Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uaai20

Anomaly Detection Using Siamese Network with Attention Mechanism for Few-Shot Learning

Hironori Takimoto, Junya Seki, Sulfayanti F. Situju & Akihiro Kanagawa

To cite this article: Hironori Takimoto, Junya Seki, Sulfayanti F. Situju & Akihiro Kanagawa (2022) Anomaly Detection Using Siamese Network with Attention Mechanism for Few-Shot Learning, Applied Artificial Intelligence, 36:1, 2094885, DOI: 10.1080/08839514.2022.2094885

To link to this article: https://doi.org/10.1080/08839514.2022.2094885

Published online: 11 Jul 2022.

APPLIED ARTIFICIAL INTELLIGENCE
2022, VOL. 36, NO. 1, e2094885 (2946 pages)
https://doi.org/10.1080/08839514.2022.2094885

Taylor & Francis
Taylor & Francis Group

Anomaly Detection Using Siamese Network with Attention Mechanism for Few-Shot Learning

Hironori Takimoto[a†], Junya Seki[b†], Sulfayanti F. Situju[c], and Akihiro Kanagawa[a]

[a]Faculty of Computer Science and Systems Engineering, Okayama Prefectural University, Okayama, Japan; [b]Graduate School of Computer Science and Systems Engineering, Okayama Prefectural University, Okayama, Japan; [c]Faculty of Engineering, Sulawesi Barat University, West Sulawesi, Indonesia

ABSTRACT

Automated inspection using deep-learning has been attracting attention for visual inspection at the manufacturing site. However, the inability to obtain sufficient abnormal product data for training deep- learning models is a problem in practical application. This study proposes an anomaly detection method based on the Siamese network with an attention mechanism for a small dataset. Moreover, attention branch loss (ABL) is proposed for Siamese network to render more task-specific attention maps from attention mechanism. Experimental results confirm that the proposed method with the attention mechanism and ABL is effective even with limited abnormal data.

ARTICLE HISTORY
Received 19 April 2022
Revised 17 June 2022
Accepted 21 June 2022

Introduction

Visual inspection is an integral part of the manufacturing process that prevents the release of defective products into the market. Manual inspection requires time, resources, and costs and lengthy visual inspections are burdensome for the inspectors. In addition, abnormalities in product appearance are difficult to identify clearly, and unification of judgment criteria among multiple inspectors is difficult. Therefore, to solve these problems, there is a need to develop a nonmanual visual inspection method for anomaly detection (Mumtaz, Mansoor, and Masood 2012).

Automation of visual inspection, such as defect inspection and anomaly detection of industrial products, is a crucial task in the field of computer vision. Detection of anomalies, such as micro-scratches on product surfaces, in image data are critical in many industries. With the development and popularization of convolutional neural networks (CNNs) (Lecun et al. 1998) based on deep learning, the practical application of CNN-based anomaly detection has been extensively studied in the past few years

CONTACT Hironori Takimoto takimoto@c.oka-pu.ac.jp Faculty of Computer Science and Systems Engineering, Okayama Prefectural University, Soja, Okayama, Japan

†These authors contributed equally to this work

This article has been republished with a minor change. This change do not impact on the academic content of the article.

(Akcay et al., 2018, 2019; Andrews et al. 2016; Bergmann et al. 1993; Cha, Choi, and Buyukzturk 2017; Eisenbach et al. 2017; Gong et al. 2019; Haselmann and Gruber 2019; Haselmann, Gruber, and Tabatabai 2018; Katafuchi et al., 2021; Ma, Xie, and Zhang 2019; Masci et al. 2012; Mei, Yang, and Yin 2018; Ren, Hung, and Tan 2017; Sakurada et al., 2014; Schlegl et al. 2017; Soukup and Huber-Mork 2014; Tang et al. 2020; Weimer, Scholz-Reiter, and Shpitalni 2016; Zenati et al. 2018). CNN-based anomaly detection can be divided into two approaches: classification-based and reconstruction-based.

Based on supervised learning, the classification-based method is a simple categorization task applied for anomaly detection. The model is generally a binary classification task that outputs a two-class classification result (normal or abnormal) for the input evaluation image. Although this approach is simple and can be applied in various areas, it requires large quantities of data including not only normal data, but considerable abnormal data as well. However, it is difficult to collect a large quantity of abnormal data as training images in a real environment, leading to a severe imbalanced dataset problem with more normal data than anomaly data. Moreover, it is also difficult to apply a classification-based method using a CNN for anomaly detection because it does not perform well in the low-data regime. On the other hand, research on unsupervised anomaly detection using generative models has attracted attention. This approach is the most popular method for anomaly detection because it does not require labeled anomalous data for training anomaly detectors. Model training without anomalous data is an optimal approach because preparing a large quantity of abnormal data is difficult in a real environment. However, the auto-encoder (AE) (Hinton and Salakhutdinov 2006) and generative adversarial network (GAN) (Goodfellow et al. 2014) output immoderate blurry structures because of the failure in reconstructing fine structures at times.

Few-shot earning is a subfield of machine learning that aims to create models that can learn the desired objective with fewer data, similar to human learning. While most classification-based methods require training on massive datasets, few-shot learning aims to learn the features of the object categories from one or only a few training images. Architectures for learning from a small dataset have been extensively researched (Jadon 2020).

Deep learning-based few-shot learning approaches can be divided into four main categories: data augmentation, metric-based, model-based, and optimization-based methods. The Siamese network (Bromley et al. 1994) is a representative metric-based method. It consists of twin networks and can use the relationship between pairs of input samples for learning. The twin networks are identical, sharing the same weights and network parameters; both refer to the same embedding network that learns efficient embedding to reveal the relationship between pairs of data points.

Although it is difficult to collect a large quantity of anomaly data, a smaller quantity can be collected. Thus, it is expected that the obtained anomaly data can be used effectively to achieve highly-accurate anomaly detection. The Siamese network learns the optimal embedding space based on pairs of data; hence, the few available anomaly data can be used efficiently for training.

Attention is used in a wide range of deep-learning applications and is an epoch-making technology in the rapidly developing field of natural language. In computer vision tasks using deep learning, attention is a mechanism to dynamically identify where the input data should be focused. It is used to improve the accuracy of the task by focusing the model on the important parts of the image (Jetley et al. 2018). The attention branch network (ABN) (Fukui et al. 2019) has been proposed as a typical classification task-specific CNN model that introduces a learnable attention mechanism. It extends the top-down visual explanation model by introducing a branch structure with an attention mechanism. By introducing a branch for attention, the ABN simultaneously achieves visualization of the gazing area through visual explanation and improves the model accuracy. The most important feature of the ABN is that the model is trained to classify images precisely using only the feature map data obtained from the attention mechanism. Thus, the attention mechanism is trained to render a more task-specific attention map.

In this study, a method based on deep-metric learning and an attention mechanism is proposed for improving the accuracy of CNN-based anomaly detection under a situation where only very few anomaly data are available. An anomaly detection model that can be efficiently and effectively trained even with a small quantity of anomaly data is first constructed using deep- metric learning combined with a Siamese network and CNN. Further, to improve the accuracy of the proposed method, an attention mechanism is applied to the feature extractor of the proposed method. With reference to the ABN concept, only the feature map obtained from the attention mechanism is used for learning to construct an optimum embedding space with the Siamese network. The attention mechanism is trained with the aim of rendering the attention map itself more task-specific to move away different classes of data in the embedding space. Furthermore, we propose pair balanced contrastive loss (PBCL) to account for the effect of training with unbalanced data on the CL, which is used to train the Siamese network. Experimental results using the benchmark dataset ”MVTec AD (Bergmann et al. 2019)” confirm that the proposed method improves the anomaly detection performance.

Related Work

Compared to the classical machine vision methods, deep-learning methods achieve automatic end-to-end learning of rules, which contribute to scientific decision-making, from the input data through network learning. The powerful

feature representation capabilities of deep-learning methods are well suited for detecting complex defects. Existing methods for anomaly detection using deep-learning can be approximately divided into classifier-based and reconstruction-based methods. The methods relevant to our work are described in brief in the following subsections.

Classifier-Based Anomaly Detection

Classification-based surface-defect detection for products has been widely proposed (Andrews et al. 2016; Cha, Choi, and Buyukzturk 2017; Eisenbach et al. 2017; Katafuchi et al., 2021; Ma, Xie, and Zhang 2019; Masci et al. 2012; Ren, Hung, and Tan 2017; Soukup and Huber-Mork 2014; Weimer, Scholz-Reiter, and Shpitalni 2016). Classification-based methods are simple classification tasks that apply a CNN for exception detection based on supervised learning. The binary classification task outputs a two-class classification result, generally normal or abnormal, for evaluating the input image. Cha et al. used two types of CNNs for detecting building damage, including cracks (Cha, Choi, and Buyukzturk 2017). Soukup et al. trained a classical CNN model to detect rail metal surface defects in a purely supervised manner (Soukup and Huber-Mork 2014). They showed that the CNN clearly outperformed traditional model-based anomaly detection. Ma et al. proposed automatic blister defect detection using a CNN with a dense block to ensure the quality and reliability of polymer lithium-ion batteries (Ma, Xie, and Zhang 2019). They added trainable weight parameters to each skip-connection for improving the dense blocks in the CNN architecture. Ren et al. adopted a generic approach for automated surface inspection (Ren, Hung, and Tan 2017). Their method extracts patch features using a pretrained CNN and then predicts the defect area by generating a defect heat map based on the patch features. However, this technique is only applicable to surface defects with localized anomalies within a homogeneous texture. Katafuchi et al. proposed a layer-wise external attention network (LEA-Net) for color anomaly detection tasks (Katafuchi et al., 2021). Their contribution is the integration of unsupervised and supervised anomaly detectors via the visual attention mechanism. Although they claimed that their proposed model improves the accuracy even on unbalanced data sets, training the model requires a relatively large quantity of anomaly data.

Classification-based methods are intended to be trained on relatively large data sets with little bias among the categories. However, it is difficult to collect considerable abnormal data as training images in a real environment. In addition, certain defect types occur very rarely. Anomaly detection datasets are heavily imbalanced, and often contain only a few anomalies for model verification and testing. Therefore, supervised

learning remains a challenge in many situations where only a few defective image patches are available along with thousands of normal image patches (Haselmann and Gruber 2019).

Reconstruction-Based Anomaly Detection

The requirement for sufficient abnormal samples can be eliminated using reconstruction-based methods trained in an unsupervised manner on exclusively normal data (Akcay et al., 2018, 2019; Bergmann et al. 1993; Gong et al. 2019; Haselmann, Gruber, and Tabatabai 2018; Mei, Yang, and Yin 2018; Sakurada et al., 2014; Schlegl et al. 2017; Tang et al. 2020; Zenati et al. 2018). This approach generally uses an AE and GAN to learn powerful reconstruction manifolds using only normal data.

The AE attempts to learn the identity function using only normal data (Hinton and Salakhutdinov 2006). It is difficult for an AE-based model to represent anomalous image structures because the model reconstructs the input image after compressing the image into a low-dimensional embedding. Thus, the model can reconstruct a plausible normal image even if an image containing anomalies is input. Therefore, anomaly detection is achieved by comparing the input image with the reconstructed image.

Mei et al. proposed a method to localize and detect using only defect-free data for model training (Mei, Yang, and Yin 2018). This method is based not only on the reconstruction of image patches with a convolutional denoising AE at different Gaussian pyramid levels but also the synthesis of the detection results from these different-resolution channels. Bergmann proposed autoencoding architectures that used pixel-wise reconstruction error metrics as unsupervised defect segmentation (Bergmann et al. 1993). Gong et al. proposed MemAE with a memory module for mitigating a problem encountered when using a simple deep autoencoder network (Gong et al. 2019).

GAN is an unsupervised learning neural network (Goodfellow et al. 2014) that learns to generate a new image with a probability distribution similar to that of the training data. To design the loss function of the neural network (generators and discriminators compete) for training, the network uses game theory.

Schlegl et al. proposed AnoGAN, which is a GAN-based anomaly detection network for images (Schlegl et al. 2017). By training the GAN using only normal samples, a model that can generate fake images with probability distributions similar to those of the normal samples is constructed. The network can then classify anomaly samples by defining the threshold of the residual score between the test and fake images. Akcay et al. proposed GANomaly, which is a conditional GAN including an extra encoder for the extraction of meaningful latent variables from

images (Akcay et al., 2018). In addition, they proposed skip-GANomaly, an improvement of GANomaly (Akcay et al., 2019). Skip-GANomaly achieves superior image reconstruction by adding a skip-connection architecture to GANomaly.

Tao et al. (Tao et al. 2022) proposed the reconstruction-based detection method for the location of anomalies using the Siamese architecture. This model was trained using only normal samples to focus on a detection task of an unknown abnormal sample. Although this method has different tasks and requisites than ours, it achieved more accurate anomaly location detection than other state-of-the-art methods.

A disadvantage of the reconstruction-based method is that the CNN-based autoencoder or generator outputs immoderate blurry structures because it sometimes fails to reconstruct fine structures. Moreover, it requires considerable computing resources.

Materials and Methods

Deep Metric Learning Using a Siamese Network and CNN

The metric learning problem, which involves learning a distance function tuned to a particular task, is beneficial when used in conjunction with nearest-neighbor methods and other techniques that rely on distances or similarities (Jadon 2020). In this study, we focus on the Siamese network (Bromley et al. 1994), a type of metric learning, and propose a visual inspection method based on deep metric learning combined with a CNN.

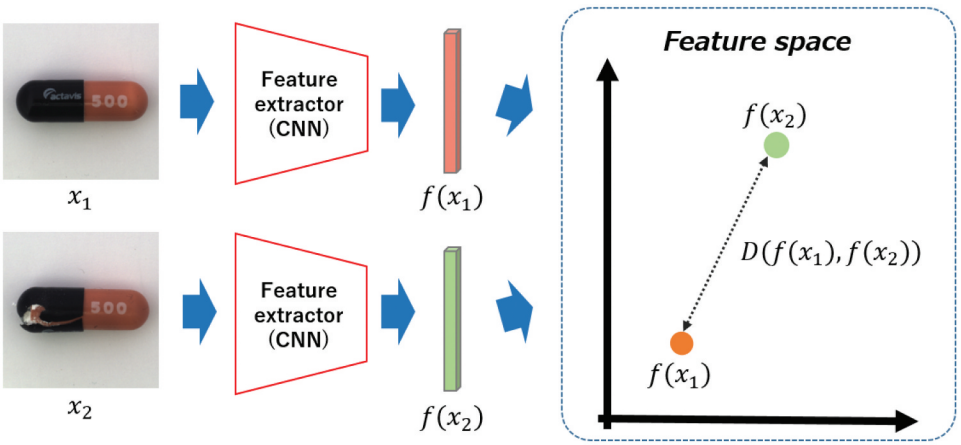

Figure 1. Overview of the siamese network using a CNN as the feature extractor.

Figure 1 presents an overview of the Siamese network. The basic concept of the Siamese network is the design of a loss function that directly pulls together the embedding of samples with the same label and pushes away those of samples with different labels. The contrastive loss (CL) defined by Equation (1) is generally used as the loss function:

$$CL(f(x_1), f(x_2)) = \frac{1}{2}\{YD^2 + (1 - Y)\max(m - D, 0)^2\} \tag{1}$$

$$D = ||f(x_1) - f(x_2)||_2 \tag{2}$$

where x_1 and x_2 are an input data pair, $f(\cdot)$ is the feature extractor, and m is a margin value greater than zero. Y is a flag indicating whether the input data pairs are in the same category (1 for the same class and 0 for different classes).

CL, which is a typical loss function for metric learning (Chopra, Hadsell, and LeCun 2005), is one of the most straightforward and intuitive training objectives. The main idea of using a Siamese network is not to classify the classes but to learn to discriminate between inputs. The network learns the distances between data using the feature vectors obtained from the feature extractor. The advantage of the Siamese network is that it can learn even when the number of original data is small because pairs of data are used for training. With a Siamese network architecture based on few-shot learning, the network can generate a feature space in which normal and abnormal data are separated by learning the normal data and a few abnormal data.

In this study, VGG16 (Simonyan and Zisserman 2014) is used as the feature extractor of the Siamese network. VGG is designed based on the fundamental concept that deeper networks are better, and has smaller filters than AlexNet (Krizhevsky, Sutskever, and Hinton 2012). Here, each filter has a size of 3×3 albeit with a lower stride of one, and effectively captures a receptive field identical to that captured by a 7×7 filter with four strides. In the VGG16 used in the proposed method, the image features related to anomaly detection are extracted from the input image, mainly in the convolutional layers (convolution (1–1) – convolution (5–4)) listed in Table 1. Note that a 224×224-pixel RGB 3CH color image is used as the input image. This feature extractor outputs a 512-dimensional feature vector, based on the attention and feature maps, as output-1.

Attention Mechanism for the CNN

When training a CNN model to handle images, the task accuracy can be improved by focusing the model on the important parts of the image. One of the methods for accomplishing this involves the usage of a learnable attention mechanism, which has been attracting considerable interest in computer

Table 1. The structure of CNN model for image aesthetic assessment.

Layer type	Kernel\Stride	Activation	Output size
Convolution(1–1)	3 × 3\1	ReLU	224 × 224 × 64
Convolution(1–2)	3 × 3\1	ReLU	224 × 224 × 64
Max Pooling	2 × 2\2	–	112 × 112 × 64
Convolution(2–1)	3 × 3\1	ReLU	112 × 112 × 128
Convolution(2–2)	3 × 3\1	ReLU	112 × 112 × 128
Max Pooling	2 × 2\2	–	56 × 56 × 128
Convolution(3–1)	3 × 3\1	ReLU	56 × 56 × 256
Convolution(3–2)	3 × 3\1	ReLU	56 × 56 × 256
Convolution(3–3)	3 × 3\1	ReLU	56 × 56 × 256
Convolution(3–4)	3 × 3\1	ReLU	56 × 56 × 256
Max Pooling	2 × 2\2	–	28 × 28 × 256
Convolution(4–1)	3 × 3\1	ReLU	28 × 28 × 512
Convolution(4–2)	3 × 3\1	ReLU	28 × 28 × 512
Convolution(4–3)	3 × 3\1	ReLU	28 × 28 × 512
Convolution(4–4)	3 × 3\1	ReLU	28 × 28 × 512
Max Pooling	2 × 2\2	–	14 × 14 × 512
Convolution(5–1)	3 × 3\1	ReLU	14 × 14 × 512
Convolution(5–2)	3 × 3\1	ReLU	14 × 14 × 512
Convolution(5–3)	3 × 3\1	ReLU	14 × 14 × 512
Convolution(5–4)	3 × 3\1	ReLU	14 × 14 × 512
Max Pooling	2 × 2\2	–	7 × 7 × 512
Global Average Pooling	–	–	1 × 1 × 512

vision (Jetley et al. 2018). The attention mechanism is generally defined as the process of refining or enhancing the image features for a recognition task. In the human perceptual system, information relevant to the current task tends to be preferentially incorporated. The attention mechanism essentially mimics this to extract features for image classification.

In anomaly detection by a Siamese network, the quality of the generated feature space affects the accuracy of anomaly detection. Therefore, we added the attention mechanism to the feature extractor in our Siamese network model. Figure 2 shows the proposed network with an attention mechanism introduced in VGG16. The green dashed line in the figure means that the model or layer shares the same parameters. The structure of the proposed

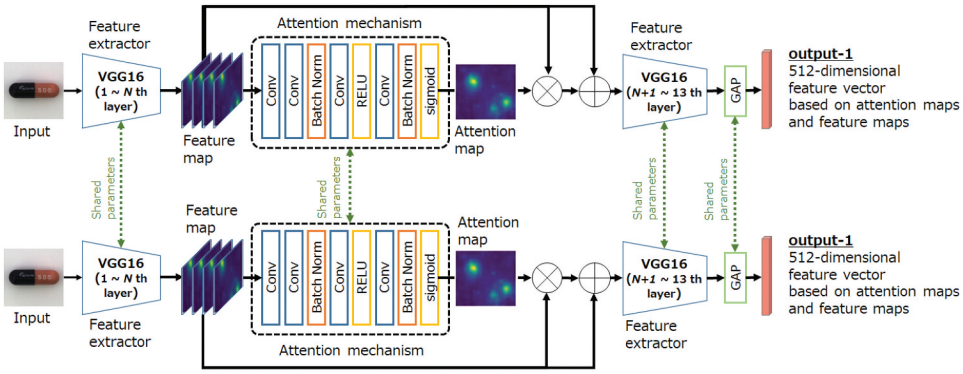

Figure 2. Architecture of the proposed feature extractor using VGG16 with an attention mechanism.

Table 2. Structure of the attention mechanism.

Layer type	Kernel\Stride	Output size
Convolution(a-1)	3 × 3\1	$h \times w \times d$
Convolution(a-2)	3 × 3\1	$h \times w \times d$
Batch Normalization	–	–
Convolution(a-3)	1 × 1\1	$h \times w \times d$
ReLU	–	–
Convolution(a-4)	1 × 1\1	$h \times w \times 1$
Batch Normalization	–	–
Sigmoid	–	$h \times w \times 1$

attention mechanism is listed in Table 2. Here, we incorporated the attention mechanism behind a single arbitrary layer of VGG16. Therefore, in Table 2, the number of channels (filters) d and the size $h \times w$ of the input feature map are indicated as variables.

On the other hand, Fukui et al. proposed an attention branch network (ABN) as a visual explanation method for image classification tasks (Fukui et al. 2019). The ABN extends the top-down visual explanation model by introducing a branch structure with an attention mechanism. By introducing a branch for attention, the ABN simultaneously achieves visualization of the gazing area through visual explanation and improves the model accuracy. With attention branch loss (ABL) in the ABN, the model is trained to precisely classify images using only the feature-map data obtained from the attention mechanism. Thus, the attention mechanism is trained to render the attention map itself more task-specific.

Therefore, to train attention maps to be more task-specific, we introduced an attention branch in the attention mechanism based on the ABN concept. Figure 3 shows the proposed network with an attention mechanism and

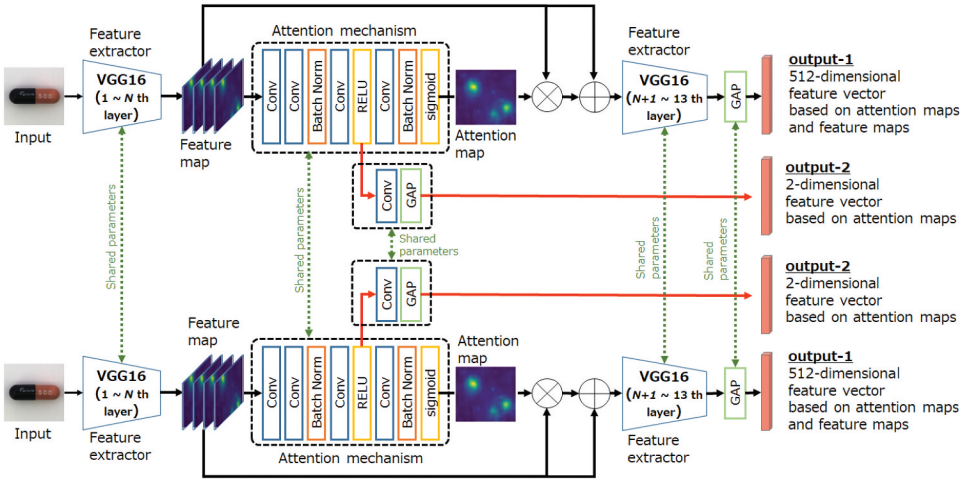

Figure 3. Architecture of the proposed feature extractor using VGG16 with an attention mechanism and attention branch.

Table 3. Structure of the attention branch.

Layer type	Kernel\Stride	Output size
Convolution	1 × 1\1	$h \times w \times d$
Global Average Pooling	–	1 × 1 × 2

attention branch. The green dashed line in the figure means that the model or layer shares the same parameters. In the proposed model, the attention branch is set after applying the ReLU activation function to convolution layer (a-3) of the attention mechanism. The structure of the proposed attention branch is depicted in Table 3. The size of the feature map input to the attention branch is $h \times w \times d$. This attention branch outputs a two-dimensional feature vector, mainly based on the attention maps, as additional output-2.

Although the ABN is trainable end-to-end using the losses at both branches, it is based on a classification model using a large dataset. This study differs from the ABN concept because it focuses on deep metric learning using the Siamese network and small datasets. Therefore, it is difficult to use the cross-entropy loss as in the ABN for training the model. To address this issue, our training loss function L_{all} is defined as the simple sum of the losses at both branches and is expressed by Equation (3):

$$L_{all} = L_{CL} + w * L_{ABL} \tag{3}$$

where L_{CL} is the CL calculated using output-1 and L_{ABL} is the ABL calculated using output-2, as depicted in Figure 3; w is a weighting parameter of the ABL.

Pair Balance Contrastive Loss (PBCL)

For anomaly detection, it is difficult to collect the same quantity of anomalous data as normal data, resulting in a biased and unbalanced dataset. Although a Siamese network can be efficiently trained even when data are scarce, accuracy degradation is a concern when a biased dataset is used for training data pairs. Therefore, we propose the PBCL, in which the number of pairs is considered, and each term of the CL is weighted. The three types of pairs used in this study during training were [normal, normal], [abnormal, abnormal], and [normal, abnormal]. The PBCL, which introduces a ratio of the numbers of data pairs belonging to each of the three types, is defined by Equation (4).

$$PBCL(f(x_1), f(x_2)) =$$

Table 4. Data in each category in dataset A used in the experiment.

Mode	Category	Capsule	Screw	Carpet	Tile
Train	Normal	80	80	80	80
	Abnormal	40	40	40	40
Test	Normal	61	71	45	40
	Abnormal	61	71	45	40

$$\frac{1}{2}\{NYD^2 + \alpha(1-N)YD^2 + \beta N(1-Y)\max(m-D,0)^2\} \quad (4)$$

where α and β are the weighting parameters determined by considering the ratio of the numbers of pairs. N is set to unity if the pair contains normal data and zero otherwise.

Experimental Setup

The MVTec AD dataset contains defect-free and anomalous images of various object and texture categories. From this dataset, we used "capsule" and "screw" as the object categories, and "carpet" and "tile" as the texture categories. Figure 4 shows an example of each product.

In this study, we evaluated the effectiveness of the proposed method by preparing two training datasets with different quantities of anomalous data. The details of both training datasets A and B are presented in Tables 4 and 5, respectively. Dataset A contains a total of four products, with two products each from the object and texture categories. In this dataset, the normal data in the training dataset are not considerable, and the abnormal data quantity is half that of the normal data. The abnormal data are limited. On the other

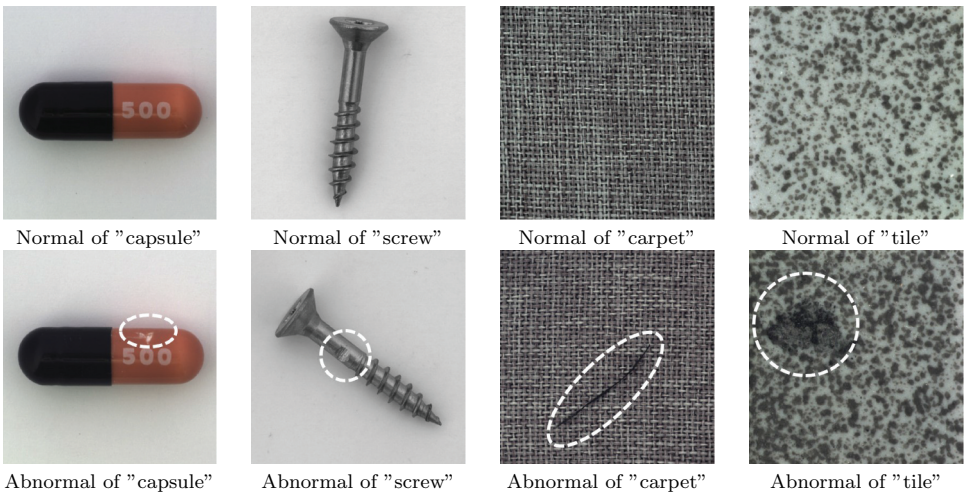

Figure 4. Examples of the MVtec AD images used in the experiment.

Table 5. Data in each category in dataset B used in the experiment.

Mode	Category	Capsule	Tile
Train	Normal	80	80
	Abnormal	20	20
Test	Normal	61	40
	Abnormal	61	40

Table 6. Number of data pairs in dataset A for training.

Pair type	# of pair
(Normal, Normal)	3,160
(Abnormal, Abnormal)	780
(Normal, Abnormal)	3,200

hand, dataset B contains two products, with a product each from the object and texture categories. In dataset B, the normal data quantity in the training dataset is similar to that of dataset A, but that of the abnormal data is half. Tables 6 and 7 list the number of data pairs used for training the Siamese network for each experiment.

As experiment A, for evaluating the effectiveness of the attention mechanism and ABL, experiments using dataset A were conducted with the following three feature extractors:

- Method I: VGG16 without an attention mechanism
- Method II: VGG16 with an attention mechanism
- Method III: VGG16 with an attention mechanism and ABL

Methods II and III are models that use the feature extractor shown in Figures 2 and 3, respectively. By incorporating an attention mechanism in one of the N-th (N=2,4,7,10,13) convolution layers of VGG16, we evaluated the differences in accuracy depending on the attention position. The ABL weight parameter was set as $w = \{1.0, 1.5\}$ and the accuracy was compared.

Table 7. Number of data pairs in dataset B for training.

Pair type	# of pair
(Normal, Normal)	3,160
(Abnormal, Abnormal)	190
(Normal, Abnormal)	1,600

Table 8. Results for "capsule" (experiment A).

Method	Model	Loss	w	N	AUC	Accuracy	Recall	Specificity
I	VGG16	CL	–	–	0.9397	0.867	**0.930**	0.803
II	VGG16 + Attention	CL	–	2	0.9409	0.865	0.910	0.820
				4	0.9406	0.854	0.898	0.810
				7	0.9373	0.861	0.915	0.807
				10	0.9433	0.859	0.916	0.802
				13	0.9381	0.854	0.928	0.780
III		CL + ABL	1.0	2	0.9464	0.871	0.908	0.833
				4	0.9486	0.867	0.878	**0.856**
				7	0.9316	0.853	0.891	0.816
				10	0.9302	0.840	0.902	0.777
				13	0.9331	0.853	0.876	0.830
			1.5	2	**0.9494**	0.869	0.910	0.828
				4	0.9477	0.869	0.884	0.853
				7	0.9429	**0.872**	0.898	0.846
				10	0.9330	0.860	0.874	0.846
				13	0.9114	0.837	0.890	0.777

Table 9. Results for "screw" (experiment A).

Method	Model	Loss	*w*	*N*	AUC	Accuracy	Recall	Specificity
I	VGG16	CL	–	–	0.8868	0.784	0.640	**0.986**
II		CL	–	2	0.8894	0.805	**0.690**	0.966
				4	0.8895	0.790	0.652	0.983
				7	0.8925	0.782	0.654	0.980
				10	0.8901	0.795	0.664	0.983
				13	0.8892	0.794	0.674	0.963
III	VGG16 + Attention	CL + ABL	1.0	2	0.9000	0.793	0.670	0.963
				4	**0.9013**	0.781	0.640	0.980
				7	0.8970	0.790	0.656	0.972
				10	0.8975	**0.809**	0.684	0.983
				13	0.8886	0.794	0.666	0.972
			1.5	2	0.8866	0.767	0.620	0.975
				4	0.8959	0.785	0.650	0.975
				7	0.8995	0.802	0.678	0.983
				10	0.8997	0.807	0.680	**0.986**
				13	0.8863	0.794	0.678	0.958

Table 10. Results for "carpet" (experiment A).

Method	Model	Loss	*w*	*N*	AUC	Accuracy	Recall	Specificity
I	VGG16	CL	–	–	0.9335	0.953	**1.000**	0.849
II		CL	–	2	0.9388	0.952	**1.000**	0.844
				4	0.9342	0.952	**1.000**	0.844
				7	0.9323	**0.966**	**1.000**	**0.889**
				10	0.9352	0.946	**1.000**	0.827
				13	0.9354	0.954	**1.000**	0.853
III	VGG16 + Attention	CL + ABL	1.0	2	0.9388	0.956	**1.000**	0.858
				4	0.9393	0.959	0.980	0.871
				7	0.9352	0.964	**1.000**	0.884
				10	0.9346	0.949	**1.000**	0.836
				13	0.9562	0.955	**1.000**	0.853
			1.5	2	0.9505	0.946	**1.000**	0.827
				4	0.9387	0.961	**1.000**	0.876
				7	0.9356	**0.966**	**1.000**	**0.889**
				10	0.9346	0.948	**1.000**	0.831
				13	**0.9579**	0.948	**1.000**	0.831

In experiment B, we used dataset B to evaluate the effectiveness of the proposed PBCL. In the experiment, the CL or PBCL combined with the ABL based on Method III were compared as the loss function. The PBCL parameters were set as $\alpha = 4.0$ and $\beta = \{1.0, 2.0\}$. These parameters were optimized from the ratio of the numbers of training pairs in Tables 6 and 7. The weight parameter w of the ABL was set to 1.0 experimentally.

In the experiments, Adam was used as the optimizer and the learning rate was set to 10^{-5}. The number of epochs and batch size were set to 100 and 16, respectively.

We describe below the determination of the validation data as normal or abnormal. Two methods were used for judging the validation data. Anomaly detection was first performed using the nearest neighbor method, wherein the

validation data is judged by calculating the distance between it and all the training data in the feature space. The accuracy, recall, and specificity were used as the evaluation criteria. Recall is an index that indicates the ratio of normal data predicted as normal, whereas specificity is an index that indicates the ratio of abnormal data predicted as abnormal. Next, the center of gravity was obtained from the normal data set used for training as a representative point of the normal class. The area under the curve (AUC) was then obtained as a criterion by applying a threshold for the distance between the validation data and the representative point.

In all the experiments, anomaly detection was performed only in the 512-dimensional embedding space obtained as output-1 of the proposed model. The embedding space obtained as output-2 was used only for training the model and not for judgment during verification.

The MVTec AD dataset contains images of 15 different products. However, as mentioned above, we not only calculated the three evaluation metrics based on the nearest neighbor method but also analyzed the accuracy of the proposed method on the ABL and some parameters. Therefore, our experiments were performed with only four products because of the enormous time cost.

Results and Discussion

Tables 8–11 present the results for each category for experiment A; the best value of each evaluation index is indicated in bold. Method III with the attention mechanism and ABL is the best in almost all the categories. This suggests that the introduction of the attention mechanism improves the quality of the feature space.

In object categories, such as the capsule and screw, the accuracy is higher when the attention mechanism is incorporated in a relatively shallow layer ($N = 2, 4$). On the contrary, the accuracy is higher for texture categories, such

Table 11. Results for "tile" (experiment A).

Method	Model	Loss	w	N	AUC	Accuracy	Recall	Specificity
I	VGG16	CL	–	–	0.9681	0.785	0.590	0.980
II		CL	–	2	0.9596	0.808	0.636	0.980
				4	0.9590	0.795	0.614	0.975
				7	0.9563	0.793	0.600	0.985
				10	0.9701	0.792	0.604	0.980
				13	0.9739	0.797	0.624	0.970
III	VGG16 + Attention	CL + ABL	1.0	2	0.9685	0.786	0.592	0.980
				4	0.9640	0.769	0.552	0.985
				7	0.9636	0.776	0.572	0.980
				10	0.9739	0.814	0.648	0.980
				13	0.9691	**0.824**	**0.668**	0.980
			1.5	2	0.9621	0.789	0.602	0.975
				4	0.9751	0.755	0.534	0.975
				7	0.9756	0.789	0.602	0.970
				10	0.9801	0.821	0.656	0.985
				13	**0.9838**	0.800	0.604	**0.995**

Table 12. Results for "capsule" (experiment B).

Method	Loss	α	β	N	AUC	Accuracy	Recall	Specificity
III	CL	–	–	2	0.9526	0.807	0.984	0.630
	+			4	0.9424	0.799	0.982	0.616
	ABL			7	0.9414	0.769	0.984	0.554
	(w=1.0)			10	0.9481	0.769	0.984	0.554
				13	0.9320	0.760	0.992	0.528
	PBCL			2	**0.9686**	0.796	0.996	0.597
	+			4	0.9675	**0.817**	0.994	**0.639**
	ABL	4.0	1.0	7	0.9611	0.759	0.986	0.531
	(w=1.0)			10	0.9604	0.704	0.996	0.413
				13	0.9632	0.718	**1.000**	0.436
	PBCL			2	0.9628	0.741	0.996	0.485
	+			4	0.9594	0.764	0.994	0.534
	ABL	4.0	2.0	7	0.9640	0.764	0.994	0.508
	(w=1.0)			10	0.9643	0.708	0.996	0.420
				13	0.9497	0.747	0.998	0.495

as the carpet and tile, when the attention mechanism is incorporated in the deeper layers ($N = 10, 13$). This indicates that in the object category, it is possible to determine whether a target region is abnormal by comparing it with a relatively narrow surrounding, but in the texture category, it is necessary to determine whether a target-site is abnormal by comparing it with a wider range of features. Thus, the accuracy is improved by incorporating the attention mechanism at a deeper layer because not only the local features, but more global features are also judged to be important. These results suggest that the optimal position for introducing the attention mechanism depends on the location of the anomaly in the product and its definition.

The results for experiment B are presented in Tables 12 and 13. When CL is used as the loss function, it can be observed that the specificity decreases significantly as the number of anomalous product data in the training data decreases. However, using the PBCL as the loss function slightly improves the degradation of the accuracy. Although the Siamese

Table 13. Results for "tile" (experiment B).

Method	Loss	α	β	N	AUC	Accuracy	Recall	Specificity
III	CL			2	0.9601	0.830	0.824	0.845
	+			4	0.9638	0.811	0.784	0.880
	ABL	–	–	7	0.9776	0.851	0.830	0.905
	(w=1.0)			10	0.9765	**0.899**	**0.920**	0.845
				13	0.9820	0.851	0.850	0.855
	PBCL			2	0.9638	0.841	0.870	0.770
	+			4	0.9597	0.821	0.808	0.855
	ABL	4.0	1.0	7	0.9829	0.845	0.826	0.895
	(w=1.0)			10	0.9699	0.823	0.836	0.790
				13	**0.9854**	0.843	0.810	**0.925**
	PBCL			2	0.9654	0.844	0.854	0.820
	+			4	0.9579	0.819	0.808	0.845
	ABL	4.0	2.0	7	0.9821	0.851	0.842	0.875
	(w=1.0)			10	0.9639	0.870	0.916	0.755
				13	0.9785	0.873	0.890	0.830

network can be trained with limited data, it is confirmed that unbalanced data sets have a negative impact on the accuracy. Therefore, the PBCL, which considers the total number of pairs for each pair, is effective for training with unbalanced data sets.

In addition, the accuracy tends to be lower when $\beta = 2.0$ compared to $\beta = 1.0$. The term related to β separates the distance between different classes of data in the feature space. This result suggests that in the learning process of the Siamese network, the function that brings features of the same class closer may be more important than the function that separates the data of different classes.

Conclusions

This study proposed an abnormality detection method based on a Siamese network with an attention mechanism, for a small dataset. An ABL was proposed for the Siamese network to render the attention maps, from the attention mechanisms, more task-specific. Using the MVTec AD dataset, we confirmed that the proposed method with the attention mechanism and ABL was effective even with few abnormal data. It was suggested that the optimal position of the attention mechanism depended on the target product. Although the Siamese network could be trained with limited training data, it was established that imbalance in the number of abnormal data in the dataset reduced its accuracy. As a countermeasure, we confirmed that the PBCL, which considers the bias in the number of pairs used for training, was effective as a loss function.

In the proposed method, the embedding space obtained by output-2 used for ABL during training is not used for verification. As a future work, it is possible that the use of two embedding spaces in the verification process can improve accuracy. In addition, it is necessary to detect anomalous regions in the image using the attention map obtained from the attention mechanism. The proposed method does not focus on detecting the location of defects. However, in our future study, we will investigate whether visualization of the attention map obtained from the attention branch contributes to defect location detection. Furthermore, visualization of the basis for evaluation is expected to lead to discovering the causes of defects.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Hironori Takimoto http://orcid.org/0000-0002-8795-7109

References

Akçay, S., A. Atapour-Abarghouei, and T. P. Breckon. 2018. GANomaly semi-supervised anomaly detection via adversarial training. *Proc. of Asian Conference of Computer Vision, pp. 622–637, Perth, Australia, 2018. arXiv:1805.06725* 622–637—.

Akçay, S., A. Atapour-Abarghouei, and T. P. Breckon. 2019. Skip-GANomaly: skip connected and adversarially trained encoder-decoder anomaly detection. *Proc. of International Joint Conference on Neural Networks, pp. 1–8, Budapest, Hungary, 2019, arXiv:1901.08954* 1–8.

Andrews, J. T. A., T. Tanay, E. J. Morton, and L. D. Griffin 2016. "Transfer representation-learning for anomaly detection", Proc. of The 33rd International Conference on Machine Learning, New York, USA.

Bergmann, P., M. Fauser, D. Sattleegger, and C. Steger 2019. "MVTec AD: a comprehensive real-world dataset for unsupervised anomaly detection", Proc. of The IEEE Computer Society Conference on Computer Vision and Pattern Recognition, California, USA, 9592–600.

Bromley, J., I. Guyon, Y. Lecun, E. Sackinger, and R. Shah. 1993. Signature verification using a "Siamese" time delay neural network. Proc. of the 6th International Conference on Neural Information Processing Systems:737–44, Colorado, USA.

Bromley, J., I. Guyon, Y. Lecun, E. Sackinger, and R. Shah. 1994. Signature verification using a "Siamese" time delay neural network. *Advance Neural Information Processing Systems* 6:737–44.

Cha, Y. J., W. Choi, and O. Buyukzturk. 2017. Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering* 32 (5):361–78. doi:10.1111/mice.12263.

Chopra, S., R. Hadsell, and Y. LeCun 2005. "Learning a similarity metric discriminatively, with application to face verification", Proc. of The IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 539–46.

Eisenbach, M., R. Stricker, K. Debes, and H. M. Gross. 2017. Crack detection with an interactive and adaptive video inspection system. *Arbeitsgruppentagung Infrastrukturmanagement* 94:94–103.

Fukui, H., T. Hirakawa, T. Yamashita, and H. Fujiyoshi 2019. "Attention branch network: learning of attention mechanism for visual explanation", Proc. of The IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 10705–14

Gong, D., L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. V. D. Hengel 2019. "Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection", Proc. of The IEEE International Conference on Computer Vision, Seoul, South Korea, 1705–14.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. *Proceeding of Advances in Neural Information Processing Systems* 27:2672–80.

Haselmann, M., D. P. Gruber, and P. Tabatabai 2018. "Anomaly detection using deep learning based image completion", Proc. of 17th IEEE International Conference on Machine Learning and Applications, Orlando, Florida, USA, 1237–42.

Haselmann, M., and D. P. Gruber. 2019. Pixel-wise defect detection by CNNs without manually labeled training data. *Applied Artificial Intelligence* 33 (6):548–66. doi:10.1080/08839514.2019.1583862.

Hinton, G. E., and R. R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313 (5786):504–07. doi:10.1126/science.1127647.

Jadon, S. 2020. An overview of deep learning architectures in few-shot learning domain. *arXiv, arXiv:2008.06365*.

Jetley, S., N. A. Lord, N. Lee, and P. Torr 2018. "Learn to pay attention", Proc. of International Conference on Learning Representations, Vancouver CANADA.

Katafuchi, R., and T. Tokunaga. 2021. LEA-Net: layer-wise external attention network for efficient color anomaly detection. *arXiv, arXiv:2109.05493*.

Krizhevsky, A., I. Sutskever, and G. E. Hinton 2012. "ImageNet classification with deep convolutional neural networks", Proc, of The 25th International Conference on Neural Information Processing Systems, NY, USA, pp. 1106–14.

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86 (11):2278–324. doi:10.1109/5.726791.

Ma, L., W. Xie, and Y. Zhang. 2019. Blister defect detection based on convolutional neural network for polymer lithium-ion battery. *Applied Sciences* 9 (6):1085–99. doi:10.3390/app9061085.

Masci, J., U. Meier, D. Ciresan, J. Schmidhuber, and G. Fricout 2012. "Steel defect classification with max-pooling convolutional neural networks", Proc. of The IEEE International Joint Conference on Neural Networks, Brisbane, Australia, 1–6.

Mei, S., H. Yang, and Z. Yin. 2018. An unsupervised-learning-based approach for automated defect inspection on textured surfaces. *IEEE Transactions on Instrumentation and Measurement* 67 (6):1266–77. doi:10.1109/TIM.2018.2795178.

Mumtaz, R., A. B. Mansoor, and H. Masood. 2012. Computer aided visual inspection of aircraft surfaces. *Proceeding of International Journal of Image Processing* 6:38–53.

Ren, R., T. Hung, and K. C. Tan. 2017. A generic deep-learning-based approach for automated surface inspection. *IEEE Transactions on Cybernetics* 48 (3):929–40. doi:10.1109/TCYB.2017.2668395.

Sakurada, M., and T. Yairi 2014. "Anomaly detection using autoencoders with nonlinear dimensionality reduction", Proc. of the Machine Learning for Sensory Data Analysis, New York, USA, 4–11.

Schlegl, T., P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *Proc. of 25th International Conference on Information Processing in Medical Imaging, pp.146–157, North Carolina, USA, 2017, arXiv:1703.05921*.146–157.

Simonyan, K., and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *Proc. of 3rd International Conference on Learning Representations, ICLR 2015, pp.1-14, San Diego, CA, USA, 2015, arXiv:1409.1556*.

Soukup, D., and R. Huber-Mork 2014. Convolutional neural networks for steel surface defect detection from photometric stereo images, Proc. of International Symposium on Visual Computing, Las Vegas, NV, USA, 668–77.

Tang, T.-W., W.-H. Kuo, J.-H. Lan, C.-F. Ding, H. Hsu, and H.-T. Young. 2020. Anomaly detection neural network with dual auto-encoders gan and its industrial inspection applications. *Sensors* 20 (12):3336. doi:10.3390/s20123336.

Tao, X., D.-P. Zhang, W. Ma, Z. Hou, Z. Lu, and C. Adak 2022. "Unsupervised anomaly detection for surface defects with dual-siamese network", IEEE Trans. on Industrial Informatics, DOI: 10.1109/TII.2022.3142326

Weimer, D., B. Scholz-Reiter, and M. Shpitalni. 2016. Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. *Cirp Annals-manufacturing Technology* 65 (1):417–20. doi:10.1016/j.cirp.2016.04.072.

Zenati, H., C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar. 2018. Efficient GAN-based anomaly detection. *arXiv, arXiv:1802.06222*.