Article

Deep Morphological Anomaly Detection Based on Angular Margin Loss

Taehyeon Kim ☑, Eungi Hong ☑ and Yoonsik Choe * ☑

Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, Korea; pyomu@yonsei.ac.kr (T.K.); hek722@yonsei.ac.kr (E.H.)
* Correspondence: yschoe@yonsei.ac.kr

Abstract: Deep anomaly detection aims to identify “abnormal” data by utilizing a deep neural network trained on a normal training dataset. In general, industrial visual anomaly detection systems distinguish between normal and “abnormal” data through small morphological differences such as cracks and stains. Nevertheless, most existing algorithms emphasize capturing the semantic features of normal data rather than the morphological features. Therefore, they yield poor performance on real-world visual inspection, although they show their superiority in simulations with representative image classification datasets. To address this limitation, we propose a novel deep anomaly detection algorithm based on the salient morphological features of normal data. The main idea behind the proposed algorithm is to train a multiclass model to classify hundreds of morphological transformation cases applied to all the given data. To this end, the proposed algorithm utilizes a self-supervised learning strategy, making unsupervised learning straightforward. Additionally, to enhance the performance of the proposed algorithm, we replaced the cross-entropy-based loss function with the angular margin loss function. It is experimentally demonstrated that the proposed algorithm outperforms several recent anomaly detection methodologies in various datasets.

Keywords: anomaly detection; angular margin loss; self-supervised learning; metric learning; morphological transformation

1. Introduction

In data analysis, anomaly detection refers to the identification of outliers in a data distribution [1]. Several visual anomaly detection algorithms based on deep neural networks (DNNs) have been proposed, including variational autoencoders (VAEs), convolutional neural networks (CNNs), and generative adversarial networks (GANs). However, DNNs can merely access the “normal” class instances. Therefore, most studies focused on representing or extracting salient features from normal instances by utilizing various methodologies, such as low-dimensional embedding, data reconstruction, and self-supervised learning. Deep anomaly detection (DAD) methodologies primarily involve the extraction of the semantically salient features of “normal” images using DNNs. Hence, most studies reported excellent results on representative image classification datasets composed of semantically distinguishable classes, e.g., CIFAR-10 [2], Fashion-MNIST [3], and cats-and-dogs dataset [4]. Figure 1a shows several images that are semantically different from each other. Generally, the semantic difference in the image domain leads to large morphological differences, such as outline and texture. Therefore, if the criterion between “normal” and “abnormal” is defined by the semantic differences, DAD tries to extract semantically important features in the training procedure. However, in common real-world anomaly detection problems, the discriminant criterion between “abnormal” and “normal” images is defined by the small morphological differences such as cracks, stains, and noise, which cannot be described semantically. Figure 1b shows two morphologically different images. In general, the criteria to distinguish between “abnormal” and “normal” classes are based...
on small spatial differences in images. Therefore, previous DAD algorithms developed to capture semantic features are not suitable for morphological anomaly detection. To address this problem, DAD models that emphasize morphological features from “normal” images are required.

Figure 1. Visual description of semantic and morphological differences in images: (a) Semantic difference. Both images are sampled from the cats-and-dogs dataset [4]. The difference between “cat” and “dog” classes is called the semantic difference. Generally, the semantic difference concerns both the semantic and morphological differences in the spatial domain of the image. To understand this difference, DNN must learn salient semantic features, such as the orientation of an object and the relation between dominant parts of a target object; (b) morphological difference. Both images are sampled from the representative industrial visual anomaly detection dataset MVTec [5]. The difference between “good grid” and “broken grid” classes is called the morphological difference. The morphological difference usually does not involve the semantic difference. Therefore, DNN, learning these semantic features, often cannot understand morphological differences between these two images.

Self-supervised learning is one of the best learning mechanisms for guiding the DAD model to understanding morphological features of “normal” images. As a subset of unsupervised learning, self-supervised learning has been proposed to learn image features without using any human-annotated labels [6]. In particular, there is a proxy objective function that enables DNN to achieve the goal of the target application. To rephrase, with a properly designed self-supervised loss function, DNN can learn the feature that we are interested in, e.g., the morphological features of an image. Several methods for self-supervised learning-based DAD have been proposed [7,8].

Existing self-supervised learning-based DAD models train DNN to recognize the geometric transformation applied to an image received as the input, e.g., 2D rotation and geometric translation. Previous studies that demonstrate this straightforward task provide a powerful supervisory signal for semantic feature learning. Therefore, these previous semantic DAD models cannot maintain their robust performance in visual morphological anomaly detection problems. More specifically, for instance, to successfully predict the 2D rotation of an image, the DAD model must learn to (1) localize salient objects in an image and (2) recognize their orientation and object type. Subsequently, it must relate the object orientation with the dominant orientation that each object tends to depict within the available “normal” images. However, the DAD model, focusing on semantic features of “normal” images, is not suitable for the morphological anomaly detection problem, as depicted in Figure 1b. This is because, in most visual morphological anomaly detection problems, not only “normal” image does not include a salient object, but the discriminate criteria between “abnormal” and “normal” images are defined by the small differences in the spatial domain of an image. To address this limitation, in this study, we propose a DAD algorithm, which trains DNN to recognize morphological transformations applied to the instance that it receives as input, including dilation, erosion, and morphological gradient. In addition, we propose a novel objective function called the kernel size prediction loss, which leads the proposed DAD model to recognize the window size of the morphological transformation filter via the transformed image only. To define this loss as a classification
loss, we define several window sizes, which facilitates the proposed DAD to learn various morphological features from “normal” images.

Although the proposed DAD model learns morphological features of a “normal” image via self-supervised learning, several challenges remain in the training procedures, including the enhancement in the discriminative performance and the stabilization of the training process. Unlike semantic feature representation-based DAD models, the proposed algorithm must capture versatile and subtle morphological features on “normal” instances to quantify the abnormality of unobserved input instances. To address these limitations, the proposed DAD model adopts an angular margin loss (AML) to augment the softmax loss, which is widely used in previous self-supervised learning-based DAD models [7–9]. The softmax loss is suitable to optimize the inter-class difference but unable to reduce the intra-class variation. To enhance the discriminative power of softmax loss, several AMLs have been developed to minimize the intra-class variation. These AMLs force the classification boundary closer to the weight vector of each class and improve the softmax loss by combining various types of margins. Because the proposed DAD is based on classification tasks, AMLs can easily be combined without any additional process. Therefore, the proposed DAD has enhanced discriminative performance in morphological feature representation learning owing to AMLs enforce intra-class compactness and inter-class discrepancy simultaneously.

In essence, the main contributions of this study are as follows:

- A novel deep morphological anomaly detection model based on straightforward morphological transformations and AML is developed. The proposed algorithm can learn the morphological features of “normal” images intensively.
- Because the proposed algorithm is based on self-supervised learning, it represents and extracts salient features with supervised learning, which often guarantees easier convergence and lower computational cost compared to unsupervised learning-based DAD models.
- To combine self-supervised learning and morphological transformations, we propose a novel objective function called the kernel size prediction loss, enabling the DAD model to recognize the morphological filter size of morphologically transformed inputs.
- The performance of the proposed DAD model is evaluated under various experimental conditions and various datasets, namely, MVTec [5], MNIST [10], and Fashion-MNIST [3].

The remainder of the paper is organized as follows. In Section 2, we briefly introduce several DAD models. In Section 3, we describe AMLs. In Section 4, we detail the proposed algorithm through theoretical analysis. In Section 5, we report and discuss the experimental results. Finally, in Section 6, we summarize the study. Notably, this study is an extension of our previous study [9] by combining self-supervised learning and AMLs.

2. Related Works

This section provides an outline of the popular reconstruction-based DAD and self-supervised learning-based DAD for visual anomaly detection.

2.1. Reconstruction-Based DAD

Reconstruction-based methods project a “normal” sample into a lower-dimensional latent space and then reconstruct it to approximate the original input. It is generally assumed that a high reconstruction error can distinguish a “normal” instance from an “abnormal” instance. Schlegl et al. argued that the discriminator of GANs, which is pre-trained on a “normal” sample, projects an “abnormal” sample far from the feature of a “normal” instance [11]. Zenati et al. tried to increase the efficiency of the test process by training the encoder and decoder simultaneously using a Bi-GAN structure [12]. Akcay et al. attempted to capture the distribution of “normal” samples by an additional encoder to the existing autoencoder structure and compared the features of the reconstructed image using an autoencoder (AE) [13]. Sabokrou et al. attempted to reconstruct a more realistic image through adversarial learning with discriminators [14]. Akcay et al. exploited adversarial
learning through an encoder–decoder network architecture with a skip connection that helps capture the detail of images [15]. Gong et al. proposed the memory-guided AE, which stores the characteristics of “normal” instances to limit the powerful generalization capabilities of CNNs [16]. Park et al. introduced loss functions that, unlike memory-guided AE, guarantee intra-class compactness and inter-class separateness of “normal” instance patterns based on a 2D convolutional AE to increase the efficiency of the memory module [17]. Perera et al. proposed a one-class GAN structure that includes two hostile discriminators and a classifier to ensure that the feature vector of the “normal” instance has a uniform distribution, thereby obtaining a high reconstruction error for “abnormal” instances [18]. Hong et al. proposed a model with high reconstruction error for “abnormal” data that does not involve complicated processes, such as adversarial learning; instead, the model uses a dispersion loss function that spreads feature vectors in a limited area [19].

2.2. Self-Supervised Learning-Based DAD

Self-supervised learning-based methods predict the transformation applied to an image or restore a damaged image, leading to the learning of the semantic features of “normal” instances. Gidaris et al. argued that the semantic characteristics of “normal” instances could be learned to recognize arbitrary geometric transformations applied to the input image without accessing the original image [8]. Golan et al. trained a model to discriminate dozens of geometric transformations, including horizontal flipping, translations, and rotations, applied to “normal” instances to learn a meaningful representation of “normal” instances [7]. Kim et al. proposed a self-supervised learning method based on morphological transformations, including dilation, erosion, and morphological gradients, to detect irregularities in data in the same semantic category [9]. Other studies have applied self-supervised learning-based methods using restoration. These methods transform the input image and restore it to learn the critical semantic features that enable the discrimination of “normal” and “abnormal” instances. Sabokrou et al. proposed adversarial visual irregularity detection using two models [20]. The first model is an AE model that considers irregular objects in the input image as noise and denoises pixel-level irregularities based on the dominant textures of the image. The other model detects irregularities in the image using patch units. These models were trained using adversarial learning. Zarvrtanik divided the original image into small patches, and randomly deleted and restored them, allowing the network to learn the semantic features of “normal” instances [21]. Fei et al. proposed an attribute restoration network, which erases some attributes from the “normal” instances and forces the network to restore the erased attributes [22].

3. Angular Margin Loss

In this section, we provide a simple description of the angular loss and the general definition of AML. The proposed algorithm enhances the anomaly detection performance by defining the objective function in the AML format.

3.1. Softmax Loss

The most popularly used classification objective function is the softmax loss $L_{\text{softmax}}$:

$$L_{\text{softmax}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{W^T y_i}}{\sum_{j=1}^{n} e^{W^T y_j}},$$ (1)

where $x_i \in \mathbb{R}^d$ is the embedded feature of the $i$th instance in the dataset, which belongs to the $c_i$th class. $W_j \in \mathbb{R}^d$ is the $j$th column vector of weight $W \in \mathbb{R}^{d \times n}$, and $N$ and $n$ are the batch size and number of classes, respectively. Despite its popularity, the softmax loss does not explicitly optimize the feature embedding to enforce higher similarity for intra-class and diversity for inter-class samples.
3.2. Angular Loss

Because $W^T x$ is equal to $\|W\| \|x\| \cos \theta$, the aforementioned softmax objective function can be reformulated as follows:

$$L_{\text{softmax}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{\|W_{ci}\| \|x_i\| \cos (\theta_{ci})}}{\sum_{j=1}^{n} e^{\|W_j\| \|x_i\| \cos (\theta_j)}}.$$  (2)

where $\|\cdot\|$ is the $l_2$ norm operation and $\theta_j$ is the angle between $W_j$ and $x_i$. To transform the softmax loss $L_{\text{softmax}}$ in Equation (2) to angular loss, following [23–25], we normalize $\|W_j\| = 1$ by $l_2$ normalization. Then, following [24–27], we set the feature $x_i$ by $l_2$ normalization and rescale it to $r$. This process makes the classifier depend only on the angle between the embedded feature $x_i$ and weight $W_j$. Therefore, the features are distributed on a hypersphere with a radius of $r$, and the softmax angular loss $L_{\text{angular}}$ can be expressed as follows [28]:

$$L_{\text{angular}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{r \cos (\theta_{ci})}}{\sum_{j=1,j \neq ci}^{n} e^{r \cos (\theta_j)}}.$$  (3)

3.3. Angular Margin Loss

To enhance its discriminative power, the softmax angular loss can be transformed into an AML $L_{\text{margin}}$ as follows:

$$L_{\text{margin}} = -\frac{1}{N} \sum_{i=1}^{N} \log e^{(m_1 \theta_{ci} + m_2) - m_3} e^{r \cos (\theta_{ci})} + \sum_{j=1,j \neq ci}^{n} e^{r \cos (\theta_j)},$$  (4)

where $m_1$, $m_2$, and $m_3$ are the margins of a multiplicative AML (MAML) [23,29], additive AML (AAML) [28], and additive cosine margin loss (ACML) [24,27], respectively. In numerical analysis, these three AMLs aim to enforce the intra-class compactness and inter-class diversity by penalizing the target logit. However, in the geometric analysis, the proposed AAML has a constant linear angular margin throughout the interval. In contrast, MAML and ACML only have nonlinear angular margins [28]. In essence, the proposed DAD model can enhance its detection performance by utilizing AML, a robust classifier based on a straightforward modification of the softmax loss function.

4. Proposed Method

In this section, we describe the proposed deep morphological anomaly detection algorithm, which effectively represents the dominant morphological features of “normal” instances via self-supervised learning and recognizes “abnormal” samples through an enhanced classifier using AML.

4.1. Morphological Image Processing

In digital image processing, a mathematical morphology transformation is a mechanism for extracting image components useful for representing and describing the shape of the regions, such as boundaries, skeletons, and convex hulls [30]. The proposed deep anomaly detection learns the morphological features by three representative morphological transformations: erosion, dilation, and morphological gradient.

4.1.1. Erosion and Dilation

The erosion at any location $(x, y)$ of image $A$ by a kernel $b$ is the minimum value of $A$ in the region covered by $b$ when the central point of $b$ is at $(x, y)$. If $b$ is an $S \times T$ kernel, obtaining the erosion at a pixel requires obtaining the smallest of the $ST$ values of
A included in an $S \times T$ region determined by the kernel when its origin was at that point. Mathematically, the erosion is defined as follows:

$$[A \ominus b](x, y) = \min_{(s,t) \in b} \{A(x+s, y+t)\}, \quad (5)$$

where $[A \ominus b]$ denotes the erosion of $A$ with filter $b$, and $A(x,y)$ denotes the $(x,y)$ pixel in image $A$, and $(s,t)$ is the $(s,t)$ pixel in filter $b$. Notably, the origin points in $A$ and $b$ are defined as the top-left corner pixel and central pixel, respectively. Because the erosion calculates the minimum pixel value of $A$ in every neighborhood of $(x,y)$ coincident with $b$, it is expected that the size of bright features will be reduced, and the size of dark features will be increased. The third column in Figures 2 and 3 show the eroded images of the “normal” and “abnormal” samples, respectively, in the “tile” class of MVTec. From these figures, it can be seen that the erosion process enlarges the dark features of an image. Additionally, it was found that the shape of the extracted features from the morphological transformation depends on the form of the kernel. If a vertical-shaped filter is used, the erosion causes the dark region to enlarge vertically.

$$[A \ominus b](x, y) = \min_{(s,t) \in b} \{A(x+s, y+t)\}, \quad (5)$$

where $[A \ominus b]$ denotes the erosion of $A$ with filter $b$, and $A(x,y)$ denotes the $(x,y)$ pixel in image $A$, and $(s,t)$ is the $(s,t)$ pixel in filter $b$. Notably, the origin points in $A$ and $b$ are defined as the top-left corner pixel and central pixel, respectively. Because the erosion calculates the minimum pixel value of $A$ in every neighborhood of $(x,y)$ coincident with $b$, it is expected that the size of bright features will be reduced, and the size of dark features will be increased. The third column in Figures 2 and 3 show the eroded images of the “normal” and “abnormal” samples, respectively, in the “tile” class of MVTec. From these figures, it can be seen that the erosion process enlarges the dark features of an image. Additionally, it was found that the shape of the extracted features from the morphological transformation depends on the form of the kernel. If a vertical-shaped filter is used, the erosion causes the dark region to enlarge vertically.

In contrast, the dilation at any location $(x, y)$ of image $A$ by a kernel $b$ is the maximum value of $A$ in the region covered by $b$ when the origin of $b$ is at $(x, y)$. The dilation transformations can be defined as follows:

$$[A \oplus b](x, y) = \max_{(s,t) \in b} \{A(x+s, y+t)\}, \quad (6)$$

where $[A \oplus b]$ denotes the dilation of $A$ with filter $b$. In contrast to the erosion process, dilation increases the size of bright features and decreases the size of the dark features. The second column in Figures 2 and 3 show the dilated figures of “normal” instances and
the anomalies in the MVTec “tile” class, respectively. It is evident from these images that
dilation has the opposite effect of erosion.

Figure 3. Morphologically transformed “abnormal” images in the “cracked tile” class of MVTec [5]: (first
column) the original “abnormal” image; (second column) dilated “abnormal” images; (third column)
eroded “abnormal” images; (fourth column) morphological gradient of “abnormal” images; (first row
except the top left image) “normal” images morphologically transformed with [13, 13] kernel; (second
row) images transformed with [1, 13] kernel; (third row) images transformed with [13, 1] kernel. [5, T]
is an S × T kernel, where S and T denote the width and height of the filter, respectively.

4.1.2. Morphological Gradient

To obtain the morphological gradient of an image, dilation and erosion can be used in
combination with image subtraction. This operation can be expressed as follows:

\[
[A \odot b] = [A \oplus b] - [A \ominus b],
\]

where \( \odot \) denotes morphological gradient operation.

Because dilation increases regions in an image, whereas erosion decreases them, the
difference between them highlights the boundaries between areas. The emphasis of edges
and suppression of homogeneous regions in an image is referred to as the “derivative-like”
(gradien) effect. The third column in Figures 2 and 3 shows the morphological gradient
images of the chosen “normal” and “abnormal” images, respectively. From these images, it
is evident that this morphological transformation emphasizes the outline of the cracked
area. In addition, the outcome of the morphological gradient depends on the shape of
the filter.

4.2. Deep Morphological Anomaly Detection via Angular Margin Loss

The proposed algorithm was developed to detect “abnormal” instances based on
morphological criteria in the anomaly detection process. To achieve this goal, the model
adopts self-supervised learning, which easily makes the DAD model learn the features of
interest, and the CNN architecture, which exhibits high performance in various computer
vision tasks. In self-supervised learning, because the learned features are chosen, we
propose a novel objective function that enforces the DAD model to represent the dominant morphological features of “normal” instances.

The problem of anomaly detection in images can be defined as follows:

\[
    f_{DAD}(\mathbf{A}) = \begin{cases} 
    1 & g(\mathbf{A}) \geq \lambda \\
    0 & g(\mathbf{A}) < \lambda '
    \end{cases}
\]  

(8)

where \( \mathbf{A} \) denotes an input image, \( f_{DAD} \) denotes an anomaly detection function that returns 1 if the input instance \( \mathbf{A} \) belongs to “normal” class, \( g(\mathbf{A}) \) is a scoring function that quantifies the normality of the input, and \( \lambda \) is the threshold parameter that controls the recall and precision of the DAD function \( f_{DAD} \). Similar to the previous algorithm [8,9], the proposed algorithm aims to estimate the scoring function \( g \) in the aforementioned equation.

Self-Supervised Learning Using Morphological Transformations

The proposed algorithm estimates the optimal scoring function using self-supervised learning. Therefore, we constructed a self-labeled dataset of images from an initial “normal” training set \( \mathcal{D} = \{ \mathbf{A}_1, \ldots, \mathbf{A}_{|\mathcal{D}|} \} \), using both a set of morphological transformations \( \mathcal{M} = \{ \mathcal{M}_1, \ldots, \mathcal{M}_{|\mathcal{M}|} \} \) (where \( \mathcal{M}_i \) denotes the \( i \)th class of morphological transformation \( \mathcal{M}_i \), and \( |\mathcal{M}| \) is the number of classes of \( \mathcal{M} \)) and a set of geometric rotation \( \mathcal{R} = \{ R_1, \ldots, R_{|\mathcal{R}|} \} \) (where \( R_i \) is the \( i \)th class of \( \mathcal{R} \) and \( |\mathcal{R}| \) denotes the number of classes of \( \mathcal{R} \)). As delineated in Figures 1 and 2, because the morphological transformation is affected by the kernel size, we define several classes for kernel width \( \mathcal{W} = \{ W_1, \ldots, W_{|\mathcal{W}|} \} \) and kernel height \( \mathcal{H} = \{ H_1, \ldots, H_{|\mathcal{H}|} \} \), respectively. Therefore, the self-labeled dataset \( \mathcal{D}' = \{ \mathbf{A}_{1|\mathcal{M}_1|\mathcal{W}_1|\mathcal{H}_1|\mathcal{R}_1|}, \ldots, \mathbf{A}_{|\mathcal{M}||\mathcal{W}||\mathcal{H}||\mathcal{R}|} \} \) is produced by applying each morphological transformation \( \mathcal{M}_i \), each kernel width \( \mathcal{W}_j \), each kernel height \( \mathcal{H}_k \), and each geometric rotation \( \mathcal{R}_p \) on all “normal” instances in \( \mathcal{D} \). Subsequently, we label each transformed image with indices of the transformation that was applied on it, i.e., \( \mathbf{A}_{k|M_i|W_j|H_o|R_p} \) denotes morphological transformed \( k \)th “normal” image in \( \mathcal{D} \), to which the transformation label \( [M_i, W_j, H_o, R_p] \) is applied.

After the creation of \( \mathcal{D}' \), the proposed DAD model \( F \) is trained to predict the transformation information \( [M_i, W_j, H_o, R_p] \) on the input-perturbed instance \( \mathbf{A}'_{k|M_i|W_j|H_o|R_p} \); Notably, the label \( [M_i, W_j, H_o, R_p] \) is unknown to \( F \). The proposed DAD model \( F \) consists of a single feature extractor and four classifiers, similar to a hard-parameter-sharing-based multi-task model. We denote the classifiers in \( F \) as \( F_M, F_W, F_H, \) and \( F_R \); they are designed to predict the class of morphological transformation \( M_i \), the class of kernel width \( W_j \), the class of kernel height \( H_o \), and the class of geometric rotation \( R_p \), respectively. All objective functions of classifiers are defined following AML fashion to enhance the discriminative power. Therefore, the proposed objective function is defined as follows:

\[
    L_{\text{proposed}}(\mathbf{A}'_{k|M_i|W_j|H_o|R_p}) = \\
    L_M(\mathbf{A}'_{k|M_i|W_j|H_o|R_p}|\theta_M) + L_W(\mathbf{A}'_{k|M_i|W_j|H_o|R_p}|\theta_W) \\
    + L_H(\mathbf{A}'_{k|M_i|W_j|H_o|R_p}|\theta_H) + L_R(\mathbf{A}'_{k|M_i|W_j|H_o|R_p}|\theta_R),
\]  

(9)

where \( \theta_M = \{ \theta_{M_1}, \ldots, \theta_{M_{|\mathcal{M}|}} \} \), \( \theta_W = \{ \theta_{W_1}, \ldots, \theta_{W_{|\mathcal{W}|}} \} \), \( \theta_H = \{ \theta_{H_1}, \ldots, \theta_{H_{|\mathcal{H}|}} \} \), and \( \theta_R = \{ \theta_{R_1}, \ldots, \theta_{R_{|\mathcal{R}|}} \} \) are the sets of angles between the embedded features of \( F \) and given classes in \( \mathcal{M}, \mathcal{W}, \mathcal{H}, \) and \( \mathcal{R} \), respectively. AMLs for \( M, W, H, \) and \( R \) are defined according to the AML definition in Section 3. We report these four objective functions in Appendix A. Through the objective function, the proposed DAD model learns dominant morphological features of “normal” images. In addition, because the proposed loss function is based on self-supervised learning and AML, the DAD model can easily train “normal” instances in an improved discriminative manner.
At the inference time, given an unseen image A, we decide whether it belongs to the “normal” class by first applying each transformation on it and then applying the classifier g which is a useful performance metric to measure the quality of the trade-off of the final normality score is defined using the combined log-likelihood of these vectors under an estimated distribution of “normal” AML output vectors.

Subsequently, we define the normality score function \( g(A) \)—fix a set of morphological transformations \( T = \{T_0, \ldots, T_{|M||W||H||R|} \} \) and define \( F(T_i(A)) \), which is the vector of AML response of the proposed model \( F \) applied on \( i \)th transformed image \( T_i(A) \). To construct the normality score, we define

\[
g(A) = \log P(F(T_i(A)) | T_i),
\]

which is the combined log-likelihood of a transformed image conditioned on each of the applied transformations in \( T \). By following [7], we approximate each conditional to be \( F(T_i(A)) \) to \( \text{Dir}(a_i) \), where \( a_i \in \mathbb{R}^{|M||W||H||R|−1} \) is the Dirichlet parameter, \( \mathbf{A} \sim P_{\text{real}}(\mathbf{A}), i \sim \text{Un}(0,\{|M||W||H||R|−1\}) \), and \( P_{\text{real}}(\mathbf{A}) \) is the real data probability distribution of “normal” images. The primary reason for the choice of the Dirichlet distribution is that it is a common choice for distribution approximation when the \( F(T_i(A)) \) reside in the unit \(|M||W||H||R|−1\) simplex. This is because, in the proposed algorithm, we modified the softmax function by applying margin penalties (scalars) on the between angle \( \theta \) and its cosine value, and the response vectors of AML and the softmax function do not differ. Therefore, the score function \( g(A) \) critically captures normality such that for two images \( A \) and \( A' \), \( g(A) > g(A') \) tend to imply that \( A \) is “more normal” than \( A' \).

5. Experimental Results and Discussion

In this section, deep anomaly detection experiments were performed to verify the performance of the proposed algorithms on several datasets, including MVTec [5], MNIST [10], and Fashion-MNIST [3]. Following [7,9], we learn the representations by transformation prediction from scratch with ResNet-34 [31]. The classes in the proposed algorithm are \( M = \{\oplus, \ominus, \circ\}, W = \{1,4,8,12\}, H = \{1,4,8,12\}, \) and \( R = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\} \). We set the classes in \( R \) by following the geometric rotation classes in [7–9]. Additionally, we follow [28] to set rescale parameter \( r \), multiplicative angular margin \( m_1 \), additive angular margin \( m_2 \), and additive cosine margin \( m_3 \) to 64, 0.9, 0.4, and 0.15, respectively. All experimental results are reported as the area under the receiver operating characteristic (AUROC), which is a useful performance metric to measure the quality of the trade-off of \( g(A) \) in (8). Because MNIST and Fashion-MNIST are not designed for anomaly detection, we trained the model for only one class as a “normal” class, and the performance was evaluated for the entire test dataset—classes other than the trained class were assumed to be “abnormal.” The proposed algorithm was actualized using PyTorch in a GPU implementation. We performed experiments with an RTX 2080Ti 11GB graphical processing unit and an Intel i7 processor. To validate the experimental results using statistical methods, we conducted all experiments five times and then calculated the average and variance.

5.1. Experimental Results on the MVTec Dataset

MVTec contains 10 object and 5 texture categories for anomaly detection, with 3629 training instances and 1725 test samples [5]. The dataset is composed of “normal” images for training and both “normal” and “abnormal” images with various industrial defects for testing. In Table 1, we present the anomaly detection performance on the MVTec dataset. In comparison with the semantic-feature-based DAD model [7], the proposed algorithm yields a higher detection performance. This result demonstrates that extracting salient morphological features on “normal” images significantly increases the detection performance on industrial images. Compared to the results of our previous study [9], the results of this study show that self-supervised learning-based anomaly detection can be improved by...
merely applying AMLs. In addition, the proposed algorithm outperforms (0.6 AUROC) the existing DAD algorithm [32], which leverages pretrained networks (87.9 AUROC). The results on this dataset confirm that the combination of morphological transformations and AML-based self-supervised learning provides satisfactory performance in industrial anomaly detection problems.

Table 1. Comparison of mean and variance AUROC performances for various algorithms on the MVTec dataset.

| Class      | Golan et al. [7] | Kim et al. [9] | Ruff et al. [32] | Kingma et al. [33] | Proposed |
|------------|------------------|----------------|------------------|--------------------|----------|
| Texture    |                  |                |                  |                    |          |
| Carpet     | 38.1 (3.7)       | 57.9 (2.4)     | 90.6 (0.4)       | 48.2 (2.3)         | 95.3 (0.2) |
| Grid       | 31.4 (2.4)       | 29.9 (4.4)     | 52.4 (1.9)       | 32.9 (1.7)         | 86.1 (0.8) |
| Leather    | 64.1 (0.8)       | 82.9 (0.3)     | 78.3 (0.7)       | 65.9 (0.8)         | 93.4 (0.4) |
| Tile       | 52.2 (1.8)       | 93.6 (0.2)     | 96.5 (0.3)       | 73.4 (0.2)         | 98.4 (0.1) |
| Wood       | 84.3 (0.6)       | 87.4 (0.5)     | 91.6 (0.1)       | 83.5 (0.2)         | 95.8 (0.7) |
| Mean       | 54.0             | 70.3           | 81.6             | 60.9               | 93.8     |
| Object     |                  |                |                  |                    |          |
| Bottle     | 83.1 (0.5)       | 95.2 (0.3)     | 99.6 (0.2)       | 90.4 (0.5)         | 98.2 (0.6) |
| Cable      | 77.8 (0.2)       | 80.3 (0.7)     | 90.9 (0.1)       | 83.8 (0.3)         | 94.0 (0.2) |
| Capsule    | 75.3 (0.6)       | 73.1 (0.2)     | 91.0 (0.4)       | 75.4 (0.7)         | 87.7 (0.1) |
| Hazelnut   | 67.1 (1.3)       | 68.0 (1.2)     | 95.0 (0.2)       | 73.1 (0.4)         | 96.8 (0.3) |
| Metal nut  | 69.3 (0.8)       | 72.7 (0.6)     | 85.2 (0.3)       | 60.4 (0.9)         | 86.9 (0.8) |
| Pill       | 62.2 (1.1)       | 57.2 (2.7)     | 80.4 (0.7)       | 81.2 (1.1)         | 83.7 (0.6) |
| Screw      | 27.7 (3.7)       | 61.9 (2.6)     | 86.9 (1.7)       | 66.4 (0.5)         | 88.3 (0.2) |
| Toothbrush | 82.7 (0.6)       | 91.7 (0.2)     | 96.4 (0.5)       | 83.4 (0.2)         | 99.2 (0.0) |
| Transistor | 88.2 (0.3)       | 83.3 (1.2)     | 90.8 (1.7)       | 82.5 (0.7)         | 89.4 (0.6) |
| Zipper     | 72.3 (0.5)       | 75.7 (0.3)     | 92.4 (0.7)       | 72.7 (1.7)         | 95.8 (0.7) |
| Mean       | 70.6             | 75.9           | 90.9             | 76.9               | 92.0     |

To illustrate the difference between the proposed algorithm and semantic feature-based DAD [7], a visual saliency map (visual representations) using Grad-CAM++ [34] is presented in Figure 4. Grad-CAM++ is a representative interpretable machine learning technique. This technology enables us to visually identify the parts of the input image that most critically influence the CNN’s output. The MVTec dataset is divided into two overarching classes: “texture” and “object.” Figure 4 shows visualizations utilizing images of the “carpet” class within the texture class and the “cable” class in the object class. The first row of Figure 4 represents visualizations for the carpet class. For images corresponding to the texture class, the DAD model must detect all regions of the input image because they have morphological characteristics relevant to the entire image. Therefore, visualizations of the results by the proposed method (shown in the second and fifth columns in Figure 4) confirm that the saliency map appears in the input image when considered as a whole. In contrast, semantic feature-based DAD shows an unequal visual expansion with a high saliency score only for the edge region of the image. This is the primary reason that the proposed method has higher AUROC performance than existing methods in the texture class of the MVTec dataset. Unlike the texture class, subclasses belonging to the object class contain crucial object information in the center of the image. The second row in Figure 4 shows visual representations of the cable class. The proposed method, similar to the output in the carpet class, has a saliency map for the entire area of the image, but its saliency map is more intense in the vicinity of the object. Conversely, semantic feature-based DADs have high saliency map concentrations on wires and on the covering parts of cables. Most notably, based on the experiments using the cable class, in the “abnormal” image, the proposed method is quite accurate—it has a high saliency score in the area expressed as defective. This implies that the proposed method has sufficiently learned various morphological features of the image and proves that it is suitable for industrial anomaly detection problems.
5.2. Experimental Results on MNIST and Fashion-MNIST Datasets

The MNIST dataset contains 10 categories labeled 0 to 9, and the Fashion-MNIST contains 10 categories of clothing. As aforementioned, in these experiments, we followed a one-class classification protocol [19]. In addition, we compared the performance of the proposed method with reconstruction-based DAD models, such as AE-, VAE-, and GAN-based algorithms. Tables 2 and 3 present the results on MNIST and Fashion-MNIST, respectively. In the MNIST experiment, the proposed algorithm achieves slightly better performance than the other algorithms because the classification criteria between different digits are based on several morphological characteristics. In the Fashion-MNIST experiments, the proposed algorithm exhibited lower performance for some classes. In particular, because this dataset contains various styles of “coat,” a high semantic feature understanding of the DAD model is required to achieve better performance. Conversely, the proposed algorithm is designed to focus on salient morphological features of “normal” images to detect small morphological differences between “normal” and “abnormal” instances. Therefore, the proposed algorithm is not appropriate for some classes in Fashion-MNIST. Notably, this phenomenon is not a limitation of the proposed model but the necessity for a proper DAD model design by considering the target.

Table 2. Comparison of mean and variance AUROC performances for various algorithms on the MNIST dataset.

| Class | Schlegl et al. [11] | Vicent et al. [35] | Akcay et al. [13] | Kingma et al. [33] | Proposed |
|-------|---------------------|--------------------|-------------------|-------------------|----------|
| 0     | 96.3 (0.2)           | 99.5 (0.1)         | 99.0 (0.4)        | 99.7 (0.2)        | 99.7 (0.0) |
| 1     | 99.3 (0.1)           | 99.9 (0.0)         | 99.9 (0.0)        | 99.9 (0.0)        | 98.8 (0.1) |
| 2     | 84.9 (0.9)           | 90.7 (0.6)         | 91.4 (1.2)        | 93.5 (0.4)        | 94.7 (0.2) |
| 3     | 88.1 (0.7)           | 94.5 (0.4)         | 93.6 (0.6)        | 95.8 (0.6)        | 95.8 (0.4) |
| 4     | 89.4 (0.2)           | 95.0 (0.6)         | 97.0 (0.2)        | 97.3 (0.3)        | 96.8 (0.2) |
| 5     | 88.3 (0.5)           | 96.1 (0.7)         | 96.6 (0.4)        | 96.2 (0.5)        | 96.3 (0.3) |
| 6     | 94.3 (0.2)           | 98.6 (0.2)         | 99.2 (0.1)        | 99.1 (0.2)        | 99.3 (0.1) |
| 7     | 94.3 (0.3)           | 96.6 (0.4)         | 97.7 (0.3)        | 97.5 (0.4)        | 98.5 (0.4) |
| 8     | 93.5 (1.2)           | 84.9 (1.2)         | 92.9 (1.0)        | 92.1 (0.9)        | 94.3 (0.7) |
| 9     | 83.9 (0.3)           | 96.6 (0.3)         | 97.9 (0.2)        | 97.6 (0.6)        | 97.8 (0.4) |
| Mean  | 92.4                | 95.2               | 96.5              | 96.6              | 97.2     |
Table 3. Comparison of mean and variance AUROC performances for various algorithms on the Fashion-MNIST dataset.

| Class    | Schlegl et al. [11] | Vicent et al. [35] | Akcay et al. [13] | Kingma et al. [33] | Proposed |
|----------|---------------------|--------------------|-------------------|-------------------|----------|
| T-shirt  | 82.4 (0.5)          | 91.0 (0.2)         | 86.6 (1.3)        | 89.6 (0.2)        | 88.4 (0.1) |
| Trouser  | 95.7 (0.7)          | 98.0 (0.1)         | 97.8 (0.4)        | 98.4 (0.3)        | 98.6 (0.3) |
| Pullover | 81.2 (0.9)          | 86.2 (0.3)         | 84.0 (0.8)        | 87.5 (0.2)        | 88.2 (0.2) |
| Dress    | 90.4 (0.6)          | 91.9 (0.2)         | 93.0 (0.8)        | 92.7 (0.4)        | 92.8 (0.2) |
| Coat     | 86.1 (0.8)          | 87.4 (0.1)         | 88.5 (0.5)        | 89.9 (0.3)        | 87.9 (0.1) |
| Sandal   | 86.6 (1.1)          | 90.8 (0.4)         | 91.8 (1.5)        | 90.6 (0.5)        | 92.3 (0.4) |
| Shirt    | 79.8 (0.5)          | 80.3 (0.8)         | 81.5 (0.6)        | 83.0 (0.6)        | 84.0 (0.2) |
| Sneaker  | 93.3 (0.2)          | 97.8 (0.3)         | 97.7 (0.4)        | 98.2 (0.1)        | 98.5 (0.1) |
| Bag      | 80.1 (0.8)          | 85.6 (0.1)         | 86.7 (1.4)        | 85.9 (0.1)        | 86.5 (0.1) |
| Ankle boot | 95.1 (0.4)      | 96.5 (0.4)         | 98.0 (0.5)        | 97.6 (0.1)        | 97.8 (0.2) |

Mean 87.1 90.6 90.6 91.3 91.5

5.3. Statistical Analysis of the Results Using the Wilcoxon Signed-Rank Test

To analyze the experimental results by a statistical method, we performed the Wilcoxon signed-rank test and reported it in Table 4. The Wilcoxon signed-rank test is a popular nonparametric statistical test for matched or paired data. This statistical test is based on both difference scores and the magnitude of the observed differences. In this paper, we define the difference score as follows:

\[
difference \text{ score} = AUROC_{\text{related}}^i - AUROC_{\text{proposed}}^i,
\]

where \(AUROC_{\text{related}}^i\) and \(AUROC_{\text{proposed}}^i\) are AUROC values of the compared algorithm and the proposed algorithm on \(i\)th class in the certain dataset. In Wilcoxon signed-rank test, the hypotheses are concerned by the population median of the difference scores. In this paper, we consider a one-sided test which is as follows:

**Hypothesis 1 (H1).** The median difference is zero.

**Hypothesis 2 (H2).** The median difference is negative.

The test static for the Wilcoxon signed-rank test \(W\) is defined as the smaller value among \(W_+\) (the sum of the positive ranks) and \(W_-\) (the sum of the negative ranks). To determine whether the observed test static \(W\) supports \(H_1\) or \(H_2\), we have to choose the \(W_{\text{critical}}\). Therefore, if the \(W\) is less than \(W_{\text{critical}}\), we reject \(H_1\), in favor of \(H_2\). In contrast, if the \(W\) exceeds the \(W_{\text{critical}}\), we do not reject \(H_1\). In this paper, to set the appropriate \(W_{\text{critical}}\), we choose the level of significance \(\alpha\) to 0.05, which is the most common value for the Wilcoxon signed-rank test.

For the MVTec dataset, the proposed algorithm yields statistically highest performance than the other algorithms. Moreover, in MNIST and Fashion-MNIST, the proposed algorithm has better performance than several algorithms. Because MNIST and Fashion-MNIST are relatively easier datasets than the MVTec dataset, it exhibits a similar performance than [33]. However, notably, the proposed algorithm is developed to address the morphological anomaly detection problem. Therefore, in the Wilcoxon signed-rank test results of the MVTec dataset, it is clearly demonstrated that the proposed algorithm achieved significantly higher performance than [33] in the morphological anomaly detection task.
Table 4. Results of the Wilcoxon signed-rank test on the proposed algorithm and related algorithms on MVTec, MNIST, and Fashion-MNIST datasets. \( W_{\text{critical}}, W_-, W_+, W, \) and \( H_1 \) are critical value, the sum of the ranks of the negative differences, the sum of the ranks of positive differences, the test statistic for the Wilcoxon signed test, and the null hypothesis. The level of significance \( \alpha \) is 0.05.

| Dataset       | Compared Algorithm       | n  | \( W_{\text{critical}} \) | \( W_- \) | \( W_+ \) | \( W \) | \( H_1 \) |
|---------------|--------------------------|----|---------------------------|-----------|-----------|--------|--------|
| MVTec         | Golan et al. [7]         | 15 | 30                        | 119       | 0         | 0      | Reject |
|               | Kim et al. [9]           | 15 | 30                        | 119       | 0         | 0      | Reject |
|               | Ruff et al. [32]         | 15 | 30                        | 106.5     | 13.5      | 13.5   | Reject |
|               | Kingma et al. [33]       | 15 | 30                        | 119       | 0         | 0      | Reject |
| MNIST         | Schlegl et al. [11]      | 10 | 11                        | 54        | 1         | 1      | Reject |
|               | Vicent et al. [35]       | 10 | 11                        | 51        | 4         | 4      | Reject |
|               | Akcay et al. [13]        | 10 | 11                        | 40.5      | 14.5      | 14.5   | -      |
|               | Kingma et al. [33]       | 8  | 6                         | 26        | 10        | 10     | -      |
| Fashion-MNIST | Schlegl et al. [11]      | 10 | 11                        | 55        | 0         | 0      | Reject |
|               | Vicent et al. [35]       | 10 | 11                        | 46        | 9         | 9      | Reject |
|               | Akcay et al. [13]        | 10 | 11                        | 45        | 10        | 10     | Reject |
|               | Kingma et al. [33]       | 10 | 11                        | 37        | 18        | 18     | -      |

6. Conclusions

We proposed a novel DAD model to extract the dominant morphological features of “normal” images. The proposed algorithm is based on a combination of morphological transformations and a self-supervised learning algorithm. In addition, to improve the discriminative power of the proposed DAD model, we adopted an angular margin on the classification loss function. The experiments confirmed that the proposed algorithm achieves higher performance in the industrial anomaly detection dataset than the previous algorithms. In addition, the proposed algorithm yields slightly better performance than the previous reconstruction-based DAD models. The results validate the satisfactory performance of the proposed algorithm in real-world industrial anomaly inspection applications. In the future, we plan to combine the proposed algorithm with the semantic feature-based algorithm.

Author Contributions: Conceptualization, T.K.; methodology, T.K.; software, T.K.; validation, T.K. and Y.C.; formal analysis, T.K.; investigation, T.K. and E.H.; resources, T.K.; data curation, T.K.; writing—original draft preparation, T.K. and E.H.; writing—review and editing, T.K.; visualization, T.K.; supervision, Y.C.; project administration, Y.C.; funding acquisition, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: This work was supported by the Technology development Program (S2798925) funded by the Ministry of SMEs and Startups (MSS, Korea).

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

Abbreviations
The following abbreviations are used in this manuscript:

- AE Auto Encoder
- AML Angular Margin Loss
- AAML Additive Angular Margin Loss
- ACML Additive Cosine Margin Loss
AUROC: Area Under Receiver Operating Characteristic
CNN: Convolutional Neural Network
DAD: Deep Anomaly Detection
DNN: Deep Neural Network
GANs: Generative Adversarial Networks
MAML: Multiplicative Angular Margin Loss
VAE: Variational Auto Encoder

Appendix A
The proposed objective function is defined as follows:

\[
L_{\text{proposed}}(A'_{[M,W,H,R,p]}) = L_M(A'_{[M,W,H,R,p]}|\theta_M) + L_W(A'_{[M,W,H,R,p]}|\theta_W) + L_H(A'_{[M,W,H,R,p]}|\theta_H) + L_R(A'_{[M,W,H,R,p]}|\theta_R),
\]

(A1)

where \(\theta_M = \{\theta_{M_1}, \ldots, \theta_{M_n}\}\), \(\theta_W = \{\theta_{W_1}, \ldots, \theta_{W_n}\}\), \(\theta_H = \{\theta_{H_1}, \ldots, \theta_{H_n}\}\), and \(\theta_R = \{\theta_{R_1}, \ldots, \theta_{R_n}\}\) are the sets of angles between the embedded features of \(F\) and the given classes in \(M, W, H,\) and \(R\), respectively. According to the AML definition in Section 3, AMLs for \(M, W, H,\) and \(R\) are, respectively, defined as follows:

\[
L_M(A'_{[M,W,H,R,p]}|\theta_M) = -\log \frac{e^{r(M_1\theta_M + M_2 - m_3)}}{e^{r(M_1\theta_M + M_2 - m_3)} + \sum_{v=1, v \neq M}^{M} e^{r(M_v \theta_M)}} e^{r(M_1 \theta_M)},
\]

(A2)

\[
L_W(A'_{[M,W,H,R,p]}|\theta_W) = -\log \frac{e^{r(W_1\theta_W + W_2 - m_3)}}{e^{r(W_1\theta_W + W_2 - m_3)} + \sum_{v=1, v \neq W}^{W} e^{r(W_v \theta_W)}} e^{r(W_1 \theta_W)},
\]

(A3)

\[
L_H(A'_{[M,W,H,R,p]}|\theta_H) = -\log \frac{e^{r(H_1\theta_H + H_2 - m_3)}}{e^{r(H_1\theta_H + H_2 - m_3)} + \sum_{v=1, v \neq H}^{H} e^{r(H_v \theta_H)}} e^{r(H_1 \theta_H)},
\]

(A4)

\[
L_R(A'_{[M,W,H,R,p]}|\theta_R) = -\log \frac{e^{r(R_1\theta_R + R_2 - m_3)}}{e^{r(R_1\theta_R + R_2 - m_3)} + \sum_{v=1, v \neq R}^{R} e^{r(R_v \theta_R)}} e^{r(R_1 \theta_R)}.
\]

(A5)

References
1. Arthur, Z.; Erich, S. Outlier Detection. In Encyclopedia of Database Systems; Springer: New York, NY, USA, 2017; pp. 1–5. ISBN 9781489979933. [CrossRef]
2. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. Master’s Thesis, Department of Computer Science, University of Toronto, Toronto, ON, Canada, 2009.
3. Xiao, H.; Rasul, K.; Vollgraf, R. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. arXiv 2017, arXiv:1708.07747.
4. Elson, J.; Douceur, J.R.; Howell, J.; Saul, J. Asirra: A CAPTCHA that exploits interest-aligned manual image categorization. In Proceedings of the ACM Conference on Computer and Communications Security, Alexandria, VA, USA, 31 October–2 November 2007; Volume 7.
5. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. MVTec AD—A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
6. Jing, L.; Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 2020. [CrossRef] [PubMed]
7. Golan, I.; El-Yaniv, R. Deep anomaly detection using geometric transformations. Adv. Neural Inf. Process. Syst. 2018, 9781–9791.
8. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised Representation Learning by Predicting Image Rotations. In Proceedings of the IEEE International Conference on Learning Representations, Vancouver, BC, Canada, 30 April 2018.
9. Kim, T.; Choe, Y. Deep Anomaly Detection via Morphological Transformations. Proceedings 2020, 67, 21. [CrossRef]
10. LeCun, Y.; Cortes, C.; Burges, C.J. Minst Handwritten Digit Database; AT&T Labs: Atlanta, GA, USA, 2010.
