# SAGAN: SKIP-ATTENTION GAN FOR ANOMALY DETECTION

*Guoliang Liu*[†], *Shiyong Lan*[⋆†*], *Ting Zhang*[†], *Weikang Huang*[†], *Wenwu Wang*[‡]

[⋆]College of Computer Science, Sichuan University, China.
[†] National Defense Key Laboratory for Synthetic Vision Graphics and Imaging Science, China.
[‡]Center for Vision Speech and Signal Processing, University of Surrey, UK.

2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506332

## ABSTRACT

Generative Adversarial Networks (GANs) have been used recently for anomaly detection from images, where the anomaly scores are obtained by comparing the global difference between the input and generated image. However, the anomalies often appear in local areas of an image scene, and ignoring such information can lead to unreliable detection of anomalies. In this paper, we propose an efficient anomaly detection network Skip-Attention GAN (SAGAN), which adds attention modules to capture local information to improve the accuracy of latent representation of images, and uses depth-wise separable convolutions to reduce the number of parameters in the model. We evaluate the proposed method on the CIFAR-10 dataset and the LBOT dataset (built by ourselves), and show that the performance of our method in terms of area under curve (AUC) on both datasets is improved by more than 10% on average, as compared with three recent baseline methods.

***Index Terms***— Anomaly Detection, Generative Adversarial Networks, Attention, Depth-wise Separable Convolutions

## 1. INTRODUCTION

Anomaly detection is an increasingly important area in computer vision, and has been extensively studied in many application fields, such as industrial anomaly detection, fraud detection and medical applications [1]. However, anomaly detection suffers from several unique challenges in practical applications. Firstly, compared with normal data, there is less amount of abnormal data available, as it is often more difficult to obtain abnormal data than normal data. Secondly, anomalies are often unpredictable, which makes it difficult to accurately define the appearance attributes of anomalous objects. Due to these challenges, supervised learning methods are often limited in anomaly detection. In contrast, unsupervised learning methods are often used to learn the distribution of normal data, where only normal data is used for training, while both normal and abnormal data are used in testing.

Recently, Generative Adversarial Networks (GANs) [2] based unsupervised learning methods [3, 4, 5, 6] have been employed for anomaly detection, showing promising performance. GAN typically consists of two modules, namely, a generator and a discriminator. The generator constantly generates images that are as realistic as possible to fool the discriminator, and the discriminator constantly tries to distinguish between the real image and the generated image. The unique network structure of GAN makes it suitable not only for encoding the image to obtain its latent representations, but also for decoding and generating image with minimal information loss via a reverse pass [7].

AnoGAN [4] is the first GANs-based method for anomaly detection, which aims to learn the mapping from a latent representation to the generated image. However, AnoGAN is computationally expensive in searching for latent representations. In contrast, EGBAD [5] does not need to search for latent representations, as it learns a mapping directly from image space to latent space. GANomaly [6] combines encoder and decoder in the generator and detects anomalies by comparing the latent representation of the input image with the latent representation of the generated image. Skip-GANomaly [8] adds the U-Net [9] structure to the generator of GANomaly, and uses skip connections to link the encoder in each layer with the decoder in the corresponding layer, which improves the reconstruction of an image from its latent representations.

In all these existing GAN-based methods, however, the anomalies are detected by comparing the difference between the global information of the input image and the generated image. In fact, relying only on global information can impact adversely on the accuracy of the latent representations of abnormal samples. Recently, some works [10, 11, 12, 13] have used the attention mechanism to capture the relevant information and suppress redundant information in the image. For example, the convolutional block attention module (CBAM) [10], which is an attention module obtained by mixing the spatial and channel attention modules, was shown to perform better than SENet [11], which uses only channel attention.

In this paper, we introduce a new method for anomaly

* Corresponding author. E-mail: lanshiyong@scu.edu.cn.

This work was funded in part by the grant 2021YFG0300 of Sichuan Science and Technology Department, China, and in part by National Defense Key Laboratory for Synthetic Vision Graphic and Imaging Science, Sichuan University, China.

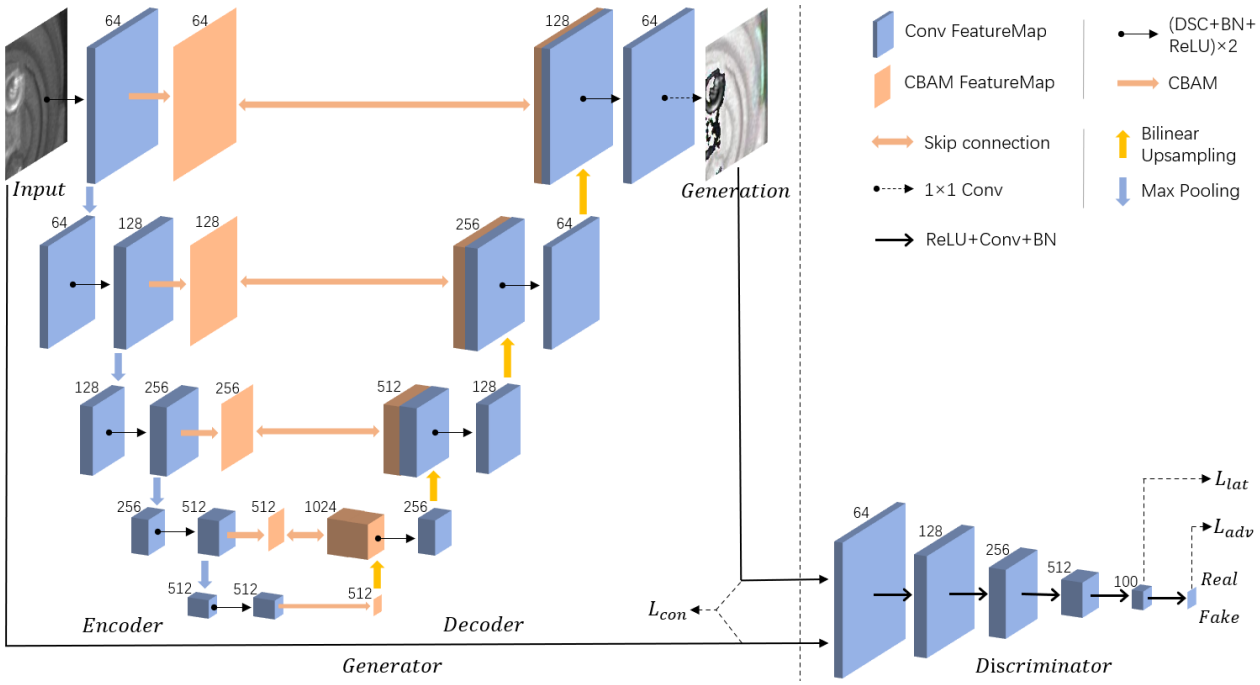



**Fig. 1**. Details of the proposed model architecture. The number above the FeatuteMap indicates the number of channels.

detection, called Skip-Attention GAN (SAGAN), which uses CBAMs to capture local information, within the Skip-GANomaly method. Our method is motivated by Attention U-Net [13], where an attention mechanism was added to the U-Net model, and was shown to perform better on image segmentation than the original U-Net model. In addition, we use depth-wise separable convolutions (DSCs), as in [14, 15, 16], to reduce the number of model parameters to improve the computational efficiency of our method.

## 2. PROPOSED METHOD

### 2.1. Model Overview

Our model structure is shown in Figure 1, which is based on the Skip-GANomaly [8]. Different from the Skip-GANomaly, we introducing CBAMs and DSCs in the generator, replace the convolution with global pooling in the downsampling process, and replace the transpose convolution with bilinear interpolation in the upsampling process.

In Figure 1, the left side shows the generator, which consists of a U-shaped encode-decoder structure. In the encoder of the generator, the feature map of each layer will first be passed through two DSCs to double the number of channels within the feature maps, and then passed through CBAM to obtain the CBAM feature map, before applying max pooling to halve its size. The CBAM feature map of each layer will be connected to the feature map in the corresponding layer in the decoder through a skip connection to obtain a new feature map with attention information. In the decoder of the generator, the feature map of each layer will first be passed two DSCs to halve the number of channels within the feature maps, and then to double their size through upsampling. Finally, the generator outputs the the generated image through a 1×1 convolution. With this method, more attention is paid to the relevant areas in the image.

The right side of Figure 1 shows the discriminator, which uses the same network structure as in DCGAN [7] to extract the latent representation of input image and distinguish whether the input image is a real image or a generated image.

### 2.2. Attention Mechanism

CBAM [10] is a mixed attention mechanism, in which the channel attention module and spatial attention module are concatenated in a specific order. The channel attention module models the importance of each feature channel, and then either enhances or suppresses different channels for different images. The goal of the spatial attention module is to locate the region of interest in the image and obtain the weight distribution map for the image.

In our proposed model, we incorporate CBAM into each layer of the encoder in the generator to obtain the attention information of the corresponding layer. The feature maps of each layer are then passed through the channel module and the spatial attention module in turn to obtain channel and spatial attention information embedded feature maps.

### 2.3. Depth-wise Separable Convolution

A network with fewer parameters is less likely over-fit with the training set. However, reducing the number of network parameters may also lead to an over-simplified network, thus degrading the learning performance. To achieve a trade-off between the model complexity and learning performance, depth-wise separable convolutions (DSCs) have been developed recently [14, 15, 16]. In this method, the conventional convolution operation is divided into per-channel and per-point convolutions, so that the network performance is retained while the network complexity (i.e the number of parameters) is reduced.

| Model | Parameters |
|---|---|
| SAGAN without DSCs | 18,498,481 |
| SAGAN with DSCs | 5,181,089 |

**Table 1**. Number of parameters in SAGAN with/without DSCs (used in our experiments).

As shown in Table 1, after adding DSCs, the number of parameters in our proposed SAGAN is greatly reduced.

### 2.4. Model Training

We introduce the same loss functions $L_{con}$, $L_{lat}$ and $L_{adv}$ as in Skip-GANomaly [8] at the positions indicated in Figure 1.

$$L_{con} = E_{x\sim p_x}|x - G(x)|_1, \tag{1}$$

$$L_{lat} = E_{x\sim p_x}|D(x) - D(G(x))|_2, \tag{2}$$

$$L_{adv} = E_{x\sim p_x}[logD(x)] + E_{x\sim p_x}[log(1 - D(G(x))], \tag{3}$$

where $L_{adv}$ is a loss function commonly used in GAN, in which the generator $G$ and the discriminator $D$ are optimized in an alternating manner through adversarial learning. The loss function $L_{con}$ is used for image reconstruction, which aims to further enhance the generated image $G(x)$ on the basis of $L_{adv}$, so that it is similar to the input image $x$. The loss $L_{lat}$ is obtained by the discriminator and represents the difference between the latent representation $z = D(x)$ of the input image $x$ and the latent representation $\hat{z} = D(G(x))$ of the generated image $G(x)$. The purpose of using $L_{lat}$ is to maintain as much consistency between $z$ and $\hat{z}$ as possible.

The overall training objective $L$ is the weighted sum of $L_{con}$, $L_{lat}$ and $L_{adv}$ as follows:

$$L = \lambda_{con}L_{con} + \lambda_{lat}L_{lat} + \lambda_{adv}L_{adv}, \tag{4}$$

where $\lambda_{con}$, $\lambda_{lat}$ and $\lambda_{adv}$ are the weight parameters of the individual loss functions in the overall loss function.

### 2.5. Anomaly Scores

We use the method in [4, 5, 8] to obtain the anomaly score for the test image as follows:

$$A(x) = \lambda R(x) + (1 - \lambda)L(x), \tag{5}$$

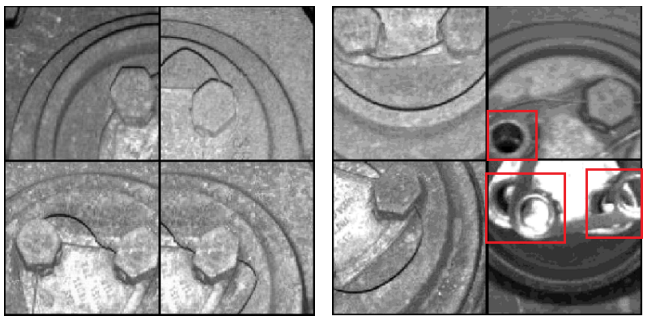
where $x$ represents the test image in the test set, and $R(x)$ is the difference between the input test image and the generated image, $L(x)$ is the difference between the latent representation of the input image and the latent representation of the generated image, $A(x)$ is the raw anomaly score of the test sample $x$, and $\lambda$ is a weighting parameter that controls the importance of $R(x)$ and $L(x)$ in $A(x)$.

According to Eq. (5), we calculate the raw anomaly scores of all the test samples in the test set, and use the vector $A$ to represent the set of anomaly scores of all the samples in the test set. Then, we use the same method as in [4] to compress each anomaly score into the range [0,1], that is, the final anomaly score $\hat{A}(x)$ of a single test sample $x$ is expressed as:

$$\hat{A}(x) = \frac{A(x) - min(A)}{max(A) - min(A)}. \tag{6}$$

## 3. EXPERIMENTAL SETUP

We have performed experiments on a Ubuntu16.04 server with 32Gb memory and a single NVIDIA RTX2080Ti GPU, and evaluated the proposed method on two datasets (the CIFAR-10 [17] dataset and the train axle bolt LBOT dataset that we constructed). Next, we introduce the two datasets, training details, the evaluation metric, before presenting the results.



(a) train data (b) test data

**Fig. 2**. Some image examples from the LBOT dataset, (a) images from the training set, and (b) images from the test set. The abnormal areas of bolts are marked in the red box.

**LBOT**: The LBOT dataset is constructed by ourselves. It is derived from a study on the train axle bolt status inspection, which defines the missing or damaged bolts as abnormal. This dataset includes 5,000 image patches of the train axle bolt status extracted by the 128×128 overlapping sliding window

| Model | CIFAR-10 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | frog | bird | cat | deer | dog | horse | ship | truck | Average |
| EGBAD [5] | 0.512 | 0.523 | 0.466 | 0.467 | 0.502 | 0.387 | 0.534 | 0.579 | 0.496 |
| GANomaly [6] | 0.777 | 0.552 | 0.647 | 0.684 | 0.815 | 0.683 | 0.818 | 0.844 | 0.728 |
| Skip-GANomaly [8] | 0.955 | 0.611 | 0.670 | 0.845 | 0.706 | 0.666 | 0.909 | 0.857 | 0.777 |
| proposed | **0.996** | **0.957** | **0.951** | **0.998** | **0.975** | **0.891** | **0.990** | **0.980** | **0.967** |

**Table 2**. The AUC results obtained on the CIFAR-10 dataset.

| Model | AUC |
|---|---|
| EGBAD [5] | 0.489 |
| GANomaly [6] | 0.900 |
| Skip-GANomaly [8] | 0.840 |
| SAGAN without DSCs | **0.960** |
| SAGAN with DSCs | **0.958** |

**Table 3**. The AUC results on the LBOT dataset.

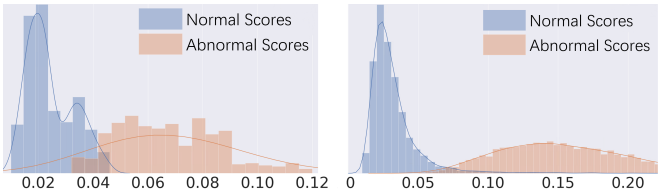



**Fig. 3**. The left histogram shows the distribution of normal and abnormal scores for the test data in the LBOT dataset. The right histogram shows the distribution of normal and abnormal scores of the test data in the CIFAR-10 dataset when the frog is defined as an anomaly.

method. As shown in Figure 2, we divided the LBOT dataset into 4,000 training images and 1,000 test images according to the ratio of 4:1. The 4,000 training images are all normal bolt images, and the 1,000 test images contain 500 normal bolt images and 500 abnormal bolt images.

**CIFAR-10**: Both GANomaly [6] and Skip-GANomaly [8] used the CIFAR-10 dataset and formulated a leave one class out anomaly detection problem. For comparison, we also used this dataset. Similar to [5, 6, 8], we divide the CIFAR-10 dataset into 8 different categories, each category has 45,000 normal training samples, 9,000 normal test samples and 6,000 abnormal test samples. Before training, one of the categories was defined as abnormal, and the other categories as normal.

**Training Details**: During the training process, the weighting parameter of $L$ in Eq. (4) is set to $\lambda_{adv} = 1$, $\lambda_{con} = 40$ and $\lambda_{lat} = 1$. The objective function $L$ is optimized by Adam [18], and the initial learning rate of Adam is $lr = 2 \times 10^{-3}$ with a lambda decay, and momentum $\beta_1 = 0.5$, $\beta_2 = 0.999$. To calculate the anomaly scores, we set $\lambda = 0.2$ in Eq. (5).

**Evaluation Metric**: We use the area under the curve (AUC) of the receiver operating characteristic (ROC) [19] as performance metric, as in [5, 6, 8].

**Baseline Methods**: We compare our method with three baseline methods, i.e. EGBAD [5], GANomaly [6], and Skip-GANomaly [8], respectively.

## 4. EXPERIMENTAL RESULTS

The results on the CIFAR-10 dataset are shown in Table 2. It can be seen that in each abnormal case, the results of the proposed method are better than the baseline methods. Especially when the anomaly category is bird, our proposed method obtains an AUC of 0.957, which is improved by more than 0.3 compared with the highest AUC of 0.611 of baseline methods.

Table 3 shows the experimental results of our proposed method SAGAN and baseline methods on the LBOT dataset. Again, SAGAN performs better than the baseline methods, and the adverse impact of adding DSCs is negligible.

Figure 3 shows the score histograms obtained by our proposed method from the CIFAR-10 dataset and the LBOT dataset. It can be seen from the Figure 3 that there are significant differences in the distribution between the normal and abnormal scores of the two datasets, which indicates that the method can separate the normal and abnormal scores well.

## 5. CONCLUSION

We have presented an anomaly detection method, i.e. SAGAN, on the basis of the Skip-GANomaly model, by incorporating an attention module and depth-wise separable convolutions. We found that adding CBAMs and DSCs to the generator of the U-Net structure allows the generator to efficiently generate images that emphasize key areas, and with the generated image, the discriminator can extract a more accurate latent representation. Our experiments on CIFAR-10 and LBOT datasets show that our method outperforms the state-of-the-art GAN-based anomaly detection methods [5, 6, 8].

## 6. REFERENCES

[1] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.

[3] F. Di Mattia, P. Galeone, M. De Simoni, and E. Ghelfi, "A survey on gans for anomaly detection," *arXiv preprint arXiv:1906.11632*, 2019.

[4] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 146–157.

[5] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient gan-based anomaly detection," *arXiv preprint arXiv:1802.06222*, 2018.

[6] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 622–637.

[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[8] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[10] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[11] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 2011–2023, 2020.

[12] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Quoc V. Le, "Attention augmented convolutional networks," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3285–3294, 2019.

[13] O. Oktay, J. Schlemper, L. L. Folgoc, M. J. Lee, M. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *ArXiv*, vol. abs/1804.03999, 2018.

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[15] Y. Guo, Y. Li, R. Feris, L. Wang, and T. Simunic, "Depthwise convolution is all you need for learning multiple visual domains," in *AAAI*, 2019.

[16] P. K. Gadosey, Yujian L, Enock Adjei Agyekum, Ting Z, Zhaoying L, Peter T. Yamak, and Firdaous E, "Sd-unet: Stripping down u-net for segmentation of biomedical images on platforms with low computational budgets," *Diagnostics*, vol. 10, 2020.

[17] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," Online: http://www. cs. toronto. edu/kriz/cifar. html, 2014.

[18] D. Kingma and J. B, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] C. Ling, J. Huang, and H. Zhang, "Auc: a statistically consistent and more discriminating measure than accuracy," in *IJCAI*, 2003.