

# Layer-wise External Attention for Efficient Deep Anomaly Detection

Tokihisa Hayakawa, Keiichi Nakanishi, Ryoya Katafuchi and Terumasa Tokunaga<sup>a</sup>

*Kyushu Institute of Technology, 680-4 Kawazu Iizuka-Shi, Fukuoka, Japan*

**Keywords:** Anomaly Detection, Visual Inspection AI, Deep Learning, Visual Attention Mechanism, Self-Attention, MVTec AD, Plant Science.

**Abstract:** Recently, the visual attention mechanism has become a promising way to improve the performance of Convolutional Neural Networks (CNNs) for many applications. In this paper, we propose a Layer-wise External Attention mechanism for efficient image anomaly detection. The core idea is the integration of unsupervised and supervised anomaly detectors via the visual attention mechanism. Our strategy is as follows: (i) prior knowledge about anomalies is represented as an anomaly map generated by the pre-trained network; (ii) the anomaly map is translated to an attention map via an external network. (iii) the attention map is then incorporated into intermediate layers of the anomaly detection network via visual attention. Notably, the proposed method can be applied to any CNN model in an end-to-end training manner. We also propose an example of a network with Layer-wise External Attention called Layer-wise External Attention Network (LEA-Net). Through extensive experiments using real-world datasets, we demonstrate that Layer-wise External Attention consistently boosts the anomaly detection performances of an existing CNN model, even on small and unbalanced data. Moreover, we show that Layer-wise External Attention works well with Self-Attention Networks.


## 1 INTRODUCTION

Anomaly detection is a technique used to identify irregular or unusual patterns in datasets. Particularly, anomaly detection for imaging data is a powerful and core technology that can be applied to various kinds of real-world problems, including medical diagnosis (Rezvantalab et al., 2018; Cao et al., 2018), plant healthcare (Ferentinos, 2018), production quality controls, and disaster detection (Minhas and Zelek, 2019; Natarajan et al., 2019). Recently, many researchers have shown great interest in the establishment of automatic anomaly detection techniques for a huge image dataset driven by breakthroughs in deep learning. Based on the various aspects of machine learning, these anomaly detection techniques can be roughly classified into three categories: supervised, semi-supervised, and unsupervised approaches. Although each approach has its advantages and disadvantages, the fundamental challenge that should be overcome is on how we detect anomalies efficiently based on a limited number of anomalous instances.

A convolutional neural network (CNN) is a commonly used artificial network for various computer vision tasks, including image recognition and image

segmentation. CNNs have realized state-of-the-art image anomaly detection in real-world applications using a huge dataset of labeled images (Hughes and Salathe, 2016; Minhas and Zelek, 2019). However, anomaly detectors based on CNN sometimes suffer from a lack of labeled images and low anomaly instances. Some studies have proposed ideas for overcoming this point by improving the learning efficiency of CNN. Particularly, one of the previous studies introduced active learning (Nico Goernitz, 2013), whereas another employed transfer learning (Minhas and Zelek, 2019) for that same purpose.

Approaches based on unsupervised learning are the most popular method for anomaly detection because they do not require labeled anomalous instances to train anomaly detectors. The simple strategy of unsupervised image anomaly detection relies on the training of reconstruction processes for normal images using a deep convolutional autoencoder (Haselmann et al., 2018). However, the autoencoder sometimes fails to reconstruct fine structures. Consequently, it outputs immoderate blurry structures. Recently, generative adversarial network (GAN) has been used for image anomaly detection to address this problem. AnoGAN (Schlegl et al., 2017) firstly employed GAN for image anomaly detection. Addi-

<sup>a</sup>  <https://orcid.org/0000-0003-1091-2022>

tionally, AnoGAN and its extensions (Zenati et al., 2018; Akcay et al., 2018) realized minute anomalies in applications of medical image processing. More recently, AnoGAN was applied to the field of color reconstructability to realize sensitive detection of color anomalies (Katafuchi and Tokunaga, 2021). It is an approximately common procedure in unsupervised anomaly detection to define the difference between the original image and the reconstructed image as *Anomaly Score*.

Although unsupervised anomaly detectors can eliminate the labeling cost of anomalous instances for a training step, there are some drawbacks. First, they tend to overlook small and minute anomalies because *Anomaly Score* is defined based on the distance between the normal and test images. In other words, the detection performance of unsupervised anomaly detectors strongly depends on whether a good *Anomaly Score* can be designed for each purpose. Second, the appropriate threshold of the *Anomaly Score* should be tuned carefully for successful classification of normal and anomalous instances. In most practices, this process requires painful trial and error.

Nowadays, the technique called visual attention mechanism is attracting much attention in the field of computer vision (Zhao et al., 2020). Attention Branch Network (ABN) is a CNN with a branching structure called *Attention Branch* (Fukui et al., 2019). Maps produced from the *Attention Branch* represent visual explanations. This attention map enables humans to understand the decision process of CNN. Further, the authors reported that the attention map contributes to improving the performances of CNN for several image classification tasks. Afterward, researchers in the field of computer vision have attempted to build a technique that finely induces the attention of CNN to informative regions in images. For the last several years, it has been shown that such technique is promising for improving the performance of CNN for several computer vision tasks, including image classification and segmentation (Fukui et al., 2019; Emami et al., 2020). Inspired by these studies, one may think of the possibility of introducing the visual attention mechanism to image anomaly detections. However, visual attention modules including ABN, rely on Self-Attention mechanisms (Hu et al., 2020a; Woo et al., 2018). Accordingly, the attention quality of the modules strongly depends on the performance of the network itself, which means that existing visual attention mechanisms could not boost the performance of anomaly detectors directly.

In this paper, we propose Layer-wise External Attention mechanism to boost CNN-based anomaly detectors. As mentioned above, purely unsupervised

anomaly detectors do not utilize anomalous instances for training. As a result, they tend to overlook small and minute anomalies, or, conversely, incur high false positives. Furthermore, purely supervised approaches suffer from a lack of labeled images, particularly from a lack of anomalous instances. In this paper, we tackle these problems by integrating supervised and unsupervised anomaly detectors via a visual attention mechanism. Recent progress on visual attention mechanism strongly implies that there must be a way to utilize prior knowledge for anomaly detections. We expect that the anomaly map generated from an unsupervised anomaly detector is useful for boosting supervised anomaly detectors but powerless by itself for efficient anomaly detections. Moreover, we expect that such a boosting approach contributes to reducing the developing cost of anomaly detectors, because we can divert existing image classifiers for image anomaly detections.

Our overall strategy is as follows: (i) Prior knowledge of anomalies is represented as an anomaly map, which is generated through unsupervised learning of normal instances; (ii) The anomaly map is then translated to an attention map by an external network; (iii) The attention map is then incorporated into intermediate layers of the Anomaly Detection Network (ADN). We note that Layer-wise External Attention can be easily applied to any CNN model in an end-to-end training manner. For a pilot study, we focused on an anomaly in colors because color anomalies are comparatively easy to represent based on CIEDE color differences (Katafuchi and Tokunaga, 2021). We examined the effectiveness of Layer-wise External Attention for image anomaly detection through extensive experiments using real-world publicly available datasets. The results demonstrated that Layer-wise External Attention consistently boosts the performance of anomaly detectors even on small and unbalanced data.

Our main contributions are as follows:

- We proposed Layer-wise External Attention mechanism for efficient image anomaly detections.
- We proposed an example of a network with Layer-wise External Attention called LEA-Net.
- We showed that Layer-wise External Attention successfully boosts the performance of image anomaly detectors, even for small and imbalanced training data.
- We showed that the combination of Layer-wise External Attention and Self-Attention realizes further improvement of anomaly detectors.

- We provide a Python implementation of LEA-Net at: [https://github.com/Tokunaga-LAB-Group/Layer-wise\\_External\\_Attention\\_Network](https://github.com/Tokunaga-LAB-Group/Layer-wise_External_Attention_Network).

## 2 RELATED WORK

The proposed method can be categorized as a semi-supervised approach. Most recently, a simpler approach was adopted in a task for automatic identification of thyroid nodule lesions in X-ray computed tomography images (Li et al., 2021). The technique uses binary segmentation results obtained from a universal network (U-Net) as inputs for supervised image classifiers. The authors showed that the binary segmentation as a preprocessing for an image classification contributes to improve anomaly detections in a real-world problem. Looking for a somewhat similar setting, Convolutional Adversarial Variational Autoencoder with Guided Attention (CAVGA) uses anomaly map in a weakly supervised setting to localize anomalous regions (Venkataraman et al., 2020). CAVGA achieved state-of-the-art results through experiments for image anomaly detections using MVTEC AD. These two studies imply that the introduction of a visual attention map has a great potential for image anomaly detections.

A visual attention mechanism refers to the process of refinement or enhancement of image features for recognition tasks. The human perceptual system tends to preferentially capture information relevant to the current task, rather than processing all information (Reynolds and Chelazzi, 2004; Chun et al., 2011). Visual attention mechanisms imitate the human perceptual mechanism for image classification (Hu et al., 2020a; Wang et al., 2017; Woo et al., 2018; Lee et al., 2019; Wang et al., 2020; Yang et al., 2021). Most image classifiers with visual attention adopt a Self-Attention module that works in plug-and-play with existing models. In such specifications, the effectiveness of visual attention mechanisms depends strongly on the performance of the main body of the model: this is a limitation of Self-Attention approaches. ABN (Fukui et al., 2019) overcame this problem by interactive editing of attention map. It enabled us to induce focus points of CNN to more informative regions on images through the correction of the attention map. Another similar visual attention mechanism is Attention Transfer, which is based on knowledge distillation (Zagoruyko and Komodakis, 2017). In the process of knowledge distillation, a smaller network called the student network receives prior knowledge from a larger network called the teacher network (Hinton et al., 2015). The idea

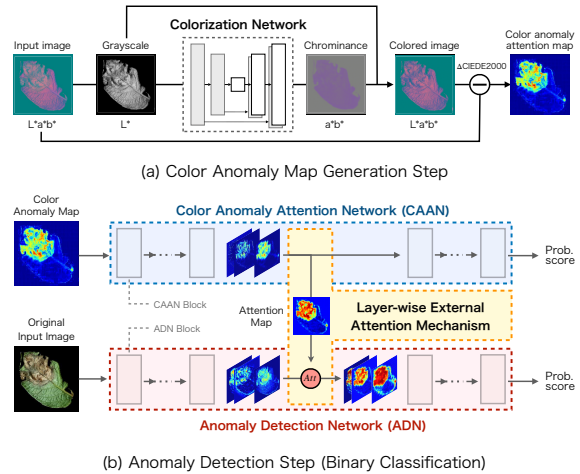


Figure 1: Overview of anomaly map generation.

of Attention Transfer relies on the assumption that a teacher network tends to focus on a more informative area than does a student network.

In this paper, we designed the structure of the network inspired by ABN and Attention Transfer network with several points of improvement. Our designed method does not require any user interaction to control focus points of CNN. The training of network with Layer-wise External Attention goes on fully automatically through a collaboration of two networks: an external network and an ADN. The external network adjusts strength of attention with the progress of training to avoid excessive effects of the visual attention mechanism in the early stage of training. Also, we note that the effectiveness of Layer-wise External Attention is not constrained by the teacher network, unlike in Attention Transfer Network.

## 3 PROPOSED METHOD

Figure 1 illustrates the overview of anomaly map generation. Details will be described below.

### 3.1 Anomaly Map Generation

Here, we focus on anomalies in colors. In the first step, a grayscale image is firstly obtained the input color image represented in  $L^*a^*b^*$  color space. Secondly, the grayscale image is reconverted to a color image using U-Net (Ronneberger et al., 2015). To supplement, the color information  $a^*b^*$  (chrominance) is predicted based on  $L^*$  (luminance) information in the  $L^*a^*b^*$  color space. Thirdly, the predicted  $a^*b^*$  is combined with the  $L^*$  of input color image to produce the resulting colored image in  $L^*a^*b^*$  color

space. The U-Net learns in advance the process of color reconstruction with normal instances only. Finally, a color anomaly map is generated by calculating CIEDE2000 (Sharma et al., 2005) color difference between the reconstructed image and the original image. For details of the color anomaly generation, see (Katafuchi and Tokunaga, 2021).

Here, we consider data augmentation, because the aim of this study is on anomaly detection with comparatively a small number of training data. Empirically, the color reconstruction often fails for quite bright or dark images. Hence, we used Fancy PCA image augmentation (Krizhevsky et al., 2012) to amplify the variation of image luminance.

### 3.2 Color Anomaly Attention Network (CAAN)

We describe the detail of Figure 1(b): Anomaly detection step. In this step, we use two different networks. One is Color Anomaly Attention Network (CAAN). This network converts the anomaly map to attention map by adjusting the complexity, certainty and sharpness of anomaly map depending on the progress in training. We refer to this network as *External Network*. The other is Anomaly Detection Network (ADN). ADN outputs the final decision for anomaly detection. It is desirable that CAAN is a lightweight and effective model so as to work in plug-and-play without a massive increase in parameters. Thus, structures of CAAN were designed based on MobileNet and ResNet. Table 1 describes the detailed structures of CAAN. The structure of MobileNet-based CAAN is the same as that of MobileNetV3-Small (Howard et al., 2019). The structure of ResNet-based CAAN is configured by simply stacking residual blocks (He et al., 2016). We note that these two models are generic for image classifiers. Therefore, they can be replaced easily by other networks.

### 3.3 Anomaly Detection Network (ADN)

The role of ADN is to make the final decision for image anomaly detection by integrating labeled images and attention map received from CAAN. We adopted ResNet18 and VGG16 for ADN, both of which are well-known and widely-used CNN for image classifications. The number of downsampling points in ADN should be the same or larger than that of CAAN due to the reasons described below.

Table 1: Structures of CAAN: MobileNet-based and ResNet-based networks. The convolution layers are denoted as  $\{\text{conv2d}, \langle \text{receptive field} \rangle \times \langle \text{receptive field} \rangle, \langle \text{number of channels} \rangle\}$ . "bneck" denotes bottleneck structure; see (Howard et al., 2019).

block name	ResNet-based CAAN	MobileNet-like CAAN
input (256×256×1 Anomaly map)		
block 1	conv2d, 7x7, 64	conv2d, 3x3, 16
Attention Output Point 1		
block 2	conv2d, 3x3, 64 conv2d, 3x3, 64	bneck, 3x3, 16
Attention Output Point 2		
block 3	conv2d, 3x3, 128 conv2d, 3x3, 128	[bneck, 3x3, 24] × 2
Attention Output Point 3		
block 4	conv2d, 3x3, 256 conv2d, 3x3, 256	[bneck, 5x5, 40] × 3 [bneck, 5x5, 48] × 2
Attention Output Point 4		
block 5	conv2d, 3x3, 512 conv2d, 3x3, 512	[bneck, 5x5, 96] × 3
Attention Output Point 5		
block 6	average pool 2-d fc sigmoid	conv2d, 1x1, 576 average pool conv2d, 1x1, 1280 conv2d, 1x1, 2 sigmoid
Params	4.491M	3.042M

### 3.4 The Overall Structure of LEA-Net

In Figure 2 we illustrate the overall structure of LEA-Net, which is an example of a network with Layer-wise External Attention. Here, the CAAN has five feature extraction blocks, and therefore it has five alternatives for outputting an attention map for Layer-wise External Attention. Accordingly, ADN can have up to five attention points. In practical, the number of the attention point should be limited to one for each anomaly detection for avoiding the performance deterioration of ADN.

In the training process, both CAAN and ADN are optimized by passing through the gradients of CAAN and ADN during back propagation. Let  $\mathbf{x}_i \in \mathbb{R}^{H \times W \times 3}$  be the  $i$ th original input image. Let  $\mathbf{x}_i^{Att} \in \mathbb{R}^{H \times W}$  be the  $i$ th color anomaly map. Also, let  $y_i \in \{0, 1\}$  be a corresponding ground-truth label. Further, let  $L_{Att}$  and  $L_{AD}$  be loss functions for CAAN and ADN, respectively. The loss function for the entire classification network can be expressed as a sum of the two loss



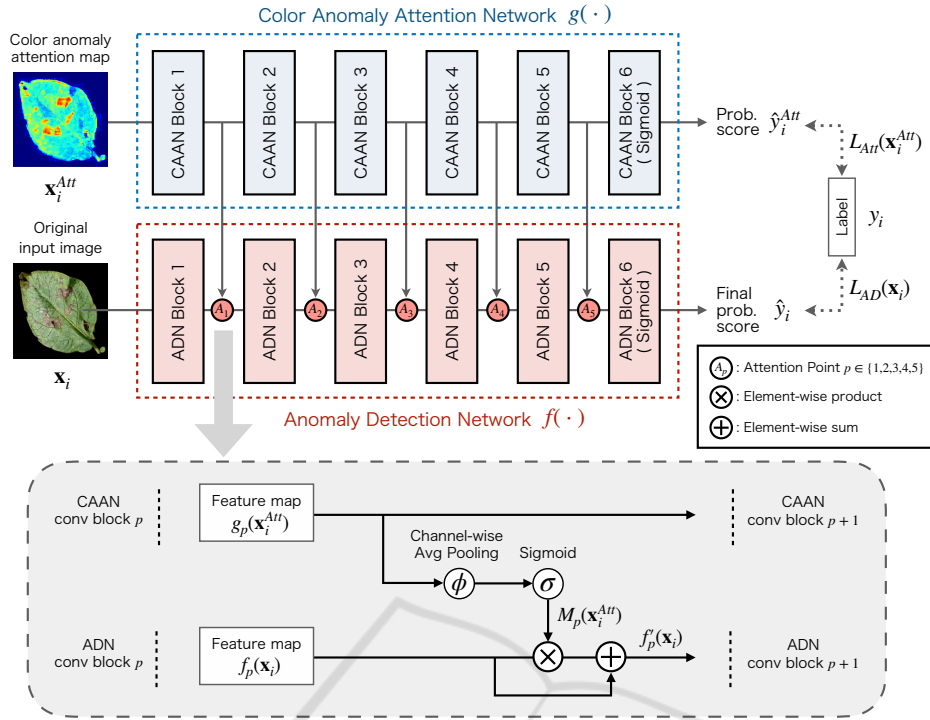


Figure 2: Overall structure of LEA-Net.

functions as:

$$\begin{aligned} L &= L_{Att} + L_{AD} \\ &= BCE(g(\mathbf{x}_i^{Att}), y_i) + BCE(f(\mathbf{x}_i), y_i) \end{aligned} \quad (1)$$

Here,  $g(\cdot)$  and  $f(\cdot)$  denote the outputs of CAAN and ADN, respectively.  $BCE(\cdot)$  denotes the binary cross entropy. We designed the loss function expecting that CAAN modifies attention maps more effectively during training, likewise ABN.

### 3.5 Attention Mechanism

Let us denote the attention map at point  $p \in \{1, 2, 3, 4, 5\}$  by  $M_p \in \mathbb{R}^{H_p \times W_p}$ . The lower half of Figure 2 illustrates the concept of Layer-wise External Attention. Let  $g_p(\mathbf{x}_i^{Att}) \in \mathbb{R}^{H_p \times W_p \times C_p}$  be the feature tensor at attention point  $p$  in CAAN for input image  $\mathbf{x}_i^{Att}$ . Then,  $M_p$  is generated by

$$M_p = \sigma(\phi(g_p(\mathbf{x}_i^{Att}))), \quad (2)$$

where  $\phi(\cdot)$  denotes channel-wise average pooling on the extracted features.  $\sigma(\cdot)$  denotes a sigmoid function. By a channel-wise average pooling on the extracted features, we obtain the single-channel feature map  $\phi(g_p(\mathbf{x}_i^{Att})) \in \mathbb{R}^{H_p \times W_p}$ . Here, we adopted channel-wise average pooling rather than a  $1 \times 1$  convolution layer, expecting the same effect reported by (Woo et al., 2018). A sigmoid layer  $\sigma(\cdot)$  normalizes the feature map  $\phi(g_p(\mathbf{x}_i^{Att}))$  within a range

of  $[0, 1]$ . It was reported that the normalization of the attention map is effective to highlight informative regions (Wang et al., 2017). Additionally, the sigmoid function prevents attention maps and ADN features from the reversal of importance by multiplying negative values. Then, we obtain the attention map  $M_p(\mathbf{x}_i^{Att}) \in \mathbb{R}^{H_p \times W_p}$ .

The role of the attention mechanism is to highlight the informative regions on feature maps, rather than erasing other regions (Wang et al., 2017). To reduce the risk of an informative region erased by attention maps, we incorporate the attention map into ADN as follows:

$$\hat{f}_p(\mathbf{x}_i) = (1 \oplus M_p) \otimes f_p(\mathbf{x}_i), \quad (3)$$

where  $\oplus$  denotes element-wise sum,  $\otimes$  denotes element-wise product, and  $\hat{f}_p(\cdot)$  is the updated feature tensor at point  $p$  in ADN after the Layer-wise External Attention.

The attention strategy described in Eq. 3 is also intended to avoid the Dying Relu Problem (Lu, 2020). The problem is that many parameters with negative values become zero when they are used in the Relu function, which will cause the vanishing gradient problem. In most cases, CAAN and ADN have significantly different feature maps. Additionally, as the layers get deeper, the feature maps tend to be sparser. If such sparse features are simply multiplied at attention points, the performance of ADN will degrade se-

riously. This is another reason why we adopt the attention strategy in Eq. 3 rather than simply multiplying the attention map.

## 4 EXPERIMENTS

We evaluate the performance of Layer-wise External Attention using several datasets for image anomaly detections. Our experiments consist of three main parts. First, we verify the effect of Layer-wise External Attention for boosting the performance of existing image classifiers using several real-world datasets. Second, we examine the performance of Layer-wise External Attention in more imbalanced settings. Further, we attempt to visualize intermediate outputs at attention points before and after the Layer-wise External Attention to understand intuitively the effects of our attention mechanism. Finally, we evaluated whether Layer-wise External Attention is effective for Self-Attention models.

### 4.1 Data Sets

Figure 3 shows the datasets used in this study. We performed the experiment for image anomaly detection using the following datasets: DR2 (Pires et al., 2016), PlantVillage (Hughes and Salathe, 2016), MVTec (Bergmann et al., 2019), and Cloud (ashok, 2020). DR2 contains 435 publicly available retina images with size of 857 pixels. Furthermore, it consists of normal (negative) and diabetic retinopathy (positive) instances. PlantVillage contains images of healthy and diseased leaves of several plants. Among these, we used *Potato* dataset. MVTec AD contains defect-free and anomalous images of various objects and texture categories. Regarding MVTec AD, we used *Leather* and *Tile*, whose anomalies are strongly reflected in color. Cloud dataset contains images without clouds (negative) and with clouds (positive). The lowest panels in Figure 3 represents the anomaly maps. We observed that anomalous regions in the Positive images were failed to reconstruct and highlighted in the color anomaly maps. Before the experiments, all images were resized to  $256 \times 256$  pixels. In order to evaluate the performance on such a practical dataset, we randomly extracted images from each dataset to construct small and imbalanced datasets. Table 2 lists the details of the datasets.

#### 4.1.1 Experimental Setup

We briefly describe the experimental setup used for the training. Parameters of U-Net were optimized by

Table 2: Small and imbalanced experimental datasets reconstructed by random sampling.

Dataset	Positive	Negative	P to N Ratio	Total
<i>Retina</i>	98	337	0.291	435
<i>Potato</i>	50	152	0.329	202
<i>Cloud</i>	100	300	0.333	400
<i>Leather</i>	92	277	0.332	369
<i>Tile</i>	84	263	0.319	347

Adam optimizer with a learning rate of 0.0001. Momentums were set to  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . To update the parameters, the total number of iterations was set to 500 epochs, and the batch size was set to 16. To reduce the computational time, early stopping was used.

In the anomaly detection experiment, we performed stratified five-fold cross-validation on each dataset without data augmentation. The parameters of ResNet18 and VGG16 were both optimized by Adam optimizer with a learning rate of 0.001 and 0.00001, respectively. To update the parameters, the total number of iterations was set to 100 epochs, and the batch size was 16. All computations were performed on a GeForce RTX 2028 Ti GPU based on a system running Python 3.6.9 and CUDA 10.0.130.

#### 4.1.2 Performance in Anomaly Detection

In Layer-wise External Attention, the attention map is incorporated into the intermediate layers of ADN via CAAN. To evaluate the effectiveness of our attention mechanism, we compared the performance of image-level anomaly detection in several settings: (i) baseline models (ResNet18, ResNet50, VGG16 and VGG19), (ii) ADN that receives anomaly map as inputs (*Anomaly Map input*), (iii) ADN that receives 4-channel images consisting of RGB images and anomaly maps as inputs (*4ch input*), (iv) ADN that receives attention maps generated from RGB images and anomaly maps according to Eq. 3 as inputs (*Attentioned input*), and (v) ADN with Layer-wise External Attention at the attention point without CAAN (*Direct Attention*). Table 3 shows the average of  $F_1$  scores with a standard deviation in each experiment. The best and second-best performances are emphasized by bold type. The values given in parentheses indicate the attention point of the best  $F_1$  score. As mentioned above, then Layer-wise External Attention was applied at only one point in each experiment, although the ADN has five attention points. For  $F_1$  scores of models with the Layer-wise External Attention, we described the result of the best attention point score.

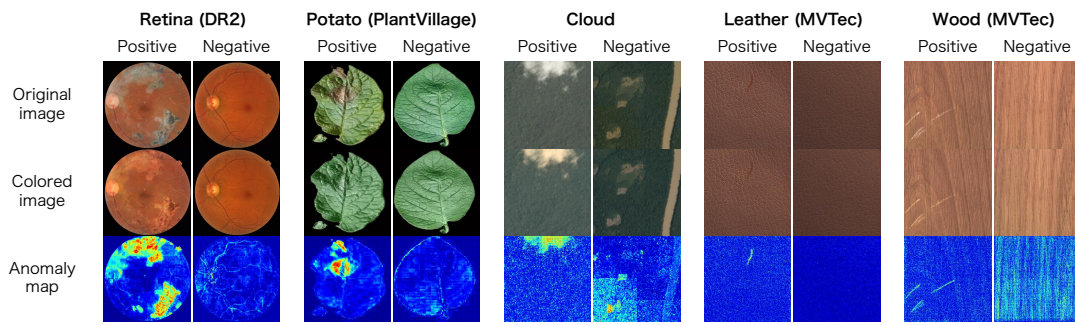


Figure 3: Examples of datasets. For each dataset, the original and colored image and the anomaly map calculated from the two images are shown for Positive and Negative.

The results in Table 3 show that the Layer-wise External Attention consistently increased the  $F_1$  scores. In most of the cases, we can see that CAAN improved the performances of ADN. For the *Potato* dataset, the  $F_1$  score increased by 0.304. Besides, the performance was comparatively low in cases where the anomaly map was directly incorporated into ADN. Interestingly, we must mention here that the most effective attention point disagrees depending on the models and datasets.

### 4.2 Performance in More Imbalanced Data

We performed the detection test for more imbalanced data. Four datasets from DR2 (*Retina*) and PlantVillage (*Potato*) were constructed by changing the proportion of the number of anomalous instances to that of all instances. Then, the performance of LEA-Net was compared to that of ResNet18 and VGG16 without the Layer-wise External Attention. The settings of LEA-Net were as follows: (i) ResNet18-based ADN with MobileNet-based CAAN, (ii) ResNet18-based ADN with ResNet-based CAAN, (iii) VGG16-based ADN with MobileNet-based CAAN, and (iv) VGG16-based ADN with ResNet-based CAAN.

Figure 4 describes the resulting  $F_1$  scores for each imbalanced dataset. The upper two panels show the results for DR2 (*Retina*), while the lower two panels show those for PlantVillage (*Potato*). All  $F_1$  scores were averaged over the five-fold cross-validation. The horizontal axes indicate a proportion of the number of anomalous instances to that of all instances. It can be seen that  $F_1$  scores of the ResNet18-based ADN with CAAN were significantly improved by the layer-wised external attention throughout all imbalanced settings. Also,  $F_1$  scores of VGG16-based ADN with CAAN for PlantVillage (*Potato*) were successfully improved in settings of 24.8% and 12.4%. However, for  $F_1$  scores of VGG16-based ADN with

CAAN for DR2 (*Retina*), recorded no improvement in all settings.

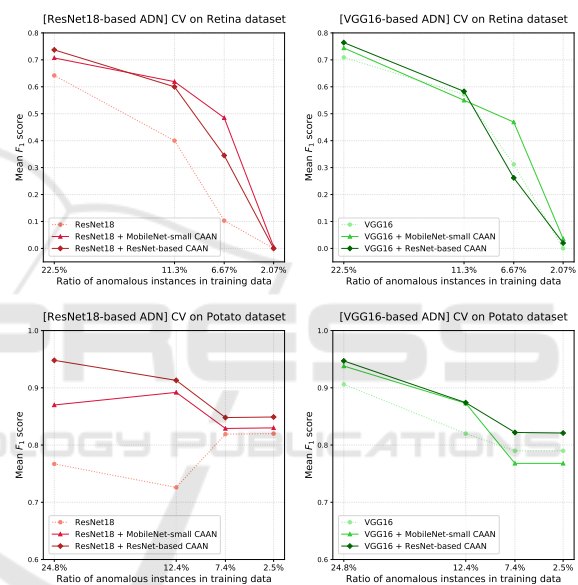


Figure 4: Performances on several imbalanced settings.

### 4.3 Visualization of Attention Effects

Here, we attempted to visualize how the Layer-wise External Attention induces the intermediate outputs of ADN. In Figure 5, we compare intermediate outputs of ADN before and after the Layer-wise External Attention at each attention point. The first column indicates intermediate outputs before the Layer-wise External Attention, the second column indicates the attention map generated by CAAN, and the third column indicates intermediate outputs after the Layer-wise External Attention. DR2 (*Retina*) and PlantVillage (*Potato*) datasets were used for inputs. For both cases, the Layer-wise External Attention notably highlighted anomalous regions from low-level to high-level features.

Table 3: Comparison of  $F_1$  scores on Retina, Potato, Cloud, Leather, and Wood datasets.

Model	Params	Retina	Potato	Cloud	Leather	Wood
ResNet18	11.2M	0.642 ± 0.097	0.767 ± 0.164	0.642 ± 0.081	0.756 ± 0.147	0.560 ± 0.230
ResNet50	23.6M	0.544 ± 0.155	0.425 ± 0.315	0.635 ± 0.147	0.631 ± 0.152	0.546 ± 0.146
ResNet18 (AnomalyMap input)	11.2M	0.412 ± 0.005	0.779 ± 0.181	0.476 ± 0.058	0.551 ± 0.300	0.329 ± 0.137
ResNet18 (4ch input)	11.2M	0.443 ± 0.054	0.850 ± 0.102	0.529 ± 0.050	<b>0.883 ± 0.066</b>	0.495 ± 0.322
ResNet18 (Attentioned input)	11.2M	0.616 ± 0.108	0.823 ± 0.157	0.532 ± 0.050	0.752 ± 0.035	0.544 ± 0.22
ResNet18 + Direct Attention	11.2M	<b>0.718 ± 0.047(3)</b>	0.884 ± 0.045(4)	<b>0.670 ± 0.084(3)</b>	0.823 ± 0.055(3)	<b>0.697 ± 0.061(4)</b>
<b>ResNet18 + MobileNet-like CAAN</b>	13.5M	0.707 ± 0.037(1)	<b>0.870 ± 0.094(4)</b>	<b>0.707 ± 0.131(2)</b>	0.826 ± 0.062(1)	<b>0.719 ± 0.247(2)</b>
<b>ResNet18 + ResNet-like CAAN</b>	16.1M	<b>0.737 ± 0.045(1)</b>	<b>0.948 ± 0.039(4)</b>	0.664 ± 0.067(4)	<b>0.828 ± 0.052(4)</b>	0.621 ± 0.142(1)
VGG16	165.7M	0.709 ± 0.119	0.906 ± 0.061	0.575 ± 0.046	0.862 ± 0.089	0.652 ± 0.142
VGG19	171.0M	0.692 ± 0.068	<b>0.941 ± 0.04</b>	<b>0.794 ± 0.041</b>	0.758 ± 0.08	0.597 ± 0.338
VGG16 (AnomalyMap input)	165.7M	0.399 ± 0.012	0.899 ± 0.093	0.438 ± 0.023	<b>0.907 ± 0.039</b>	0.397 ± 0.111
VGG16 (4ch input)	165.7M	0.458 ± 0.031	0.823 ± 0.073	0.479 ± 0.092	<b>0.925 ± 0.033</b>	0.575 ± 0.042
VGG16 (Attentioned input)	165.7M	0.726 ± 0.056	<b>0.947 ± 0.039</b>	0.576 ± 0.119	0.877 ± 0.053	0.593 ± 0.269
VGG16 + Direct Attention	165.7M	<b>0.754 ± 0.060(5)</b>	0.927 ± 0.030(1)	<b>0.596 ± 0.072(4)</b>	0.877 ± 0.102(3)	<b>0.746 ± 0.069(3)</b>
<b>VGG16 + MobileNet-like CAAN</b>	168.0M	0.744 ± 0.068(3)	0.938 ± 0.058(1)	0.577 ± 0.069(3)	0.867 ± 0.077(2)	0.707 ± 0.074(4)
<b>VGG16 + ResNet-like CAAN</b>	170.6M	<b>0.764 ± 0.069(5)</b>	<b>0.947 ± 0.064(1)</b>	0.584 ± 0.076(5)	0.899 ± 0.091(3)	<b>0.762 ± 0.035(5)</b>

#### 4.4 External Attention for Self-Attention Models

Finally, we evaluated whether Layer-wise External Attention can be used with Self-Attention for image

anomaly detection. In Figure 6, we compare the detection performances of Self-Attention models with and without Layer-wise External Attention. Here, SE module (Hu et al., 2020b), SimAM module (Yang et al., 2021), and SRM module (Hyunjae et al., 2019)



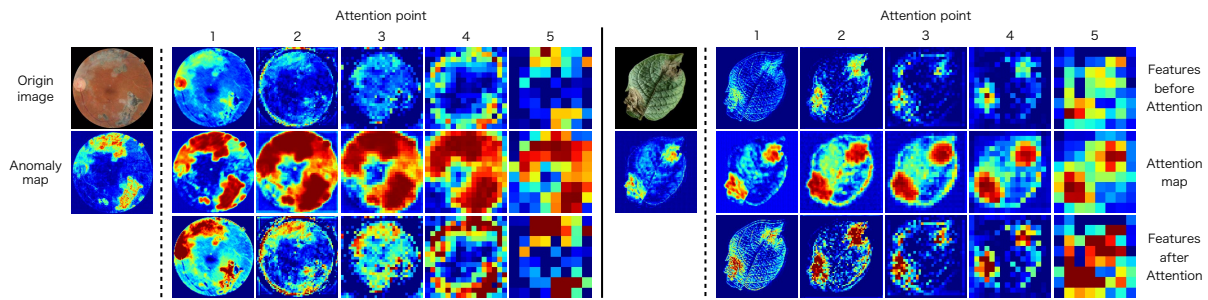


Figure 5: Visualization of intermediate outputs at attention points before and after the Layer-wise External Attention.

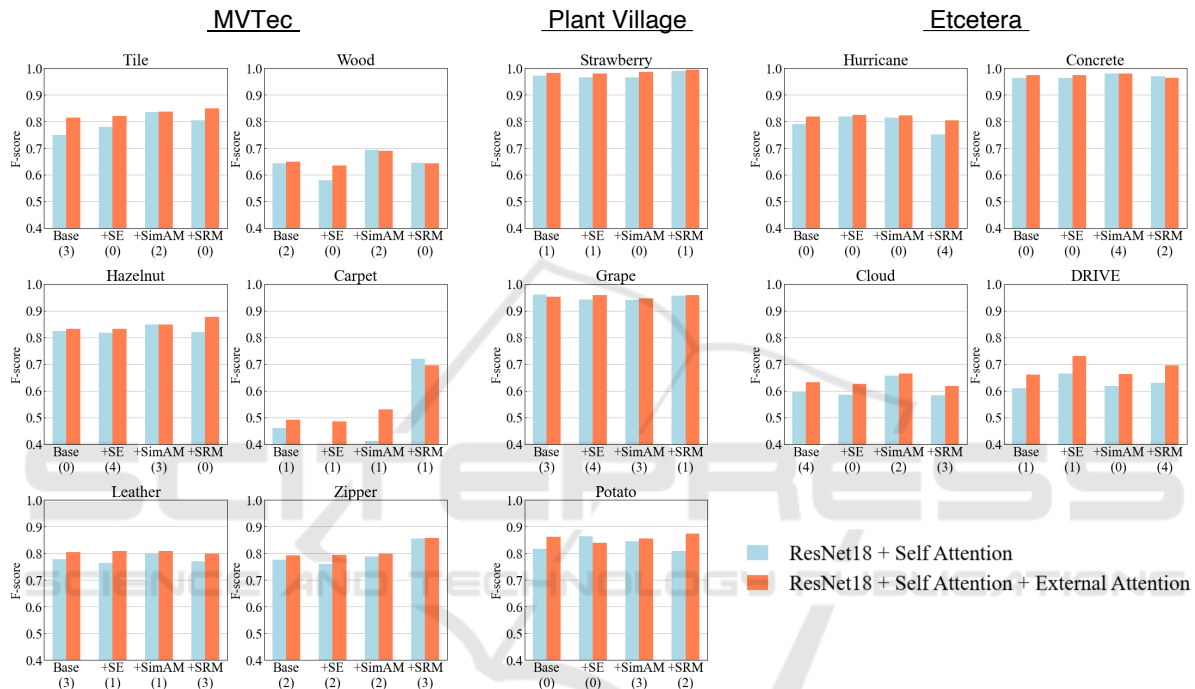


Figure 6: Effects of Layer-wise External Attention for Self-Attention Models.

were attached to ADN for Self-Attention mechanism. Numbers in parentheses indicate the adopted attention point. The blue bars indicate the averages of  $F_1$  scores of Self-Attention models without the Layer-wise External Attention for MVTEC, PlantVillage, Hurricane, Concrete, Cloud and DR2 (DRIVE) datasets. The red bars indicate those of Self-Attention models with the Layer-wise External Attention. The results clearly show that the Layer-wise External Attention successfully boosts detection performances of Self-Attention models.

## 5 CONCLUSION

In this paper, we proposed a concept of Layer-wise External Attention for efficient image anomaly detection. Especially, our concern was on whether

the introduction of Layer-wise External Attention enables us to utilize prior knowledge about anomalies for deep anomaly detection. Through comprehensive experiments using real-world datasets, we demonstrated that the Layer-wise External Attention successfully boosts anomaly detection performances of CNNs. Also, we found that the benefits are still active and working even for small and imbalanced training data. For an additional experiment, we attempted to attach the Layer-wise External Attention module on Self-Attention models. Somewhat surprisingly, the anomaly detection performances of Self-Attention models were also improved by Layer-wise External Attention, which suggests that Layer-wise External Attention complementary works with Self-Attention. So we conclude that Layer-wise External Attention mechanism has an enough potential to be the new trend in visual attention.

The present study has not yet clarified the most reasonable way in generating attention map. Results shown in Figure 5 implies that External Attention sometimes emphasizes noises. Such excessive attentions could inhibit training progresses of ADN. So we need to establish some methodologies for adjusting the strength of external attention at each attention point. One simple idea is to introduce sparsity assumptions on attention map. At the same time, the effectiveness of Layer-wise External Attention most likely depends on the quality of anomaly map. In experiments that we conducted in this study, we use CAAN to generate anomaly map for all datasets. However, for more practical use, it is reasonable to choose the most appropriate Anomaly Attention Network for each dataset and task. The LEA-Net system was originally designed to appropriately replace Anomaly Attention Network to suit for the target domain. In this regard, we can say that Layer-wise External Attention is more versatile system in comparison with Self-Attention.

## REFERENCES

- Akçay, S., Atapour-Abarghouei, A., and Breckon, T. P. (2018). Ganomaly: semi-supervised anomaly detection via adversarial training.
- ashok (2020). Cloud and non-cloud images (anomaly detection). <https://www.kaggle.com/ashoksrinivas/cloud-anomaly-detection-images>. Accessed: 2020-08-22.
- Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. (2019). MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X., and Xie, Z. (2018). Deep learning and its applications in biomedicine. *Genomics, proteomics & bioinformatics*, 16(1):17–32.
- Chun, M. M., Golomb, J. D., and Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual review of psychology*, 62:73–101.
- Emami, H., Aliabadi, M. M., Dong, M., and Chinnam, R. B. (2020). Spa-gan: Spatial attention gan for image-to-image translation. *IEEE Transactions on Multimedia*, 23:391–401.
- Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318.
- Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2019). Attention branch network: Learning of attention mechanism for visual explanation.
- Haselmann, M., Gruber, D. P., and Tabatabai, P. (2018). Anomaly detection using deep learning based image completion.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network.
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., and Adam, H. (2019). Searching for mobilenetv3.
- Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2020a). Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023.
- Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2020b). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023.
- Hughes, D. P. and Salathe, M. (2016). An open access repository of images on plant health to enable the development of mobile disease diagnostics.
- Hyunjae, L., Hyo Eun, K., and Hyeonseob, N. (2019). Srm: A style-based re-calibration module for convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*.
- Katafuchi, R. and Tokunaga, T. (2021). Image-based Plant Disease Diagnosis with Unsupervised Anomaly Detection based on Reconstructability of Colors. pages 112–120.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.
- Lee, H., Kim, H. E., and Nam, H. (2019). SRM: A style-based recalibration module for convolutional neural networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:1854–1862.
- Li, W., Cheng, S., Qian, K., Yue, K., and Liu, H. (2021). Automatic Recognition and Classification System of Thyroid Nodules in CT Images Based on CNN. *Computational Intelligence and Neuroscience*, 2021.
- Lu, L. (2020). Dying relu and initialization: Theory and numerical examples. *Communications in Computational Physics*, 28(5):1671–1706.
- Minhas, M. S. and Zelek, J. (2019). Anomaly detection in images. *arXiv preprint arXiv:1905.13147*.
- Natarajan, V., Mao, S., and Chia, L.-T. (2019). Salient textural anomaly proposals and classification for metal surface anomalies. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (IC-TAI)*, pages 621–628.
- Nico Goernitz, Marius Micha Kloft, K. R. U. B. (2013). Toward supervised anomaly detection. *Journal Of Artificial Intelligence Research (JAIR)*, 46:235–262.
- Pires, R., Jelinek, H. F., Wainer, J., Valle, E., and Rocha, A. (2016). Advancing Bag-of-Visual-Words Representations for Lesion Classification in Retinal Images.
- Reynolds, J. H. and Chelazzi, L. (2004). Attentional modulation of visual processing. *Annu. Rev. Neurosci.*, 27:611–647.

- Rezvantalab, A., Safigholi, H., and Karimijeshni, S. (2018). Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms. *arXiv preprint arXiv:1810.10348*.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.
- Schlegl, T., Seebock, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157.
- Sharma, G., Wu, W., and Dalal, E. N. (2005). The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations. *Color Research and Application*, 30(1).
- Venkataramanan, S., Peng, K.-C., Singh, R. V., and Mahalanobis, A. (2020). Attention guided anomaly localization in images. In *European Conference on Computer Vision*, pages 485–503. Springer.
- Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., and Tang, X. (2017). Residual attention network for image classification. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua(1):6450–6458.
- Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 11531–11539.
- Woo, S., Park, J., Lee, J. Y., and Kweon, I. S. (2018). CBAM: Convolutional block attention module. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11211 LNCS:3–19.
- Yang, L., Zhang, R.-Y., Li, L., and Xie, X. (2021). Simam: A simple, parameter-free attention module for convolutional neural networks. In *International Conference on Machine Learning*, pages 11863–11874. PMLR.
- Zagoruyko, S. and Komodakis, N. (2017). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pages 1–13.
- Zenati, H., Foo, C.-S., Lecouat, B., Manek, G., and Chandrasekhar, V. R. (2018). Efficient gan-based anomaly detection. In *6th International Conference on Learning Representations (ICLR2018)*.
- Zhao, H., Jia, J., and Koltun, V. (2020). Exploring Self-attention for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 10073–10082.