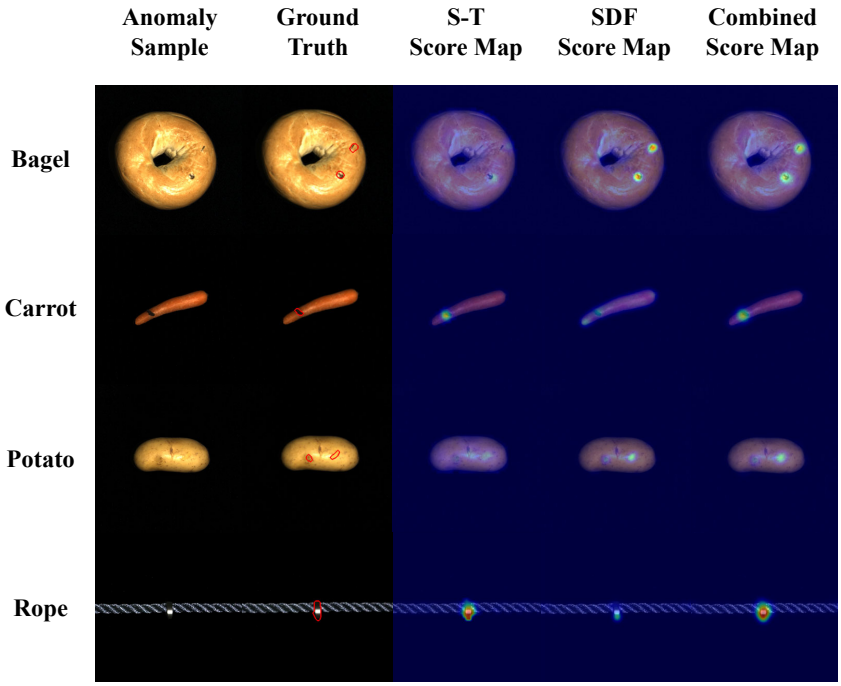


Figure 2. The visualization of our model's detection in different modalities. The red circles pinpoint the ground truth.

This paper extends our previous work (It has been accepted in The 7th Chinese Conference on Pattern Recognition and Computer Vision, PRCV 2024) [20] as follows. Firstly,

we reviewed the theoretical foundation of SDF, analyzed the interpretability of its application in 3D anomaly detection through an experimental study, and investigated the reasons why the detection performance benefits from the combination of two modules, a student–teacher network and signed distance learning. Secondly, we discussed and compared more recently proposed methods for anomaly detection, including CFM [8], ITNM [13], MMRD [14], CPMF [15]. Lastly, we discussed the influence of different settings on MDSS by conducting some experiments, including an ablation study, as well as hyper-parameter analysis, and we also made a comparison between the different alignment strategies.

The remainder of this paper is organized as follows. Section 2 offers a brief review of related work. In Section 3, we introduce the new method MDSS and analyze its interpretability. Then, the experimental results along with their discussion are included in Section 4. Finally, the main conclusions of this work are given in Section 5.

2. Related Work

2.1. Two-Dimensional Image Anomaly Detection

Many types of methods have been presented to solve 2D image anomaly detection and localization, such as reconstruction-based methods [21–25], normalized flow-based methods [26–29], student–teacher network-based methods [10,18,30,31], and memory bank-based methods [32–34]. However, not all of these methods can be transferred and applied to multimodal anomaly detection. The most related are student–teacher network-based methods and memory bank-based methods, so we will provide a detailed introduction to these two types of methods.

A student–teacher network is originally an approach for knowledge distillation [35]. Bergmann et al. [36] first applied the student–teacher network to 2D anomaly detection. They assumed significant regression errors between students and the teacher network, and the substantial uncertainty among multiple student networks can be used as anomaly representation for detection. The STPFM [30] introduces the integration of a multi-scale feature-matching strategy into the student–teacher framework. The hierarchical feature-matching strategy enables the student network to receive a mixture of multi-level knowledge from the feature pyramid under better supervision, which enables the detection of anomalies for diverse sizes. Deng et al. [31] present a novel student–teacher network comprising a teacher encoder and a student decoder, along with the introduction of a straightforward yet powerful "reverse distillation" paradigm. Their student network accepts the one-class embedding from the teacher model as input and aims to reconstruct the teacher's multi-scale representation. EfficientAD [18], respectively, utilizes an autoencoder and a lightweight student–teacher network trained with an asymmetric loss function to obtain a global and a local anomaly map and then normalizes them to generate a combined anomaly map complementing information to improve detection performance and computational efficiency. It is worth mentioning that the global semantic information extracted from the global anomaly map enhances the model's performance in detecting logical anomalies. Rudolph et al. [10] propose asymmetric student–teacher networks (AST). To be specific, they train a normalizing flow for density estimation as the teacher and a conventional feed-forward network as the student to induce significant distances for anomaly detection.

Memory bank-based methods initially originate from 2D anomaly detection. Cohen et al. [34,37] propose to store the deep pretrained features and use the K nearest neighbors of features extracted from a new sample to conduct both anomaly detection and localization. This approach later becomes the foundation of many memory bank-based methods. Many researchers have begun to investigate such methods in 2D anomaly detection. The most representative method is PatchCore [33]. PatchCore utilizes a variant of ResNet as the backbone pretrained on the ImageNet dataset and employs a maximally representative memory bank characterized by greedy coreset subsampling for storing nominal patch-level features. During the inference, samples are input into the backbone to obtain features,

which are then compared with the features in the memory bank to calculate corresponding similarity values for determining whether they are anomalous.

2.2. Multimodal Anomaly Detection

In the field of multimodal anomaly detection, researchers explore the applicability of the student–teacher network. Bergmann et al. [6] construct an expressive teacher network that extracts dense local geometric descriptors from the input image. In their student–teacher network, the regression errors between the teacher and the student are utilized to achieve reliable localization of anomalous structures. Another student–teacher network, AST [10], can also be used in the multimodal setting where the RGB features and the depth from point clouds are concatenated to train the student–teacher network for multimodal anomaly detection. MMRD [14] employs a frozen multimodal teacher encoder to produce distillation targets, which the learnable student decoder is designed to restore. This approach incorporates information from the auxiliary modality into both the frozen teacher encoder and the student decoder across multiple feature levels.

In addition, researchers also migrate the memory bank-based methods from 2D anomaly detection. BTF [9] combines handcrafted 3D representations (FPFH [38]) with a deep, color-based method (PatchCore [33]). BTF outperforms the baseline provided by the author of MVTec 3D-AD by a large margin and justifies that there are complementary benefits from using both 3D and color modalities. M3DM [11] employs different backbones for RGB images and point clouds, respectively, and designs an unsupervised feature fusion with patch-wise contrastive learning to encourage the interaction of different modal features, which are stored in multiple memory banks for final detection. However, with the memory banks, the memory cost of M3DM in inference can be 6.52 GB and the FPS is only 0.51 which may be unacceptable in some real application scenarios [8]. Shape-guided [7] uses the two experts (ResNet for RGB images and signed distance function for point clouds) to build the dual memory banks from the anomaly-free training samples and perform shape-guided inference. Wang et al. [13] point out that most advanced multimodal methods rely on one-time training for anomaly detection, overlooking the ongoing generation of new samples in industrial settings. To address this, they propose ITNM, a method that allows for efficient incremental updates of the model by incorporating new nominal samples, enabling the template set used for generating anomaly score maps to be continuously refreshed. CPMF [15] combines local geometric features from 3D handcrafted descriptors with global semantic information from 2D pretrained neural networks. It aligns 2D and 3D modalities by projecting multi-view 2D features back into 3D space and aggregating them into point-wise 2D features. These point-wise features are then fused and stored in a memory bank for point cloud anomaly detection.

In conclusion, the drawback of current student–teacher network-based methods is that the RGB features and point clouds are directly concatenated, which may result in interference between modalities. Furthermore, using only depth leads to the loss of some 3D information in the point clouds. Moreover, memory bank-based methods suffer from the disadvantage of excessive memory consumption.

In addition to the two mainstream methods mentioned above, there are also some other novel methods. EasyNet [12] integrates a multi-scale, multimodality feature encoder–decoder with a multimodality anomaly segmentation network and introduces an attention-based information entropy fusion module for feature fusion during inference, allowing it to effectively reconstruct segmentation maps of anomalous regions without relying on pretrained models or memory banks, which is well-suited for real-time deployment. Costanzino et al. propose CFM [8], in which a light and fast framework is introduced for learning to map features from one modality to the other on normal samples. During inference, anomalies are detected by pinpointing inconsistencies between observed and mapped features. CFM achieves faster inference and occupies less memory than memory bank-based methods. These two methods will also be selected as two baseline methods in the present study.

2.3. Signed Distance Function

The signed distance function (SDF) is a continuous function for outputting the point's distance to the closest surface. Therefore, SDF could be used to describe the shape of an object's surface in three-dimensional space with positive values indicating the point is outside, negative values signifying that it is inside, and zero meaning the point is on the surface. The formula of SDF is as follows:

$$SDF(x) = s : x \in R^3, s \in \mathbb{R} \tag{1}$$

In recent years, the signed distance function has become a key representation of 3D shapes in deep learning-based shape analysis [39], which offers a significant advantage over other methods in capturing high-resolution shapes with complex topology. In deep learning-based shape analysis, to learn a signed distance function, it is common to train a deep neural network to predict signed distance values for specific 3D locations based on data with ground truth. Such a neural network is referred to as a neural implicit field [40–42]. However, these methods need ground truth to supervise the training of neural networks, which makes them inapplicable to unsupervised anomaly detection. In 2022, Ma et al. [39] introduced a method for learning the signed distance function directly from raw point clouds without ground truth. Their approach involves mapping the surrounding 3D space onto the surface represented by the point cloud, thereby facilitating a self-supervised learning paradigm that aligns well with the requirements of unsupervised anomaly detection. Thereafter, shape-guided [7] is the first method to incorporate a signed distance function into unsupervised anomaly detection, which primarily employs the middle features to construct a memory bank. However, it neglects the potential of the signed distance function to capture detailed anomaly information within 3D shape geometry.

3. Method

In our memoryless multimodal anomaly detection method MDSS, three modules are included, respectively, student–teacher network, signed distance learning and score map alignment. Figure 3 provides the detailed framework of the proposed method MDSS.

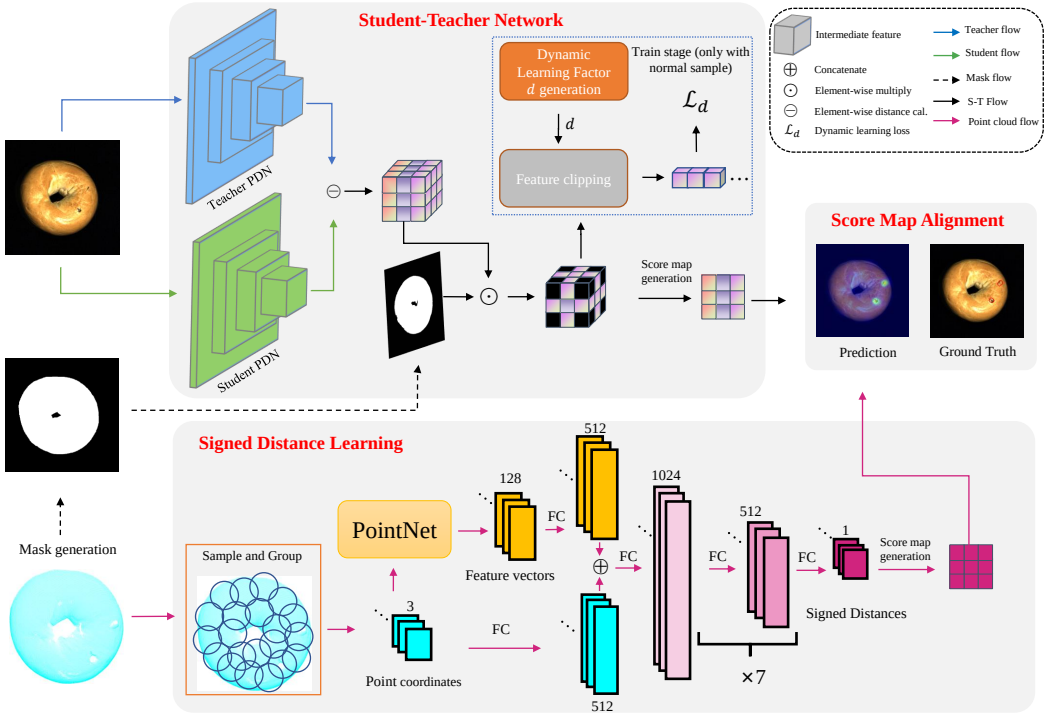


Figure 3. The framework of our method MDSS.

To be specific, the student–teacher network is trained with normal RGB images and masks generated from point clouds by a dynamic loss, and the anomaly score map can be obtained from the discrepancy between the output of the student network and teacher network. Moreover, in signed distance learning, the signed distance function learns from normal point clouds to predict the signed distances between points and the surface, and the obtained signed distances are used to generate the anomaly score map. Subsequently, the two previously obtained anomaly score maps are aligned to generate the final anomaly score map for detection.

In the following, Section 3.1 provides the details of our student–teacher network. Section 3.2 presents the process of signed distance learning. Finally, anomaly score map alignment will be introduced in Section 3.3.

3.1. Student–Teacher Network

Student–teacher networks have been widely used in 2D anomaly detection. In MDSS, the structure of the student network is the same as the teacher. Particularly, a lightweight student–teacher network PDN (Patch Description Network) [18] is selected to detect the anomaly in RGB images. The PDN is fully convolutional, which contains four convolution layers and some pooling layers for a very low overall latency. Therefore, PDN could be applied to an image of variable size to generate all feature vectors in a single forward pass. Each output neuron of PDN has a receptive field of 33 × 33 pixels, and thus, each output feature vector describes a 33 × 33 patch, which prevents the anomaly information in different patches from interfering with each other and helps localize context-related anomalies.

In the proposed method MDSS, a training image I is applied to the Teacher T and Student S, and corresponding features $T(I) \in \mathbb{R}^{C\times H\times W}$ and $S(I) \in \mathbb{R}^{C\times H\times W}$ are obtained. The square difference of each tuple (c, w, h) is computed as $D_{c,w,h} = (T(I)_{c,w,h} - S(I)_{c,w,h})^2$. Moreover, as the object is presented in a 3D perspective with a static background, it is straightforward and reasonable to remove the irrelevant background, which is the case for almost all real applications. Therefore, a binary mask M is generated from 3D point clouds for extracting the foreground of the object. Then, the element-wise multiplication $\odot$ is applied to D and M for setting the output elements belonging to the background as zero.

Moreover, excessive training may cause the student to mimic the output of the teacher beyond normal samples, thereby hindering the detection accuracy [18]. PDN employs a hard feature loss for training, which uses the output elements with the highest loss for back-propagation to encourage the student to focus on emulating the most underfitting regions. However, they ignore that the proportion of these regions will change dynamically during the training process. Furthermore, the fixed proportion may not be appropriate for different datasets. Therefore, a dynamic learning factor d is proposed in MDSS to solve this problem.

Specifically, the dynamic learning factor d varies between 0.99 and 0.999 in a cosine annealing [43] way. Given a dynamic learning factor d, the d-quantile of the elements in $D \odot M$ is computed, and the elements larger than d-quantile are averaged as our dynamic loss, referred to as $\mathcal{L}_d$ and calculated in Equation (2).

$$\mathcal{L}_d = \frac{1}{n}\sum_{i=1}^{n} DM_i \tag{2}$$

where $DM_i \in \{D^{'} \mid D^{'} \in D \odot M, D^{'} > d\text{-quantile}\}$, n represents the number of elements in the set, and $\odot$ stands for the element-wise multiplication.

In the inference stage, the trained student–teacher network is employed to generate a corresponding anomaly score map for a new image. Particularly, the distance between the output of the teacher and the student is first calculated and averaged along the channel dimension to obtain the anomaly score map. Then, the score map is resized to the size of the input image with bilinear interpolation. Each pixel value in the map represents the

likelihood of an anomaly pixel, and the maximum value of the score map is regarded as the anomaly score on the image level.

3.2. Signed Distance Learning

Chu et al. [7] introduce the signed distance function (SDF) for 3D anomaly detection and propose the shape-guided method. In the shape-guided method, SDF is applied to the point clouds, and the SDF features are stored in the memory bank for inference. SDF is a continuous function to output the distance of a point to the closest surface, in which the sign represents whether the point is inside or outside the watertight mesh, and the underlying surface boundary is implicitly represented by the zero-level set with the distance being zero. Due to the spatial locality of the occurrence of the anomaly, shape-guided employs PointNet [44] and Multilayer Perceptron (MLP) [7,39] to obtain the local geometry representation and store them in the memory bank. Our SDF uses the same structure as the shape-guided method.

However, in our opinion, if the SDF is trained only with normal samples, the model will be learned to predict the distances of normal points to the implicit surface. Therefore, in inference, the distance from abnormal points to the surface is expected to be larger than that from normal points, and it can be used as the anomaly score. Therefore, MDSS directly uses SDF for 3D anomaly detection without the usage of a memory bank. The signed distance is taken as the anomaly score. Particularly, MDSS trains the SDF model for each category, respectively, and the output signed distance is directly used to detect anomalies. We have demonstrated our assumption through experiments conducted on the bagel class in the MVTec 3D-AD dataset. The corresponding result is provided in Figure 4.

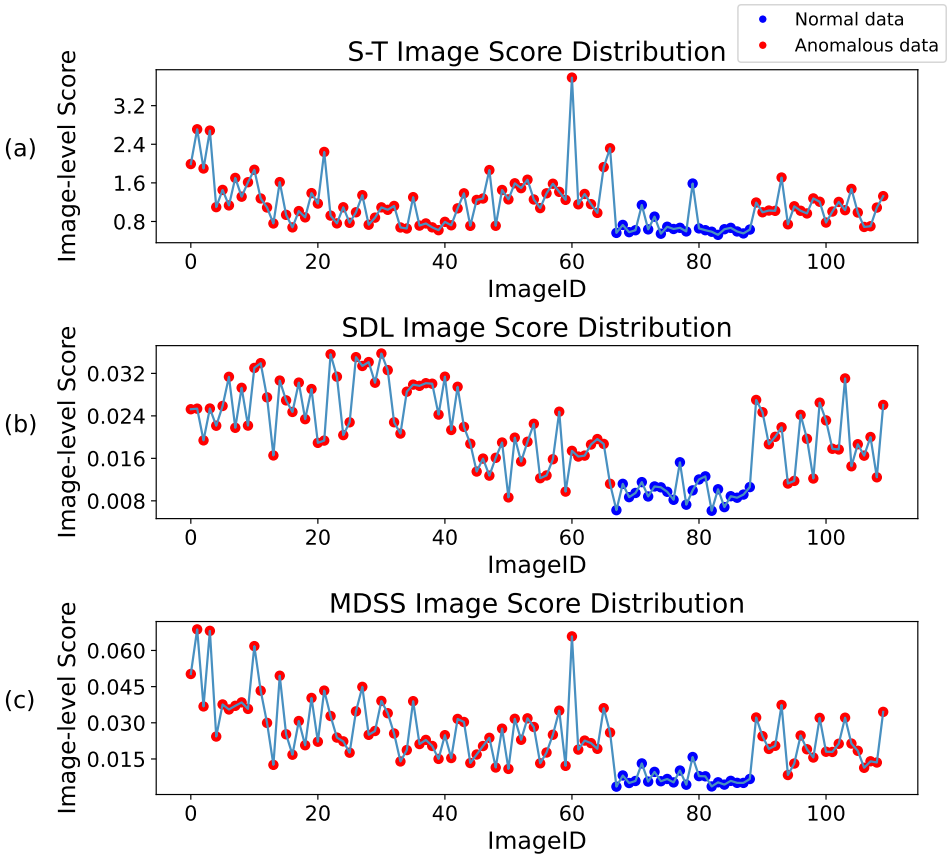


Figure 4. Image-level anomaly score distribution of bagel. (**a**) The distribution of image-level anomaly score of S-T. (**b**) The distribution of image-level anomaly score of SDL. (**c**) The distribution of image-level anomaly score of MDSS.

Figure 4a depicts the distribution of the output distance between the student and the teacher network, which we regard as the anomaly score. The output of the signed distance function, i.e., the signed distance, is taken with its absolute value to serve as the anomaly score, and Figure 4b illustrates its distribution. Figure 4c shows the distribution of the aligned anomaly scores. In Figure 4a, the S-T image score of normal and anomalous data is difficult to distinguish since many of them are concentrated around 0.8. From Figure 4b, we can observe that the anomaly scores for normal samples are generally smaller than those for anomalous samples. This suggests that the distance can be used as an anomaly score to distinguish normal samples from anomalous ones. However, the scores of some abnormal data are still close to normal data, even closer than some normal data, such as the data with the imageID around 50 or 60. In Figure 4c, we can see that, after the two anomaly scores are aligned and combined, the distinction between normal and anomalous samples is improved. Almost all the scores of normal data are lower than those of the anomalous samples. Therefore, the usage of SDF in MDSS is reasonable, and our alignment strategy, which will be elaborated on in the next section, is effective.

For the inference stage, a new point cloud sample is passed into the SDF model, and the signed distances between points and the surface are output. With the corresponding 2D index, the signed distances of each point can be assigned to a pixel in the anomaly score map. Furthermore, Gaussian blur is applied to the anomaly score map for improving the relevance between the anomaly point and its neighbors. Similar to Section 3.1, the maximum value of the score map is regarded as the anomaly score on the image level.

3.3. Score Map Alignment

Our method, MDSS, is built upon the capability of these two models and their collaborative nature to more effectively tackle the challenge of multimodal anomaly detection. To be specific, the student–teacher network considers the RGB information to identify any appearance irregularities in the aspect of color, and the signed distance learning utilizes 3D information to probe possible anomalies in shape geometry. As mentioned in Sections 3.1 and 3.2, the two models will, respectively, generate an anomaly score map. The product of the maximum of these two maps is used for image-level anomaly detection.

However, for pixel-level anomaly detection (also referred to as anomaly segmentation), the anomaly scores in the two score maps should be firstly transformed into a similar scale due to their significant numerical gap. In MDSS, a statistical method is used to align them, which is similar to the shape-guided method [7]. Specifically, the validation set is employed to simulate the distribution of the two score maps in real scenarios. Then, we compute the mean value and standard deviation of RGB and 3D anomaly score maps, respectively. During the inference stage, the RGB score map for a new sample will be aligned to its 3D score map with the previously obtained mean value and standard deviation, such that the mean $\pm$ 3 $\times$ standard deviation of the RGB score map is aligned to the mean $\pm$ 3 $\times$ standard deviation of the 3D score map. The pixel-wise maximum of two score maps is selected to construct the final anomaly score map for anomaly segmentation.

4. Experiments

4.1. Dataset

The proposed method MDSS is validated for its effectiveness with the popular MVTec 3D-AD [5] dataset in the experimental study. This dataset is the first publicly available multimodal anomaly detection dataset and comprises 10 categories, including natural objects and industrial components. Particularly, MVTec 3D-AD contains 2656 training, 294 validation, and 1197 test samples, where the test data are split into 249 normal samples and 948 abnormal samples. The abnormal test samples include 4 to 5 different types of defects in each category.

To be specific, each category is represented by both RGB images and high-resolution 3D point clouds. The 3D point clouds are obtained using structured light from an industrial sensor and store position information in 3-channel tensors (x, y, and z coordinates), while

RGB information is recorded for each point. Since all samples in the dataset are viewed from the same angle, the RGB information for each sample can be stored in a single image. Furthermore, the label and pixel-level ground truth are also provided to conduct image-level detection and pixel-level detection.

The overview of the MVTec 3D-AD dataset is listed in Table 1.

Table 1. Dataset statistics for different categories in MVTec 3D-AD.

Category	Train	Val	Test (Good)	Test (Anomalous)	Defect Types	Annotated Regions	Image Size (Width × Height)
bagel	244	22	22	88	4	112	800 × 800
cable gland	223	23	21	87	4	90	400 × 400
carrot	286	29	27	132	5	159	800 × 800
cookie	210	22	28	103	4	128	500 × 500
dowel	288	34	26	104	4	131	400 × 400
foam	236	27	20	80	4	115	900 × 900
peach	361	42	26	106	5	131	600 × 600
potato	300	33	22	92	4	115	800 × 800
rope	298	33	32	69	3	72	900 × 400
tire	210	29	25	87	4	95	600 × 800
Total	2656	294	249	948	41	1148	

4.2. Evaluation Metrics

As is common for anomaly detection, we adopt the area under the receiver operating characteristic curve (AUROC) to evaluate the detection performance of our method at the image level (I-AUROC). For segmentation evaluation, the per-region overlap (PRO) metric [45] is employed, which is defined as the average relative overlap of the binary prediction with each connected component of the ground truth. Similar to I-AUROC, the area under PRO curve (AUPRO) is computed to evaluate the pixel-level detection performance. Note that both metrics range from 0 to 1, and higher values indicate better performance.

4.3. Implementation Detail

In the experiment, the background planes of the point clouds are removed according to the method in [9]. Then, all of the point clouds are cut into different patches and each patch includes 500 points. Specifically, a set of points from the original points is sampled with the farthest point sampling [46] and the K-nearest points to each of them are searched to construct a patch. In the present experiment, we set a default value of K to 500, similar to in [7]. Note that the patches may overlap with each other, and each point should belong to at least one patch.

Furthermore, inspired by [10], the point clouds are used to generate a corresponding mask M for RGB images. Specifically, if there is a non-zero pixel in point cloud images, we set it to 1 at the same position in the mask for the foreground and otherwise 0 for the background. In order to fill the missing values, the foreground mask is dilated using a square structural element of size 8. Both of the point clouds and RGB images are resized to 256×256.

For the student–teacher network, we utilize a learning rate of 0.001 and a batch size of 4. For the signed distance function, we employ a cosine annealing learning rate with a batch size of 32. The batch size and learning rate are selected based on preliminary experiments for optimal performance.

4.4. Experimental Results

4.4.1. Detection Performance

In the experiment, we compare MDSS with many different methods on 10 categories of MVTec 3D-AD, including memory bank-based methods (BTF [9], M3DM [11], shape-guided [7], ITNM [13], CPMF [15]), student–teacher network-based methods (3D-ST [6],

AST [10], MMRD [14]), and other methods (EasyNet [12], CFM [8] and so on). Table 2 and Table 3, respectively, provide the detection and segmentation performance in terms of I-AUROC and AUPRO. Note that in addition to multimodal detection results, we also provide the corresponding detection results, respectively, based on RGB images and 3D point clouds in Tables 2 and 3.

Table 2. Image-level anomaly detection performance in terms of I-AUROC. The blue circled numbers represent the performance ranking of MDSS among the corresponding methods.

	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
RGB	SPADE [34]	0.771	0.793	0.760	0.531	0.848	0.683	0.646	0.460	0.879	0.502	0.687
	FastFlow [29]	0.624	0.472	0.654	0.694	0.501	0.667	0.595	0.632	0.816	0.731	0.639
	DifferNet [26]	0.859	0.703	0.643	0.435	0.797	0.790	0.787	0.643	0.715	0.590	0.696
	PADiM [32]	0.975	0.775	0.698	0.582	0.959	0.663	0.858	0.535	0.832	0.760	0.764
	STFPM [30]	0.930	0.847	0.890	0.575	0.947	0.766	0.710	0.598	0.965	0.701	0.793
	CSFlow [28]	0.941	0.930	0.827	0.795	0.990	0.886	0.731	0.471	0.986	0.745	0.830
	RD4AD [31]	0.975	0.987	0.943	0.575	0.999	0.830	0.863	0.618	0.984	0.899	0.867
	BTF [9]	0.876	0.880	0.791	0.682	0.912	0.701	0.695	0.618	0.841	0.702	0.770
	AST [10]	0.947	0.928	0.851	0.825	0.981	0.951	0.895	0.613	0.992	0.821	0.880
	M3DM [11]	0.944	0.918	0.896	0.749	0.959	0.767	0.919	0.648	0.938	0.767	0.850
	Shape-guided [7]	0.911	0.936	0.883	0.662	0.974	0.772	0.785	0.641	0.884	0.706	0.815
	EasyNet [12]	0.982	0.992	0.917	0.953	0.919	0.923	0.840	0.785	0.986	0.742	0.904
	MMRD [14]	0.987	0.937	0.943	0.770	0.981	0.847	0.913	0.753	0.993	0.853	0.898
	MDSS	0.915	0.894	0.907	0.780	0.963	0.793	0.869	0.743	0.953	0.856	0.867 ⑤
3D	3D-ST [6]	0.862	0.484	0.832	0.894	0.848	0.663	0.763	0.687	0.958	0.486	0.748
	BTF [9]	0.825	0.551	0.952	0.797	0.883	0.582	0.758	0.889	0.929	0.653	0.782
	AST [10]	0.881	0.576	0.965	0.957	0.679	0.797	0.990	0.915	0.956	0.611	0.833
	M3DM [11]	0.941	0.651	0.965	0.969	0.905	0.760	0.880	0.974	0.926	0.765	0.874
	Shape-guided [7]	0.983	0.682	0.978	0.998	0.960	0.737	0.993	0.979	0.966	0.871	0.916
	EasyNet [12]	0.735	0.678	0.747	0.864	0.719	0.716	0.713	0.725	0.885	0.687	0.747
	MMRD [14]	0.829	0.686	0.937	0.804	0.972	0.865	0.947	0.806	0.967	0.849	0.866
	MDSS	0.969	0.691	0.959	0.906	0.849	0.865	0.966	0.989	0.898	0.926	0.902 ②
RGB + 3D	BTF [9]	0.918	0.748	0.967	0.883	0.932	0.582	0.896	0.912	0.921	0.886	0.865
	AST [10]	0.983	0.873	0.976	0.971	0.932	0.885	0.974	0.981	1.000	0.797	0.937
	M3DM [11]	0.994	0.909	0.972	0.976	0.960	0.942	0.973	0.899	0.972	0.850	0.945
	Shape-guided [7]	0.986	0.894	0.983	0.991	0.976	0.857	0.990	0.965	0.960	0.869	0.947
	EasyNet [12]	0.991	0.998	0.918	0.968	0.945	0.945	0.905	0.807	0.994	0.793	0.926
	ITNM [13]	0.992	0.951	0.988	0.950	0.999	0.876	0.919	0.965	0.991	0.850	0.948
	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 [8]	0.994	0.888	0.984	0.993	0.980	0.888	0.941	0.943	0.980	0.953	0.954
	CPMF [15]	0.983	0.889	0.989	0.991	0.958	0.809	0.988	0.959	0.979	0.969	0.952
	MDSS	0.983	0.911	0.984	0.927	0.955	0.962	0.973	0.978	0.962	0.930	0.956 ①

Table 3. Pixel-level anomaly detection performance in terms of AUPRO. The blue circled numbers represent the performance ranking of MDSS among the corresponding methods.

	Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
RGB	PADiM [32]	0.980	0.944	0.945	0.925	0.961	0.792	0.966	0.940	0.937	0.912	0.930
	CFlow [27]	0.855	0.919	0.958	0.867	0.969	0.500	0.889	0.935	0.904	0.919	0.871
	BTF [9]	0.901	0.949	0.928	0.877	0.892	0.563	0.904	0.932	0.908	0.906	0.876
	M3DM [11]	0.952	0.972	0.973	0.891	0.932	0.843	0.970	0.956	0.968	0.966	0.942
	Shape-guided [7]	0.946	0.972	0.960	0.914	0.958	0.776	0.937	0.949	0.956	0.957	0.933
	MMRD [14]	0.970	0.983	0.982	0.924	0.976	0.875	0.981	0.975	0.984	0.973	0.962
	MDSS	0.917	0.967	0.975	0.873	0.951	0.808	0.935	0.969	0.949	0.977	0.932 ④
3D	3D-ST [6]	0.950	0.483	0.986	0.921	0.905	0.632	0.945	0.988	0.976	0.542	0.833
	BTF [9]	0.973	0.879	0.982	0.906	0.892	0.735	0.977	0.982	0.956	0.961	0.924
	M3DM [11]	0.943	0.818	0.977	0.882	0.881	0.743	0.958	0.974	0.950	0.929	0.906
	Shape-guided [7]	0.974	0.871	0.981	0.924	0.898	0.773	0.978	0.983	0.955	0.969	0.931
	MMRD [14]	0.926	0.806	0.965	0.858	0.904	0.731	0.962	0.958	0.966	0.936	0.901
	MDSS	0.973	0.818	0.979	0.911	0.874	0.801	0.982	0.983	0.949	0.960	0.923 ③
RGB + 3D	BTF [9]	0.976	0.969	0.979	0.973	0.933	0.888	0.975	0.981	0.950	0.971	0.959
	M3DM [11]	0.970	0.971	0.979	0.950	0.941	0.932	0.977	0.971	0.971	0.975	0.964
	Shape-guided [7]	0.981	0.973	0.982	0.971	0.962	0.978	0.981	0.983	0.974	0.975	0.976
	ITNM [13]	0.980	0.973	0.982	0.947	0.981	0.967	0.981	0.983	0.982	0.974	0.975
	MMRD [14]	0.986	0.990	0.991	0.951	0.990	0.901	0.990	0.990	0.987	0.982	0.976
	CFM [8]	0.979	0.972	0.982	0.945	0.950	0.968	0.980	0.982	0.975	0.981	0.971
	CPMF [15]	0.958	0.946	0.979	0.868	0.897	0.746	0.980	0.981	0.961	0.977	0.929
	MDSS	0.979	0.968	0.981	0.949	0.958	0.969	0.982	0.983	0.970	0.978	0.972 ④

As shown in Table 2, our method MDSS obtains the state-of-the-art performance in terms of average I-AUROC over all categories for multimodal anomaly detection. To be

specific, MDSS achieves the mean I-AUROC with 0.956, outperforming the best student–teacher network method MMRD and the memory bank-based method CPMF by 0.6% and 0.4%, respectively. Compared with the latest method, CFM, also without a memory bank, MDSS still demonstrates superior performance.

From Table 3, it can be observed that MDSS obtains the mean AUPRO with 0.972 for anomaly segmentation in a multimodal setting, which is slightly worse than three of the methods (shape-guided, MMRD and ITNM) but better than the other four methods (including the latest method CFM). Moreover, the MMRD performance is unstable for different models. For example, MMRD ranks first for RGB images but is penultimate for 3D point clouds.

In summary, MDSS outperforms other SOTA methods for multimodal anomaly detection in terms of I-AUROC. For the purpose of comparing MDSS with other SOTA methods more comprehensively, Figure 5 provides a detailed comparison for each category.

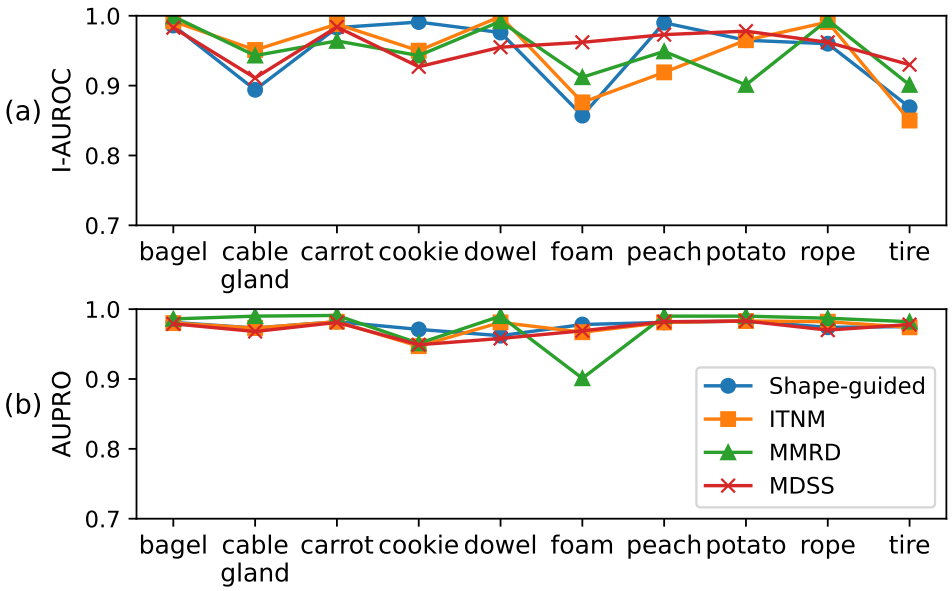


Figure 5. Comparison of MDSS with SOTA methods. (**a**) The image-level detection performance of different methods in terms of I-AUROC. (**b**) The pixel-level detection performance of different methods in terms of AUPRO.

From Figure 5, it can be observed that for anomaly segmentation, MDSS and other SOTA methods show relatively close performance in terms of AUPRO for different categories, which means that these methods are both stable and effective for anomaly segmentation. However, for anomaly detection, we can observe the performance fluctuation for other SOTA methods in terms of I-AUROC, indicating their instability regarding anomaly detection. This suggests that our method exhibits high robustness across various data types and is also more likely to be applicable to other unknown scenarios.

To conclude, MDSS is comparable to several SOTA methods, regardless of memory bank-based or student–teacher network-based methods, and furthermore, it performs better than other baseline methods in multimodal anomaly detection.

4.4.2. Ablation Study

Two ablation studies are conducted in the present study, including one to demonstrate the effectiveness of combining the student–teacher network with signed distance learning and another to show the effectiveness of using a mask. Figure 6 provides a detailed comparison of MDSS, a student–teacher (S-T) network and signed distance learning (SDL) for different categories in terms of I-AUROC and AUPRO. Figure 7 provides the average performance in all categories with and without a mask.

From Figure 6, it can be observed that MDSS usually performs better than S-T and SDL for most categories. Specifically, MDSS obtains the best I-AUROC for eight categories (except dowel and potato) and achieves the best AUPRO for all categories. As is known to all, some anomalies may manifest in terms of color while others may manifest in terms of 3D geometry, combining the two types of modality information for detecting anomalies will be more effective. In Figure 7, with the mask added to the model, both I-AUROC and AUPRO are higher compared to that without the mask.

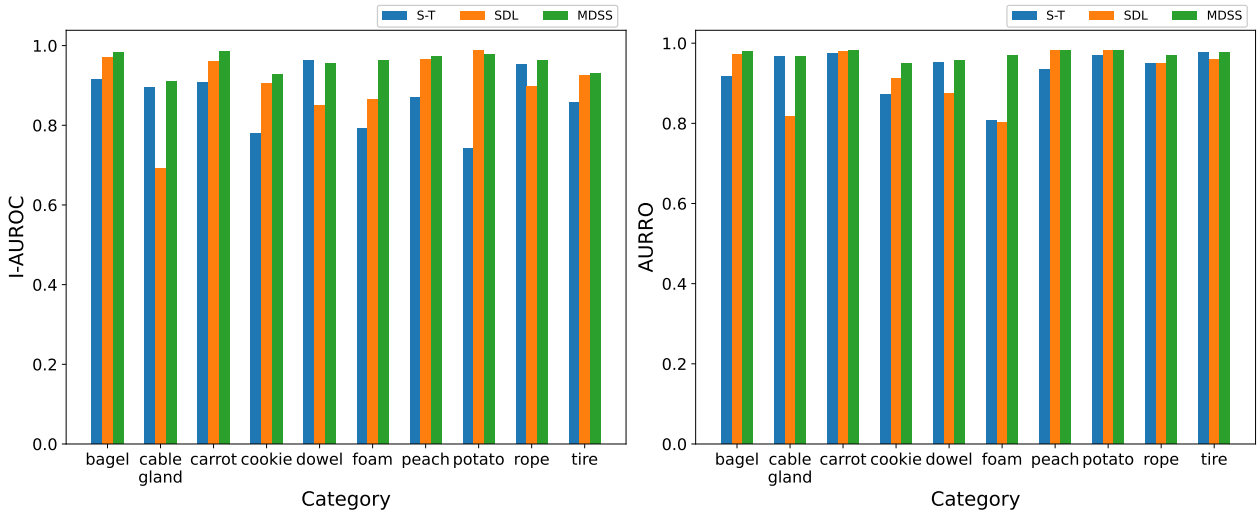


Figure 6. Ablation study for MDSS. **Left** figure shows I-AUROC of S-T, SDL, and MDSS. **Right** figure shows AUPRO of S-T, SDL, and MDSS.

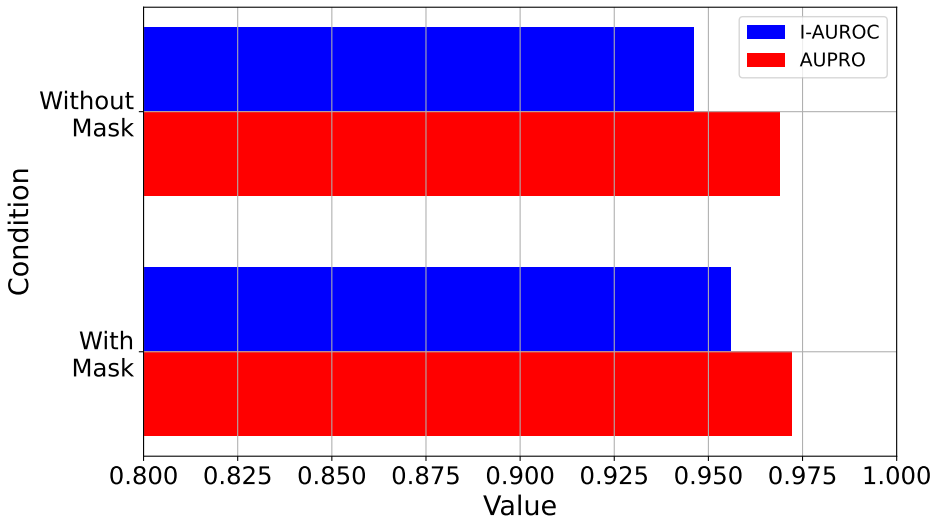


Figure 7. The average performance on all categories with and without a mask.

In conclusion, MDSS activates the interaction between information from the two modalities and usually performs better than the two individual modules in MDSS, namely student–teacher network and signed distance learning. In addition, adding the mask in MDSS could improve the performance of anomaly detection and segmentation.

4.4.3. Hyper-Parameter Analysis

We test the sensitivity of our method MDSS to the hyper-parameter, specifically the dynamic learning factor d, and search for the optimal parameter combination. Figure 8 provides the corresponding values of I-AUROC and AUPRO for the dynamic learning factor and some fixed learning factors. The horizontal axis represents I-AUROC, while

the vertical axis represents AUPRO, both ranging from 0 to 1. This means that the closer a point is to the top-right corner, the better the model performs. In Figure 8, the different points correspond to the performance of various learning factors, including 0, 0.5, 0.75, 0.9, 0.95, 0.99, 0.999, and 0.9999. In addition, "dynamic" in Figure 8 indicates the performance of our dynamic learning factor. In Figure 8, the average I-AUROC increases as the dynamic learning factor d increases from 0.9 to 0.999, but decreases when d is further increased to 0.9999. We then apply a cosine annealing strategy [43] to adjust d from 0.99 to 0.999, which achieves the best performance in both detection and segmentation. This demonstrates the effectiveness of the dynamic learning strategy in MDSS.

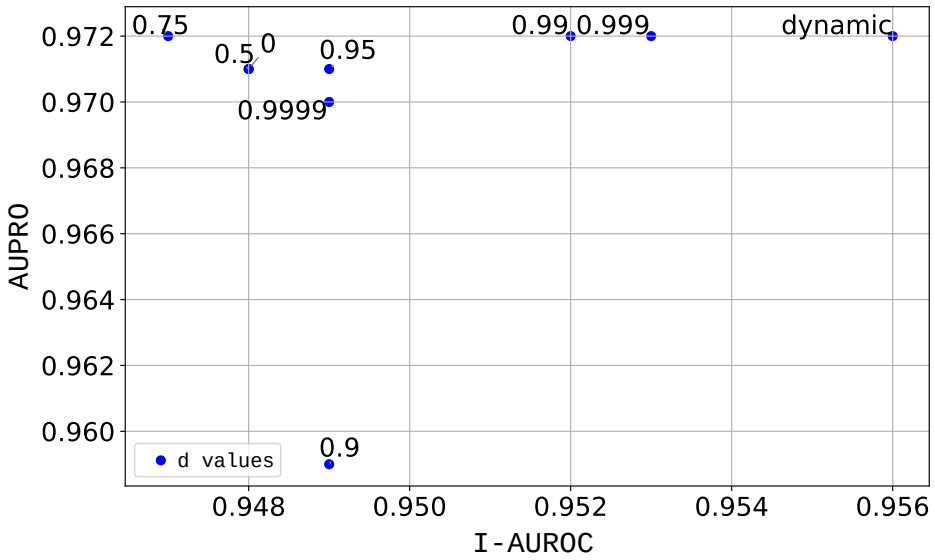


Figure 8. Performance of MDSS with different learning factor d.

4.4.4. Alignment Strategies Comparison

In the present study, we validate three different alignment strategies [7,11,18] to combine the output from different modalities. In addition to the alignment strategy proposed by the shape-guided method [7] that we used in MDSS, M3DM [11] applies two learnable One-Class Support Vector Machines (OCSVM) on the concatenated score map to integrate the results of two modalities. We imitate them and concatenate the anomaly score maps of two modalities to train two OCSVMs for anomaly detection and segmentation. In EfficientAD [18], for each of the two anomaly map types, they compute the set of all pixel anomaly scores across the validation images and determine a linear transformation that maps the 0.9 quantile to an anomaly score of 0 and the 0.995 quantile to a score of 1. We perform the same operation and average the anomaly score maps of the two modalities by pixel. The averaged map and its maximum value are used for anomaly detection and segmentation, respectively.

For clarity, the alignment strategies described above have been assigned the following names for reference: Concat-OCSVM Alignment for the approach using concatenated score maps and OCSVMs [11], Quantile-Interval Alignment for the quantile-based linear transformation [18], and Mean-Std Alignment for the method employed in MDSS. These names are used for convenience in this work. The performances of different alignment strategies are shown in Figure 9.

As shown in Figure 9, the Mean-Std Alignment obtains the best performance, and the Concat-OCSVM Alignment obtains the worst performance. The most significant drawback of the Concat-OCSVM Alignment strategy is to train a OCSVM and the concatenating operation may cause a disturbance between modalities, which may be the reason why the corresponding performance is not good. Moreover, the Quantile-Interval Alignment strategy is hindered by the need to select two quantiles, which requires extensive

experimentation for selecting appropriate quantiles and is challenging to apply across various datasets.

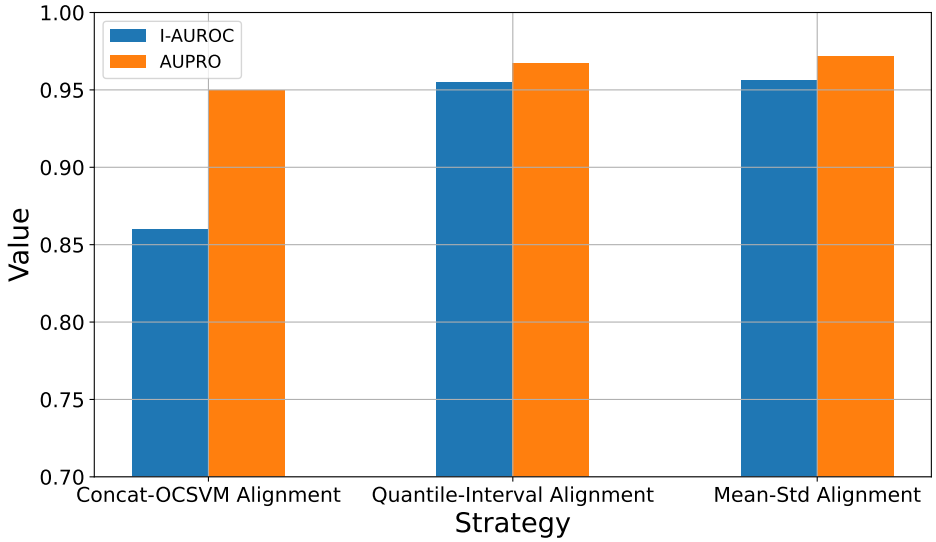


Figure 9. Average performance on all categories when using different alignment strategies.

5. Conclusions

In the present study, a novel memoryless multimodal anomaly detection method MDSS is proposed, which includes three different modules, namely student–teacher network, signed distance learning and score map alignment. Specifically, the student–teacher network aims to learn RGB images and masks generated from 3D point clouds to obtain the RGB anomaly score map. In signed distance learning, we employ the signed distance function to reconstruct the surface from normal point clouds and adopt the distance from the point to the surface to generate a corresponding 3D anomaly score map. Finally, a statistical approach is employed to align the RGB anomaly score map and the 3D anomaly score map, and then, the aligned score map, which combines the anomaly information from both RGB and 3D, is used for anomaly detection. The experimental results with the popular MVTec 3D-AD dataset demonstrate that MDSS is comparable but more stable than SOTA methods and, furthermore, performs better than other baseline methods. Our method demonstrates that for multimodal anomaly detection, performance is still on the rise without the usage of memory banks, making it more suitable for some real-world application scenarios.

However, we find that MDSS underperforms in certain categories like "dowel" and "potato". This phenomenon may be because our method conducts the posterior fusion of the two modalities after training and may be unsuitable for these types of datasets. Therefore in the forthcoming research, we will explore whether interacting information from different modalities during the training process can achieve better performance for such categories. Moreover, the present study does not explore the robustness of MDSS and we intend to employ more datasets from various scenarios to validate the robustness of MDSS in the future.

Author Contributions: Funding acquisition, Z.S.; investigation, Z.S. and X.L.; methodology, Z.S. and X.L.; supervision, Z.S.; validation, X.L.; writing—original draft, Z.S. and X.L.; writing—review and editing, Y.L. and Y.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Fundamental Research Funds for the Central Universities under Grant No. 2021QN1075.

Data Availability Statement: This study used open data sources.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ruff, L.; Kauffmann, J.R.; Vandermeulen, R.A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, T.G.; Müller, K.R. A unifying review of deep and shallow anomaly detection. *Proc. IEEE* **2021**, *109*, 756–795. [CrossRef]
2. Siddique, M.F.; Ahmad, Z.; Kim, J.M. Pipeline leak diagnosis based on leak-augmented scalograms and deep learning. *Eng. Appl. Comput. Fluid Mech.* **2023**, *17*, 2225577. [CrossRef]
3. Siddique, M.F.; Ahmad, Z.; Ullah, N.; Ullah, S.; Kim, J.M. Pipeline Leak Detection: A Comprehensive Deep Learning Model Using CWT Image Analysis and an Optimized DBN-GA-LSSVM Framework. *Sensors* **2024**, *24*, 4009. [CrossRef] [PubMed]
4. Liu, J.; Xie, G.; Wang, J.; Li, S.; Wang, C.; Zheng, F.; Jin, Y. Deep industrial image anomaly detection: A survey. *Mach. Intell. Res.* **2024**, *21*, 104–135. [CrossRef]
5. Bergmann, P.; Jin, X.; Sattlegger, D.; Steger, C. The MVTec 3D-AD dataset for unsupervised 3d anomaly detection and localization. In Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Virtual, 6–8 February 2022; pp. 202–213.
6. Bergmann, P.; Sattlegger, D. Anomaly detection in 3d point clouds Using deep geometric descriptors. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 2613–2623.
7. Chu, Y.M.; Liu, C.; Hsieh, T.I.; Chen, H.T.; Liu, T.L. Shape-guided dual-memory learning for 3d anomaly detection. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 6185–6194.
8. Costanzino, A.; Ramirez, P.Z.; Lisanti, G.; Di Stefano, L. Multimodal industrial anomaly detection by crossmodal feature mapping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 17234–17243.
9. Horwitz, E.; Hoshen, Y. Back to the feature: Classical 3d features are (almost) all you need for 3d anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2967–2976.
10. Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; Wandt, B. Asymmetric student-teacher networks for industrial anomaly detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 2592–2602.
11. Wang, Y.; Peng, J.; Zhang, J.; Yi, R.; Wang, Y.; Wang, C. Multimodal industrial anomaly detection via hybrid fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 8032–8041.
12. Chen, R.; Xie, G.; Liu, J.; Wang, J.; Luo, Z.; Wang, J.; Zheng, F. Easynet: An easy network for 3d industrial anomaly detection. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 7038–7046.
13. Wang, J.; Wang, X.; Hao, R.; Yin, H.; Huang, B.; Xu, X.; Liu, J. Incremental Template Neighborhood Matching for 3D anomaly detection. *Neurocomputing* **2024**, *581*, 127483. [CrossRef]
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. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 8445–8453.
15. Cao, Y.; Xu, X.; Shen, W. Complementary pseudo multimodal feature for point cloud anomaly detection. *Pattern Recog.* **2024**, *156*, 110761. [CrossRef]
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 9650–9660.
18. Batzner, K.; Heckler, L.; König, R. EfficientAD: Accurate visual anomaly detection at millisecond-level latencies. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 1–6 January 2024; pp. 128–138.
19. Béthune, L.; Novello, P.; Boissin, T.; Coiffier, G.; Serrurier, M.; Vincenot, Q.; Troya-Galvis, A. Robust One-Class Classification with Signed Distance Function using 1-Lipschitz Neural Networks. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 2245–2271.
20. Sun, Z.; Li, X.; Li, Y.; Ma, Y. Memoryless Multimodal Anomaly Detection via Student-Teacher Network and Signed Distance Learning. *arXiv* **2024**, arXiv:2409.05378.
21. Schlegl, T.; Seeböck, P.; Waldstein, S.M.; Schmidt-Erfurth, U.; Langs, G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In Proceedings of the International Conference on Information Processing in Medical Imaging, Boone, NC, USA, 25–30 June 2017; pp. 146–157.
22. Schlegl, T.; Seeböck, P.; Waldstein, S.M.; Langs, G.; Schmidt-Erfurth, U. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Med. Image Anal.* **2019**, *54*, 30–44. [CrossRef] [PubMed]
23. Kristan, M.; Zavrtanik, V.; Skočaj, D. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognit.* **2021**, *112*, 107706.
24. Zavrtanik, V.; Kristan, M.; Skočaj, D. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 8330–8339.

25. Zavrtanik, V.; Skočaj, D.; Kristan, M. Dsr–a dual subspace re-projection network for surface anomaly detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 539–554.
26. Rudolph, M.; Wandt, B.; Rosenhahn, B. Same same but differNet: Semi-supervised defect detection with normalizing Flows. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 1907–1916.
27. Gudovskiy, D.; Ishizaka, S.; Kozuka, K. 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, Waikoloa, HI, USA, 3–8 January 2022; pp. 98–107.
28. Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; Wandt, B. Fully convolutional cross-scale-flows for image-based defect detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 1088–1097.
29. Yu, J.; Zheng, Y.; Wang, X.; Li, W.; Wu, Y.; Zhao, R.; Wu, L. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv* **2021**, arXiv:2111.07677.
30. Wang, G.; Han, S.; Ding, E.; Huang, D. Student-teacher feature pyramid matching for anomaly detection. In Proceedings of the 32nd British Machine Vision Conference, Online, 22–25 November 2021; p. 306.
31. Deng, H.; Li, X. Anomaly detection via reverse distillation from one-class embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9737–9746.
32. Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. Padim: A patch distribution modeling framework for anomaly detection and localization. In Proceedings of the International Conference on Pattern Recognition, Milano, Italy, 10–15 January 2021; pp. 475–489.
33. Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; Gehler, P. Towards total recall in industrial anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14318–14328.
34. Cohen, N.; Hoshen, Y. Sub-image anomaly detection with deep pyramid correspondences. *arXiv* **2020**, arXiv:2005.02357.
35. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
36. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4183–4192.
37. Reiss, T.; Cohen, N.; Bergman, L.; Hoshen, Y. PANDA: Adapting pretrained features for anomaly detection and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 2806–2814.
38. Rusu, R.B.; Blodow, N.; Beetz, M. Fast point feature histograms (FPFH) for 3D registration. In Proceedings of the 2009 IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 3212–3217.
39. Ma, B.; Liu, Y.S.; Zwicker, M.; Han, Z. Surface reconstruction from point clouds by learning predictive context priors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6326–6337.
40. Park, J.J.; Florence, P.; Straub, J.; Newcombe, R.; Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 165–174.
41. Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; Geiger, A. Occupancy networks: Learning 3d reconstruction in function space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4460–4470.
42. Chen, Z.; Zhang, H. Learning implicit fields for generative shape modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5939–5948.
43. Loshchilov, I.; Hutter, F. SGDR: Stochastic gradient descent with Warm restarts. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
44. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
45. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9592–9600.
46. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5099–5108.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.