A Transformer-based GAN for Anomaly Detection*

Caiyin Yang†1[0000−0002−9057−4256], Shiyong Lan*1, Weikang Huang†1, Wenwu Wang[2], Guoliang Liu[1], Hongyu Yang[1], Wei Ma[1], and Piaoyang Li[1]

[1] College of Computer Science, Sichuan University, Chengdu 610065, China
lanshiyong@scu.edu.cn
[2] University of Surrey, Guildford, GU2 7XH, United Kingdom
w.wang@surrey.ac.uk

Abstract. Anomaly detection is the task of detecting outliers from normal data. Numerous methods have been proposed to address this problem, including recent methods based on generative adversarial network (GAN). However, these methods are limited in capturing the long-range information in data due to the limited receptive field obtained by the convolution operation. The long-range information is crucial for producing distinctive representation for normal data belonging to different classes, while the local information is important for distinguishing normal data from abnormal data, if they belong to the same class. In this paper, we propose a novel Transformer-based architecture for anomaly detection which has advantages in extracting features with global information representing different classes as well as the local details useful for capturing anomalies. In our design, we introduce self-attention mechanism into the generator of GAN to extract global semantic information, and also modify the skip-connection to capture local details in multi-scale from input data. The experiments on CIFAR10 and STL10 show that our method provides better performance on representing different classes as compared with the state-of-the-art CNN-based GAN methods. Experiments performed on MVTecAD and LBOT datasets show that the proposed method offers state-of-the-art results, outperforming the baseline method SAGAN by over 3% in terms of the AUC metric.

Keywords: Anomaly Detection · Transformer · Generative Advertise Network

1 Introduction

Anomaly detection is an important field in computer vision. The detection of abnormal images plays an increasingly important role due to the growing demand

* This work was funded in part by the Key R&D Project of Sichuan Science and Technology Department, China (2021YFG0300), and in part by 2035 Innovation Pilot Program of Sichuan University, China.
* Corresponding author. E-mail:lanshiyong@scu.edu.cn.
† Equal contribution.

in various applications, such as video surveillance, risk management and damage detection [8, 14]. Current state of the art methods in this area are based on deep learning methods such as Deep-anomaly [17] and ADCNN [10]. However, the performance of these methods is limited by the lack of labelled data. On the one hand, it is hard to collect abnormal images due to unbalanced distribution of normal and abnormal data. On the other hand, abnormal data is difficult to be defined clearly [15]. To solve these problems, a number of abnormal detection methods have been proposed based on unsupervised learning which consider anomaly detection as a one-class classification problem [20]. These methods learn the feature distribution of normal data and the data whose distribution is substantially different from the learned distribution in terms of a predefined threshold is regarded as containing abnormal objects.

A recent method for anomaly detection is based on unsupervised learning with generative adversarial network (GAN) [9]. The adversarial learning process facilitates the generator to learn normal data distribution [6]. AnoGAN [19] is the first GAN-based representation learning method for anomaly detection [16]. In EBGAN [25] and Fast-AnoGAN [18], a network is built to learn feature representations in a latent space with an inverse of the generator. GANomaly [1] introduced an encoder-decoder-encoder network for the generator to learn image representations within the latent space of images. Skip-GANomaly [2] uses an U-net structure as the generator to improve detection performance. SAGAN [12] uses an attention module in skip connection to capture additional local information. However, due to the limited receptive field induced by the convolution operation, the aforementioned methods can only model local information but are limited in capturing long-range information within the data, thus are ineffective in detecting the abnormal information distributed both locally and globally.

The self-attention mechanism in transformer has been widely used in computer vision tasks, offering state-of-the-art performance. The attention module in a transformer can associate the input sequence to learn long range information globally. ViT [7] firstly applied a transformer to computer vision tasks by directly processing image as patch sequences. Swin Transformer [13] proposed a method to calculate the attention in local windows to reduce its computational complexity. U-shaped transformers which are similar to SwinUnet [4] and Uformer [23] have also been proposed. These models take advantage of the self-attention mechanism in capturing long range dependency that CNN lacks for representation learning.

In this paper, we propose a novel anomaly detection framework to address the limitation of CNN in modelling the long-range information within the data by leveraging the strength of the transformer model. We build our model based on GAN, and introduce the self-attention mechanism to capture long-range information within the image data. The key idea of our framework is that we build an Unet-Shaped Encoder-Decoder structure with Transformer-based blocks. A limitation with the transformer model is that it may ignore local information while learning long-range dependencies. To address this issue, we propose to modify the skip-connection for capturing multi-scale information from local fea-

tures. Experimental results show that our method outperforms state-of-the-art anomaly detection methods on datasets such as CIFAR10 [11], and STL10 [22] on outer-class task, LBOT [12] and MVTecAD [3] considering inter-class task, additionally. The main contributions of this paper are summarized as follows:

- We propose a novel anomaly detection framework AnoTrans which combines the Transformer-based module with existing GAN-based method to address the limitation of the CNN encoder used in GAN-based method for modelling the long range information within image data.
- We design a new method for fusing the global attention with the local attention, which enables the global and local information to be captured simultaneously when performing anomaly detection.
- Experimental results on four datasets, CIFAR10, STL10, LBOT and MVTecAD, respectively, demonstrate the superiority of our method over the state-of-the-art CNN-based methods in anomaly detection.

2 Proposed Method

2.1 Model Overview

To enhance the feature representation with global information as well as the local details, we propose AnoTran (Fig. 1) based on SAGAN [12], where an attention module (i.e., CBAM [24]) is incorporated into the depth-wise CNN-based encoder of GAN to enhance the latent representation of input images. To introduce long-range dependency within the representation, we replace the convolution block by the self-attention module as used in the encoder of the transformer.

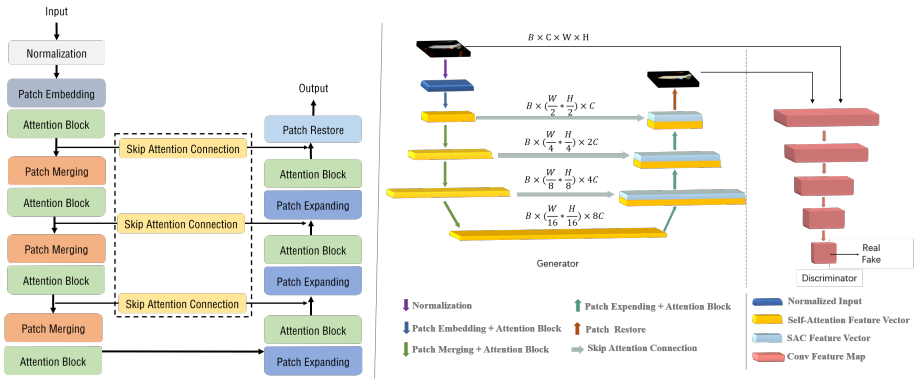


Fig. 1: The structure of our proposed model, where the attention block is used to replace the convolution module. The left part is the logical composition of the generator, and the right part is the overall network structure of our model.

In addition, we design a new skip attention connection, which introduces an attention mechanism to modify skip connections in SAGAN. In our proposed attention connection, the global dependencies captured by the Transformer-based

module can be used to relate each local feature, so as to enhance the feature representation of the input image. Moreover, we add a batch normalization in the beginning of the input, which can mitigate the impact of the overall offset of each batch of the input data, thus facilitate the generator to learn a better representation for normal data.

As shown in Fig. 1, our model is composed of a generator and discriminator. The generator is implemented by a U-shaped encoder-decoder structure which will be described in detail in Section 2.2. The Transformer-based self attention module is used to replace the convolution in the encoder of the generator, as discussed in Section 2.3. The improved skip connection module with self-attention mechanism is introduced, as discussed in Section 2.4. The loss function used in our model and the criteria for calculating the anomaly score will be described in Sections 2.5 and 2.6, respectively. The discriminator which is the same as that in SAGAN is used in our model to distinguish the label of the extracted latent representation of the input image.

2.2 U-Shape Generator

To simulate the convolution operation, inspired by [13] and [4], we use a reshape operation to change the dimension of the feature vector from the transformer. Patch merging and expanding are applied to change the scales of the vector obtained from the patch embedding of the input image. The operation retains the same data as the input features, but with a new specified shape to achieve the scale transformations.

In our model, the feature vector plays the same role as feature map in the convolution network. The feature vector has three dimensions $B \times L \times C$, where B is the batch size, L denotes the number of patches in this vector, and C stands for the dimension of the features in each patch. In other words, $L = W_{patch} \times H_{patch}$. After patch merging, the number of patches is decreased to L', where $L' = \frac{W_{patch}}{2} \times \frac{H_{patch}}{2}$. After the merging, the dimension of the feature vector is increased to $C' = 4C$. Then, a linear layer is applied to project the vector to the dimension of $2C$. The workflow and data format can be seen in Fig. 2.

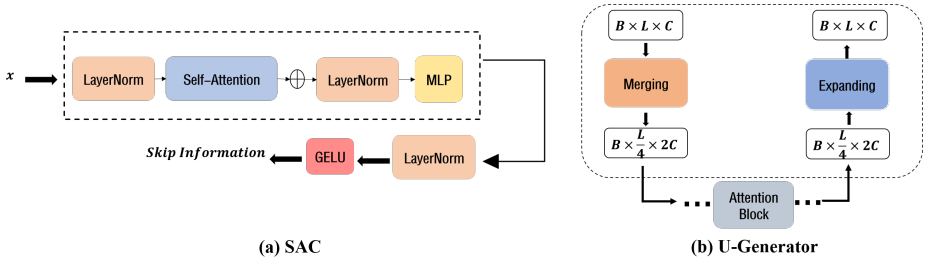


Fig. 2: (a) Transformer-based module in Skip Attention Connection (SAC) which provides multi-scale information through self-attention. (b) The workflow and data format in the U-shape Generator.

2.3 Swin Transformer Block

Transformer captures long-range dependency within image data while increasing the number of tokens at the same time. When the images are represented in high-resolution, the tokens may lead to high computational complexity. To reduce the complexity, we introduce Swin Transformer [13] by replacing the conventional multi-head self-attention module with shifted windows, and calculating the self-attention within the local windows. The Swin Transformer blocks in our model are computed as:

$$\hat{x}^l = W\text{-}MSA(LN(x^{l-1})) + x^{l-1} \tag{1}$$

$$x^l = MLP(LN(\hat{x}^l)) + \hat{x}^l \tag{2}$$

$$\hat{x}^{l+1} = SW\text{-}MSA(LN(x^l)) + x^l \tag{3}$$

$$x^{l+1} = MLP(LN(\hat{x}^{l+1})) + \hat{x}^{l+1} \tag{4}$$

where $\hat{x}^l$ and x^l represent the outputs of the Window Multihead Self-Attention (i.e., W-MSA) and the MLP of the l-th block, respectively. $\hat{x}^{l+1}$ is output of the Sift Window Multihead Self-Attention (i.e., SW-MSA).

2.4 Skip Attention Connection

The Transformer-based structure offers promising results, as demonstrated in Section 3. However, we empirically found (in Table 5) that the Transformer-based U-Generator is not effective in capturing some critical local information in feature representation. Inspired by SAGAN [12] where the CBAM module is incorporated into the skip connection to capture local information, we propose a Skip Attention Connection (SAC) to further improve the performance of our method.

CBAM is a mixed attention mechanism involving convolution operation which can be limited in the receptive field. The work in [5] illustrates the benefit of using positional encoding in a single multi-head self-attention layer. This inspired us to employ the SAC as the module in skip-connection to replace the CBAM module. In the self-attention module, different heads can pay attention to different pixels and areas in the image via the attention mechanism during training. Thus, our improved skip-connection can capture the local information to complement the Transformer-based structure, without being limited by the convolution receptive field as in CBAM.

As shown in Fig. 2, we incorporate self-attention into the output of the encoder in each layer to obtain a feature vector. The feature vector can focus on pixels in different areas in multi-scale by self-attention blocks. The output of each self-attention block will be sent to an MLP followed by a layer normalization. Finally, the vector is passed through a GELU activation function and transferred to the decoder. Therefore, our proposed skip connection block with self-attention can offer multi-scale local information without using the transforms (e.g. reshape operation and convolution) as performed in the CBAM module. Our empirical results in Section 3 show that it performs better than the original CBAM module.

2.5 Loss Function

The aim of our GAN-based anomaly detection method is to train the model on normal data and correctly reconstruct the normal data on both image and latent space. On the contrary, the model should fail to reconstruct the abnormal data as if it is never trained on the abnormal data. Thus we use the loss from [2] in our model as follows.

Adversarial Loss: L_{adv} is the standard loss used in GAN to optimize the generator G and discriminator D in the adversarial process which ensures the generated image from G to be as realistic as possible with the help of classification result from D.

$$\mathcal{L}_{adv} = E_{x\sim p_x}[logD(x)] + E_{x\sim p_x}[log(1 - D(G(x))] \tag{5}$$

where $E_{x\sim p_x}[.]$ indicates expectation.

Contextual Loss: Contextual loss $\mathcal{L}_{con}$ is defined as the error between the generated image $G(x)$ and the input image x, as follows:

$$\mathcal{L}_{con} = E_{x\sim p_x}||x - G(x)||_1 \tag{6}$$

where $||\cdot||_1$ is an L_1 norm. This loss helps the algorithm learn contextual information from images.

Latent Loss: The latent loss aims to reduce the reconstruction loss in latent representation. We choose the feature in the last layer of D to get the latent representation. The latent loss $\mathcal{L}_{lat}$ is formulated as:

$$\mathcal{L}_{lat} = E_{x\sim p_x}||f(x) - f(G(x))||_2 \tag{7}$$

where $||\cdot||_2$ is an L_2 norm, the $f(x)$ and $f(G(x))$ are the latent representation of the input image x and the generated image $G(x)$, respectively. The final loss function is shown as a weighted sum of the loss functions mentioned above.

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{con}\mathcal{L}_{con} + \lambda_{lat}\mathcal{L}_{lat} \tag{8}$$

where λ_{adv}, λ_{con}, and λ_{lat} are the weighting parameters chosen empirically in our experiments.

2.6 Inference

Anomaly score is often used to determine whether a test image is an anomaly or not. Image with a score higher than a predefined threshold is considered as an anomaly. We use the method in SAGAN [12] and Skip-Anomaly [2] to obtain the anomaly score as follows:

$$A(x) = \lambda R(x) + (1 - \lambda)L(x) \tag{9}$$

where x represents the test image, $A(x)$ is the raw anomaly score of x, $R(x)$ is the reconstruction score between x and the generated image x', $L(x)$ is the difference between the latent representations of x and x' which are obtained

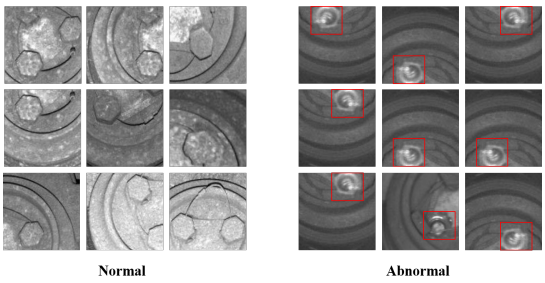
from the discriminator, and λ is the weight that controls the relative importance of $R(x)$ and $L(x)$ in $A(x)$. After calculating the raw anomaly score for all the test images in the test set and denoting them as a vector A, we use the equation below to normalize the scores to the range of $[0, 1]$. Thus, the final anomaly score for an individual test image is obtained as:

$$A'(x) = \frac{A(x) - min(A)}{max(A) - min(A)} \tag{10}$$

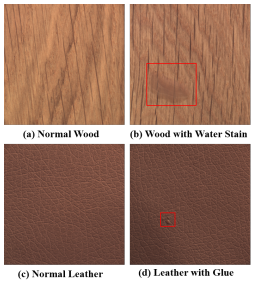
3 Experiment

We evaluate our model[1] in a way of leave-one-class-out anomaly detection, with datasets CIFAR10 [11], STL10 [22], LBOT [12] and MVTecAD [3]. We use SAGAN[2] and Skip-Anomaly as baseline methods in our comparison. The area under the curve (AUC) of the receiver operating characteristic (ROC) is used as the performance metric.

3.1 Dataset



(a) Examples on the left are from normal data, and those on the right are abnormal data with red boxes containing damaged or missing bolts.



(b) Examples in MVTecAD with slight abnormal part which can be detected by SAC

Fig. 3: Examples in the dataset.

CIFAR10: CIFAR10 is a benchmark dataset which consists of color images in 32x32 pixels from 10 classes. We choose one class of images as anomaly and the other images as normal data. Then we train our model on the normal data and test on both normal data and abnormal data.
STL10: STL10 is a dataset similar to CIFAR10. The difference between them is that STL10 has less labeled training data than CIFAR10 in each class. In addition, image resolution in STL10 is 96×96 pixels. We train our model on STL10 in the same way as CIFAR10.
LBOT: The LBOT dataset is used in [12] which focuses on the inspection of axle bolts. The dataset includes 5,000 image patches of the train axle bolt status extracted by the 128×128 overlapping sliding window method. In training, we

[1] https://github.com/SYLan2019/Transformer-Gan-Anomaly-Detection
[2] https://github.com/SYLan2019/Skip-Attention-GAN

define the missing or damaged bolts as anomalies. We split the LBOT dataset into 4,000 training images and 1,000 test images. The training images are all normal bolt images, and the 1,000 test images contain 500 normal bolt images and 500 abnormal bolt images.

MVTecAD: MVTecAD is a benchmark anomaly detection method which focuses on industrial inspection. It contains 5000 images of fifteen different objects and texture categories. The abnormal images in MVTecAD have partial differences from the normal ones in terms of details. We select the images in certain classes as abnormal data and others as normal data.

3.2 Training Details

Our experiments are performed on an NVIDIA GeForce RTX 3090 GPU with 24Gb Memory. In the training process, the objective function is optimized by Adam optimizer with momentum $\beta_1 = 0.5$, $\beta_2 = 0.999$ and initial learning rate $l_r = 2 \times 10^{-4}$. We set $\lambda = 0.1$ in Eq. 9 when calculating the anomaly score, $\lambda_{adv} = 1$, $\lambda_{con} = 50$ and $\lambda_{lat} = 1$ in Eq. 8 when calculating the loss function. We use a patch size of 2×2 in the patch embedding with positional encoding and a four-head self-attention. The window size in the windowed attention is set to 2 with a shift size of 1. Data augmentation is applied to increase the amount of training data.

Due to the instability of GAN, our Transformer-based model may not always converge on the dataset. To alleviate this issue, we introduce a training strategy where in each epoch, we would train the generator once but the discriminator twice. This strategy can help the model to get a relatively strong discriminator first, which helps guide the optimization of the generator in the right direction. The loss function in Fig. 4 shows our model converges faster with this training strategy.

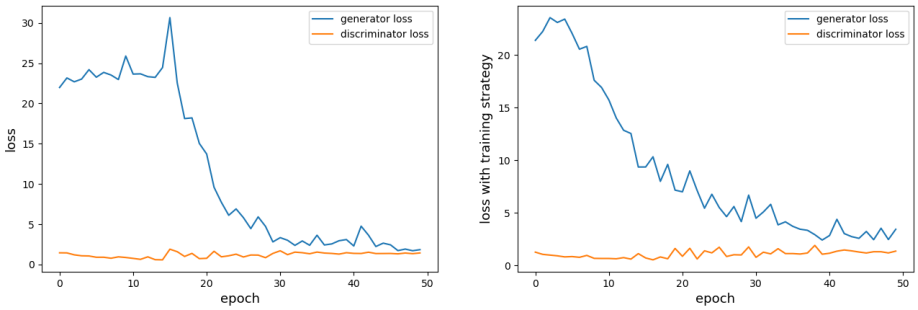


Fig. 4: Convergence of the loss functions. The left plot shows the loss curve on the CIFAR10 with cat as the abnormal class, while the right plot shows that our training strategy can help the model converge faster.

3.3 Encoder of Our Transformer-based Structure

We tested two different self-attention modules of the transformer as our encoder block, respectively, the self-attention module from classical ViT [7] and the window attention of Swin Transformer [13]. The results are shown in Table 1. It can

be seen that the shifted windows in Swin Transformer reduce the complexity on computation and also improve the performance over the ViT, as the Swin Transformer performs the self-attention in parallel. In the following experiments, we use the window attention module of Swin Transformer as the encoder in our proposed method for its computational efficiency.

Table 1: The average AUC with different transformer-based blocks

Module	dataset	Average AUC
self-attention of [7]	CIFAR10	0.963
window attention of [13]	CIFAR10	0.978
self-attention of [7]	STL10	0.946
window attention of [13]	STL10	0.982

Table 2: The AUC results on the LBOT dataset.

Model	AUC
GANomaly [1]	0.900
Skip-GANomaly [2]	0.840
SAGAN [12]	0.960
Proposed with SAC	**0.996**

Table 3: The AUC results on the CIFAR10 dataset.

Model	frog	bird	cat	deer	dog	horse	ship	truck	Average
GANomaly [1]	0.512	0.523	0.466	0.467	0.502	0.387	0.534	0.579	0.496
Skip-GANomaly [2]	0.955	0.611	0.670	0.845	0.706	0.666	0.909	0.857	0.777
SAGAN [12]	0.996	**0.957**	0.951	0.998	**0.975**	0.891	0.990	0.980	0.967
Proposed with CBAM module	**1.000**	0.932	**0.977**	0.998	0.940	0.941	**1.000**	0.969	0.970
Proposed with SAC	**1.000**	0.944	0.960	**0.999**	0.968	**0.949**	0.999	**0.990**	**0.976**

3.4 Experimental Analysis

We compare our model with Skip-GAnomaly [2] and SAGAN on CIFAR10, STL10, LBOT and MVTecAD datasets, using the AUC metric. In addition, we evaluate our model with different skip-connection modules.

Table 3 shows our results on CIFAR10. Our model with modified CBAM performs better than all the CNN-based models in average score. With SAC, the proposed method performs even better. Table 4 shows the results on STL10 in which we can see that SAGAN performs better than the CBAM module on 96×96 images. However, the model with SAC offers the best performance which shows that our proposed SAC can adapt to the resolution change better than the CBAM module.

Table 4: The AUC results on STL10.

Model	bird	car	cat	deer	dog	horse	monkey	ship	truck	Average
SAGAN [12]	0.929	**1.000**	0.963	0.996	0.859	0.947	0.979	**0.999**	**0.998**	0.963
Skip-GANomaly [2]	0.588	0.902	0.556	0.664	0.581	0.726	0.590	0.568	0.770	0.661
Proposed with CBAM module	0.916	0.999	0.937	0.991	0.821	0.922	0.938	0.993	0.997	0.951
Proposed with SAC	**0.966**	**1.000**	**0.985**	**0.998**	**0.942**	**0.975**	**0.980**	**0.999**	0.997	**0.984**

The experimental results on both CIFAR10 and STL10 show that our proposed method performs better in detecting a certain class as an anomaly. This can be attributed to the long-term dependencies within the image that help obtain a more accurate feature representation, in which local details are associated with global information. In addition, the results on MVTecAD (which is mainly used for local texture anomalies) are also greatly improved as compared to SAGAN, further demonstrating that our Transformer-based method outperforms CNN-based methods.

Table 5: The AUC results on MVTecAD dataset.

Model	bottle	capsule	carpet	grid	leather	nut	pill	screw	brush	transistor	wood	zipper	Average
SAGAN	0.873	0.951	0.966	0.918	0.984	0.922	0.806	0.991	0.824	0.832	0.931	0.742	0.896
Skip-GANomaly	0.882	0.869	0.964	**0.966**	0.955	0.954	0.862	0.976	0.900	0.956	0.930	0.898	0.926
Proposed	1.000	**0.988**	0.959	0.945	0.928	**0.991**	0.994	**0.992**	**0.999**	0.984	0.769	1.000	0.962
Proposed with SAC	**1.000**	0.978	**0.974**	0.938	**0.991**	0.965	**1.000**	0.986	0.981	**0.989**	**0.970**	**1.000**	**0.981**

Table 5 shows the results on the MVTecAD dataset. From this table, we can see that our proposed Transformer-based method gives better performance than the SAGAN due to its effectiveness in capturing the long range dependency in image data. However, the table also shows our proposed method with pure skip connection gives relatively low accuracy in some types (such as wood in Table 5). By inspecting the images in these images, we found that the defects in them are so small that the standard skip connection may miss the subtle details. In contrast, with the skip connection method described in Section 2.4, the results can be significantly improved on MVTecAD. Experiments on MVTecAD and LBOT datasets show that the proposed method outperforms the baseline SAGAN by over 3% in terms of AUC metric (e.g. see details in Table 2 and Table 5).

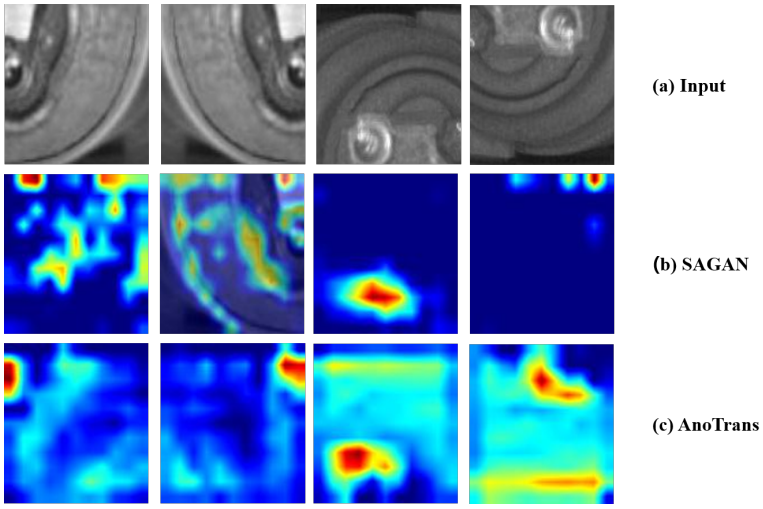


Fig. 5: The heat-map from attention features in our proposed model AnoTrans and SAGAN.

We also train the model with skip attention connection on LBOT, which is a real anomaly dataset. The results are shown in Table 2. We can see that the proposed model with transformer has a stronger representation ability than the convolution model. Fig. 5 visualizes the attention features on LBOT using Grad-Cam [21]. We can observe that our Transformer-based method has a wider horizon on image detection. Compared to the SAGAN which focuses on the local area when detecting anomalies, our method picks the abnormal area at a larger scale through the long-range dependency. The result shows that a wider horizon

with long-range dependency can produce a better representation and locate the anomaly more precisely considering the semantic of the whole image.

4 Conclusion

In this paper, we have presented a Transformer-based method for anomaly detection from images. The experiments show that the proposed method outperforms the CNN-based methods in capturing long-range dependency, and the limited receptive field of CNN can be effectively mitigated by the self-attention mechanism of the transformer. In addition, using the modified skip connection with self-attention in our Transformer-based encoder can further improve the performance, due to the advantage of skip connection in exploiting the multi-scale information. Compared with the state-of-the-art CNN-based anomaly detection methods, our method achieves better results on four datasets evaluated. In the future, we will further study representation enhancement in anomaly detection from images.

References

1. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: Ganomaly: Semi-supervised anomaly detection via adversarial training. In: Asian conference on computer vision. pp. 622–637. Springer (2018)
2. Akçay, S., Atapour-Abarghouei, A., Breckon, T.P.: Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2019)
3. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9592–9600 (2019)
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation (2021)
5. Cordonnier, J.B., Loukas, A., Jaggi, M.: On the relationship between self-attention and convolutional layers. arXiv preprint arXiv:1911.03584 (2019)
6. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE Signal Processing Magazine **35**(1), 53–65 (2018)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Ghazal, M., Vázquez, C., Amer, A.: Real-time automatic detection of vandalism behavior in video sequences. In: 2007 IEEE International Conference on Systems, Man and Cybernetics. pp. 1056–1060. IEEE (2007)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

10. Kwon, D., Natarajan, K., Suh, S.C., Kim, H., Kim, J.: An empirical study on network anomaly detection using convolutional neural networks. In: 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). pp. 1595–1598. IEEE (2018)
11. Li, H., Liu, H., Ji, X., Li, G., Shi, L.: Cifar10-dvs: an event-stream dataset for object classification. Frontiers in neuroscience **11**, 309 (2017)
12. Liu, G., Lan, S., Zhang, T., Huang, W., Wang, W.: Sagan: Skip-attention gan for anomaly detection. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 2468–2472. IEEE (2021)
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
14. Mould, N., Regens, J.L., Jensen III, C.J., Edger, D.N.: Video surveillance and counterterrorism: the application of suspicious activity recognition in visual surveillance systems to counterterrorism. Journal of Policing, Intelligence and Counter Terrorism **9**(2), 151–175 (2014)
15. Pang, G., Shen, C., Cao, L., Hengel, A.V.D.: Deep learning for anomaly detection: A review. ACM Computing Surveys (CSUR) **54**(2), 1–38 (2021)
16. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
17. Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., Klette, R.: Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding **172**, 88–97 (2018)
18. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis **54**, 30–44 (2019)
19. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International conference on information processing in medical imaging. pp. 146–157. Springer (2017)
20. Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C., et al.: Support vector method for novelty detection. In: NIPS. vol. 12, pp. 582–588. Citeseer (1999)
21. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
22. Singh, H., Swagatika, S., Venkat, R.S., Saxena, S.: Justification of stl-10 dataset using a competent cnn model trained on cifar-10. In: 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA). pp. 1254–1257. IEEE (2019)
23. Wang, Z., Cun, X., Bao, J., Liu, J.: Uformer: A general u-shaped transformer for image restoration. arXiv preprint arXiv:2106.03106 (2021)
24. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
25. Zenati, H., Foo, C.S., Lecouat, B., Manek, G., Chandrasekhar, V.R.: Efficient gan-based anomaly detection. arXiv preprint arXiv:1802.06222 (2018)