Expert Systems With Applications 228 (2023) 120284

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Anomaly-GAN: A data augmentation method for train surface anomaly detection

Ruikang Liu [a], Weiming Liu [a,*], Zhongxing Zheng [a], Liang Wang [b], Liang Mao [b], Qisheng Qiu [b], Guangzheng Ling [b]

[a] South China University of Technology, Guangzhou 510641, China
[b] Guangdong Pearl River Delta Intercity Railway Co., Ltd, Guangzhou 510000, China

ARTICLE INFO

Keywords:
Generative adversarial networks
Data augmentation
Image-to-image translation
Train surface anomaly detection

ABSTRACT

Train surface anomaly detection is an essential task in vision-based railway safety inspection. Although existing deep learning methods show great potential, their anomaly detection accuracy is affected by the lack of abnormal images. The number of abnormal images on the train surface is far less than that of normal images. One effective way to solve this problem is to expand the abnormal sample. However, most of train surface anomaly image generation methods faces difficulties in producing images with high quality and rich diversity at the same time. In addition, generating small-area anomalies is not ideal. An Anomaly-GAN based on mask pool, abnormal aware loss, and local versus global discriminators is thus proposed in this paper to solve these problems. The mask pool consists of prior-knowledge-based masks and expert-experience-based masks that guide model in generating anomalies with different shapes, rotation angles, spatial locations, and part numbers. The anomaly aware loss focuses on small-area anomalies, thereby promoting the generated anomalies with more detailed textures and richer semantics. The local versus global consistency discriminators combine local and global feature expressions that lead to the generation of more realistic and natural abnormal samples. The experiments show that, compared with other advanced data augmentation algorithms, the images generated by Anomaly-GAN achieve the best FID and LPIPS scores in all anomaly categories. In addition, compared with the case without data augmentaion, the proposed data enhancement method improves the performance of CNN on mAP and mIOU by 25.6% and 24.2%, respectively. Test code is available in https://github.com/AI-Dream-Chaser/Anomaly-GAN.

1. Introduction

In train surface anomaly inspection, using intelligent technology can effectively reduce labor costs and improve operational efficiency. Under the high-speed operation of the electrical multiple units (EMUs), the foreign objects, such as wool fabric and paper scraps, near the track are easily drawn into the bottom bogie, cable, and equipment gaps, thereby generating smoke and foul odor, and may even trigger equipment short circuits and fires. In addition, the long-term operation of the train can loosen the screws of some key components on the train surface and subsequently reduce the safe operation of the train. Recently, many railway system started installing the trouble of moving EMU detection system (TEDS) along railways to capture images of critical parts of moving trains from the bottom and two-side views. All images captured from TEDS need to be transmitted to the control center in real time and then inspected by experienced visual inspectors. However, human factors, such as technician stress and eyestrain, can trigger lead to numerous false alarms and missed inspection.

Recently, deep learning (DL)-based methods have shown great advantages in traditional visual inspection fields, including defect classification (Zhang et al., 2022a) and detection (Dai et al., 2022; Zheng & Cui, 2022). Some DL algorithms have also been applied to detect anomalies on the train surface key components. However, these DL-based methods usually require a large amount of training samples, which are difficult to acquire. Chen et al. (2020a) proposed a hybrid DL-based framework for detecting defects in the component of moving trains. However, this framework only acquires a limited number of defect images and was only trained on 1156 with strong data imbalance. Chen et al. (2020b) copied and pasted the defective components of trains multiple times in each image to augment their training data. They also attempted to address the poor accuracy and robustness

* Corresponding author.
E-mail addresses: liuruikanglin@163.com (R. Liu), wmliu@scut.edu.cn (W. Liu), zxzheng318@foxmail.com (Z. Zheng), 253572784@qq.com (L. Wang), 390677143@qq.com (L. Mao), qiuqisheng@gzmtr.com (Q. Qiu), lingguangzheng@126.com (G. Ling).

https://doi.org/10.1016/j.eswa.2023.120284
Received 25 December 2022; Received in revised form 14 March 2023; Accepted 24 April 2023
Available online 8 May 2023

of their DL model by increasing the number of images and sample diversity. Train component defect detection (Zhang et al., 2022c), railway insulator surface defect detection (Kang et al., 2018), and EMU key components defect detection (Zhao et al., 2020) also adopt simple geometric transformation methods for data augmentation, such as scale change, rotation, translation, and flipping. However, the lack of anomaly data is inevitable in anomaly (or defect) detection tasks, hence underscoring the importance of using data augmentation to improve the anomaly detection performance.

The difficulties in acquiring anomaly data of moving trains can be ascribed to two reasons. First, the improvements in manufacturing and logistics maintenance technologies have extended the service life of train components, and reduced the frequency of abnormal phenomena. Therefore, anomalies that cover all possible features, including defective component and foreign object samples, are rarely trained by DL detection models. Second, the collection and labeling of anomalous samples involve a high cost.

The lack of anomaly samples (Yang et al., 2021) usually introduces two problems, namely, data imbalance and weak model generalization ability. On the one hand, if the number of training samples of different types of anomalies is not uniform, then a DL model cannot easily learn the representation of some rare anomalies, thereby reducing its detection accuracy. On the other hand, the scarcity of abnormal features can prevent the model from learning diverse features at the training stage, thereby resulting in overfitting. Although the trained model performs well on the training set, when the abnormal features of the test and training set are inconsistent, the model performs poorly on the test set. To alleviate this problem, many researchers have successfully applied data augmentation methods in various fields, including rail fastener anomaly detection (Su et al., 2022), medical image anomaly detection (Zhang et al., 2022b), and industrial parts surface defect detection (Niu et al., 2021).

Image augmentation is a popular method in image processing whose fundamental purpose is to include more potentially never-seen-before images during training and enhance the ability of DL models to learn diverse features. The traditional geometric-transformation-based augmentation methods mainly include scale change, rotation, random cropping, translation, flipping, and copy pasting. However, the images produced by simple linear transformations do not reveal the new distributions and characteristics of unknown anomalies, such as random changes in shape or texture. Given that DL models cannot fundamentally learn different abnormal features, they cannot fundamentally improve their detection accuracy.

With the outstanding performance recently demonstrated by image generation models (such as autoencoders [AE] (Kingma & Welling, 2013) and generative adversarial networks [GAN] (Goodfellow et al., 2020)), these models have been utilized in different industries to help generate images that are otherwise difficult to obtain to supplement their real data. The anomaly detection performance can be improved by augmenting the training set. However, previous research on image-generation-based anomaly data augmentation still suffers from several shortcomings. First, the existing generation methods typically focus on the entire image rather than local regions and can easily ignore anomalies, especially the smaller ones. Our application scenario is particularly concerned with the task of small anomalies generation. Second, the generated images are random and uncontrolled and have shape and location features that are similar to those of existing anomalous images. Therefore, the diversity of a training set augmented in this way can hardly be improved. Third, the image quality needs to be improved. The GAN training process is prone to mode collapse, which blurs the details of the generated small anomalies. In addition, if the generated image is unreal and unstable, then the connection between the generated local anomalies and the global image is obviously unnatural. In sum, fine-tuning anomaly detection models using unclear and unreal synthetic images with poor diversity can be detrimental.

To address these problems, this paper proposes an Anomaly-GAN for train surface anomaly image generation. The main contributions of this work are summarized as follows:

1. An Anomaly-GAN based on mask pool, abnormal aware loss and local versus global discriminators is proposed to solve the shortage of anomaly image and the poor diversity of abnormal features.
2. Anomaly-GAN improves the quality of the generated anomaly images and generates anomalies with different shapes, rotation angles, spatial locations, and numbers in a controllable manner. This method also generates high realistic synthetic anomalies and small anomalies with shaper details.
3. A train surface anomaly dataset is collected from the real scene. Extensive experiments show that Anomaly-GAN improves the performance of the existing DL models in train surface anomaly detection tasks.

The rest of this paper is organized as follows. Section 2 briefly reviews the previous works on train surface anomaly detection and surface anomaly generation. Section 3 introduces the pipeline of the proposed method. Section 4 discusses Anomaly-GAN in detail. Section 5 presents the experiment results. Section 6 concludes the paper.

2. Related work

This section reviews the related work on surface anomaly detection and anomaly image generation based on DL models, especially in train surface anomaly augmentation scenarios, and discusses similar applications in industrial surface anomaly image augmentation.

2.1. Train surface anomaly detection

DL has made remarkable achievements in computer vision and resulted in the generation of a number of advanced algorithms, such as Faster R-CNN (He et al., 2016), YOLO (Redmon & Farhadi, 2018), and Mask R-CNN (He et al., 2017). These DL methods also achieve significant advantages in the challenging task of detecting surface anomalies, such as on train surface anomalies and railway fastener. He et al. (2019) designed a variant of the Faster R-CNN network to detect the plastic bags drawn into the bottom of high-speed trains. However, their model can only detect one class of anomalies because cases of foreign objects found under these trains are very rare. To obtain more real defect samples for model testing, Zhang et al. (2021a) manufactured artificial defects on key components of experimental trains in an EMU operation. However, the man-made destruction of components is costly, and the number of abnormal samples obtained is still very limited (only 200). Chen et al. (2020b) proposed a multistate fault diagnosis and location method for detecting anomalies in the key components of high-speed trains. They copied and pasted small objects multiple times in each anomaly image and performed rotary transformation to enhance the number and diverse scale of small anomalies. Dong et al. (2019) also used the copy-pasted data augmentation strategy (Kisantal et al., 2019) to improve the detection accuracy of Faster R-CNN for defective railway fasteners. However, these simple data augmentation methods for the training dataset prevent DL models from learning new and diverse anomalous features.

Although DL possesses great advantages in surface anomaly detection, this method requires a large number of diverse, well-annotated surface anomaly images for model training, which are rare in the train surface anomaly detection scenario. Therefore, a surface anomaly generation approach must be proposed to augment the surface anomaly images for DL training.

2.2. Image generation based data augmentation

Since their introduction by Goodfellow, GANs (Goodfellow et al., 2020) and their variants have rapidly become mainstream image generation algorithms due to their superior generative capabilities. The basic GAN consists of a generator and a discriminator, both of which

are involved in a game process for Nash equilibrium. The goal of the generator (G) is to "fool" the discriminator (D) by generating real images as much as possible, whereas D does its best to distinguish the generated images from the real ones. DCGAN (Heusel et al., 2017a), LSGAN (Mao et al., 2017), InfoGAN (Chen et al., 2016), and CGAN (Mirza & Osindero, 2014) are typical representatives of early image generation models that generate an image from randomly distributed noise, such as Gaussian noise. For example, DCGAN (Heusel et al., 2017a) replaces the fully connected layers with the convolutional architecture to ensure the stability and convergence of the network during training. CGAN (Mirza & Osindero, 2014) controls the attributes of generated images by providing additional category information, thereby improving the image generation performance. However, these methods are often plagued by the notorious mode collapse problem, where G tries to generate some fixed patterns to fool D, hence leading to the generation of low-quality images with repetition anomalies. In addition, the generation process has limited controllability, while the synthetic anomalies are not realistic and cannot be used to fine-tune the anomaly detection model.

More powerful variants of GANs have recently been used for the augmentation of industrial anomaly images. Among these variants, image-to-image approaches are currently the mainstream. According to the type of input, these methods can be categorized into supervised and unsupervised models. Pix2Pix (Isola et al., 2017) and pix2pixHD (Wang et al., 2018) are classic representatives of supervised algorithms. The pix2pix-based approach applies GANs in supervised image-to-image translation tasks, such as labels to street scenes, aerial to map, and day to night. These methods essentially attempt to optimize the mapping from pixel to pixel through the L1 loss, hence requiring them to pair the input images. Despite improving the quality of the generated images, the Pix2pix series cannot easily collect paired samples for training in anomaly image generation tasks. Furthermore, these methods can effectively generate sharp details of small areas but fail to generate realistic large area textures. Unsupervised algorithms, such as CycleGAN (Zhu et al., 2017), can successfully achieve image-to-image translation without the need for paired images and can freely convert normal pictures into abnormal ones. Therefore, unsupervised models have more applications in surface anomaly generation.

On the basis of these architectures, Yu and Liu (2021) proposed a multi-granularity GAN with auxiliary feature extractor for defective wafer map generation. However, this method is only applicable to large-scale anomalies occupying the whole image. In Liu et al. (2019), Liu et al. proposed a typical supervised method for surface anomaly generation that must maintain high consistency in the anomalous attributes (e.g., shape, location, and scale) of the paired images during the training process. On the basis of the unsupervised model, Niu et al. (2020) proposed a SDGAN network to expand the commutator cylindrical surface defect image dataset. Similarly, Liu et al. (2021) proposed FD-Cycle-GAN to generate defective railway fastener images using a large number of defect-free ones. However, while these CycleGAN-based methods can efficiently synthesize repetitive textures in a large area, they are generally unable to generate some details of small areas. In addition, they have no control over the shape, scale, angle, and number of anomalies during image generation. To obtain sharper details of small areas, a Mask2Defect network (Yang et al., 2021) was used as an image augmentation method for improving metal surface defect inspection. This network initially renders various defect details according to the mask by an AE and then applies CycleGAN to transform the rendered samples from the fake domain to the real defect domain. Despite their improved image generation ability, using multiple complex networks also consumes much time. Overall, most GAN-based models have achieved promising results in surface anomaly generation, but some remaining problems need to be addressed. In the generation of train surface anomaly samples, anomalies usually occupy a small part of the image. Therefore, enhancing the details of small-area anomalies while synthesizing realistic large-area textures remains a challenge. A large number of normal images should also be used to controllably generate anomalies and increase the diversity of the generated images.

3. Overview of the proposed method

Fig. 1 presents an overview of the proposed approach for train surface anomaly samples generation. This method consists of two parts, namely, anomaly image generation module and effect verification module.

The proposed unsupervised Anomaly-GAN lies at the core of the anomaly image generation module. The input of Anomaly-GAN includes unpaired images and a mask pool. In the original dataset of unpaired images, the normal dataset contains a large number of normal samples obtained from the key components of the train surface, whereas the anomaly dataset only contains a small number of abnormal images. Anomaly-GAN can be converted between normal and abnormal images. To generate anomalies with varied features, a mask pool is designed to integrate expert empirical knowledge into the generation process. DL based detection models differ greatly from human experts in feature learning. Specifically, DL based models often fail to identify abnormal features for which they are not trained, whereas human experts have a rich imagination to anticipate future anomalies. For example, when they see a short cloth strip, they can recognize other strips of different shapes and lengths. Therefore, we can reasonably use expert experience to generate additional new anomalies. To achieve such, we design a mask pool to encode the expert experience knowledge and the real anomalous features-based prior knowledge into binary masks, which are then used as guidance vectors to generate new anomalies with different shapes, rotation angles, spatial locations, and part numbers.

Anomaly-GAN attempts to generate high-quality small-area anomalies and large-area background textures simultaneously. Anomaly-GAN contains two generators, four discriminators, and two abnormal aware networks. This method follows the symmetrical cyclic processes of "Normal → Abnormal → Normal" and "Abnormal → Normal → Abnormal". The cycle-consistent generation process ensures the transition from normal to abnormal images. The proposed abnormal aware networks and local versus global discriminators guarantee that the generated anomalies not only have sharp local details but also have realistic background textures.

The anomaly images generated by Anomaly-GAN along with the partial real anomalies will be used as teacher samples to finetune the DL-based anomaly detection model, thereby compensating for the various features that human experts want the detection model to learn. As a result, the detection model can adaptively learn a wider range of anomaly features and achieve a better performance in identifying rare anomalies. The effect verification module is then used to verify the performance improvement of the data augmentation method over the DL method.

4. The proposed method

We describe the proposed data augmentation method in this section in three aspects, namely, unpaired images and mask pool, structure of Anomaly-GAN, and objective functions.

4.1. Unpaired images and mask pool

4.1.1. Unpaired images collected by TEDS system

Fig. 2(a) shows the actual installation of our image acquisition equipment. The TEDS system uses high-speed area scan charge coupled device (CCD) cameras and linear scan CCD cameras installed on the rail side and under the gantry to obtain the key components of moving trains from the bottom view and two-side view. This methodology can realize the troubleshooting of different trains with high efficiency, low cost, and without affecting the safety of the train operation. Fig. 2(b) shows some of the collected normal and abnormal images, including those of hanging cloth strips, missing screws, and attaching paper scraps. In actual train operations, the train surface

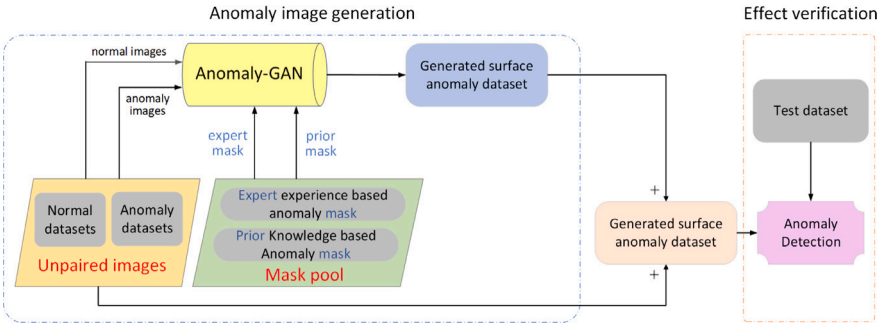


Fig. 1. Overview of the proposed data augmentation method, which has two main components, namely, the anomaly image generation module and the effect verification module.

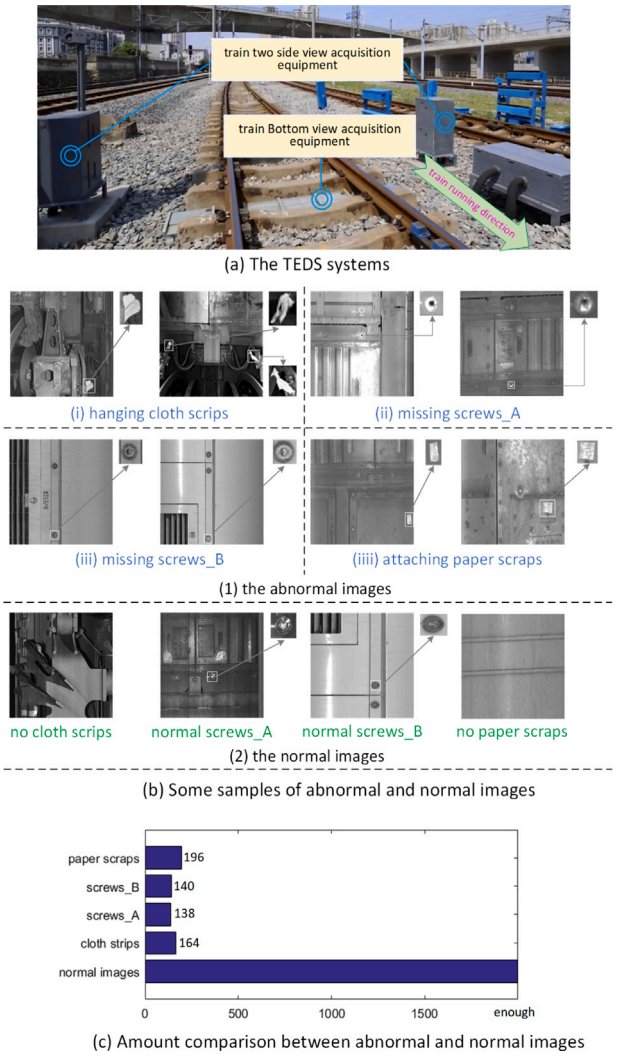


Fig. 2. Data collection system. (a) TEDS platform; (b) some image samples; and (c) comparison of the number of normal and abnormal images.

components are rarely abnormal. We collect our negative samples in real scenarios, thereby explaining the small number of negative samples in this study. Fig. 2(c) compares the volume of normal and abnormal images collected over the past 3 years.

The supervised data augmentation methods based on paired images and pixel-level loss (such as L1 and L2 loss) generally perform better at generating shaper small-area anomalies. However, these methods require the paired images in the training set to be completely consistent in content and style for the algorithm to achieve pixel-to-pixel mapping. When we want to generate anomalous style images, the anomalies must be relatively rare. In addition, the normal and abnormal images we collected are difficult to register with each other in terms of content and style. Therefore, using supervised algorithms is not ideal. We thus use unpaired images in our proposed Anomaly-GAN to realize the transition from the normal to abnormal image domains. Our proposed data augmentation algorithm generates various local abnormal components while keeping the background content unchanged.

4.1.2. Proposed mask pool

Identifying new and previously unseen anomalies is a challenge to DL-based anomaly detection models. Merely recognizing the labeled anomalies is far from sufficient for real-world needs. We propose using a mask pool to increase the diversity of the anomaly samples. By introducing expert empirical knowledge to create a large number of "previously unseen" anomalies as augmented data, our proposed model can learn more diverse features. As a result, the DL-based detection model can be improved in terms of its generalization ability and recognition of unseen anomalies. The elements in the mask pool include some pictorial, feature-rich binary masks that represent the compensation information that human experts want the model to learn. Therefore, we use these binary masks as guide vectors for subsequent image generation models to synthesize diverse anomalous images.

Fig. 3 illustrates the creation process of mask pool. The abnormal elements in the mask pool have two main sources, namely, the prior-knowledge-based anomaly masks and expert-experience-based anomaly masks. The prior-knowledge-based anomaly masks are directly cut out from the original data in the collected anomaly image dataset, whereas the expert-experience-based anomaly masks are drawn by the experts through labelme (Russell et al., 2008) tools and vary in shape, scale, and spatial location.

To guarantee the rationality of the expert hand-drawn masks, we add some weak constraints to the drawing process according to the actual scenarios. For morphologically fixed abnormalities (e.g., screws), the expert simply cut out the anomaly-free screws in the collected normal images without changing their shape. The position and number of anomalous elements in the binary anomaly masks can vary randomly according to the situation in the normal images. For anomalies with complex textures (e.g., cloth strips and paper scraps), experts can initially create new irregular shapes. After these masks are created, an expert evaluation system is applied according to fuzzy set theory (Isomoto & Yoshine, 1995) to measure the rigor and rationality of hand-drawn anomalies. A total of 10 experts score the rationality of the hand drawings, and the final score of each hand drawing is taken as the average of the expert scores. The top 10% hand-drawn masks unanimously approved by senior experts will be put into the mask pool.

To further increase the number and diversity of new samples, the anomalous elements (excluding screws) in these two groups are randomly repositioned, rotated, and scaled. Different anomalous elements can also be randomly combined to increase the number of anomalies in a single mask. This operation reduces the workload of drawing. These binary masks can also be used as the ground-truth of the anomalies for the subsequent anomaly detection baseline.

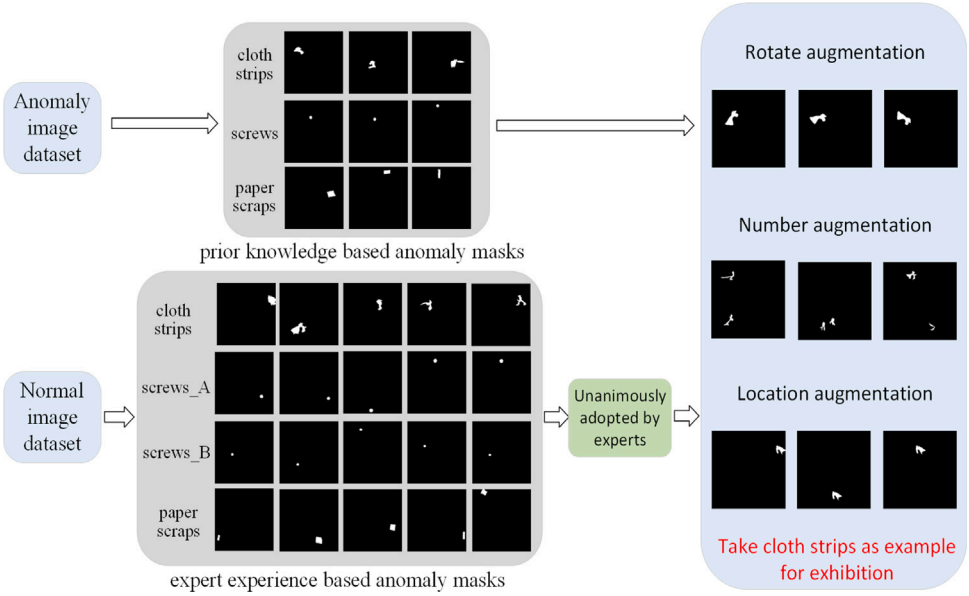


Fig. 3. Mask pool creation process.

4.2. The structure of anomaly-GAN model

Proposed by Goodfellow et al. GAN is a widely used model for image generation (Goodfellow et al., 2020). However, anomalous data from TEDS are extremely scarce, thereby making it difficult to meet the training requirements of the original GAN. Moreover, the generated anomaly images in most existing GAN exhibit ambiguous local details and insufficient image diversity, but any anomaly detection model needs to be trained using anomaly images with high diversity and sharp details. To address these problems, we propose Anomaly-GAN, whose structure is illustrated in Fig. 4. Anomaly-GAN consists of two generators ($G_{norm2abnl}$, $H_{abnl2norm}$) and four local versus global consistency discriminators ($D1_{abnl}$, $D2_{abnl}$, $D1_{norm}$, $D2_{norm}$).

Anomaly-GAN solves an image-to-image problem. To learn the feature map between the normal domain "Norm" and abnormal domain "Abnl", two processes are carried out simultaneously, namely, "Normal → Abnormal → Normal" (Fig. 4(a)) and "Abnormal → Normal → Abnormal" (Fig. 4(b)). The input normal images I_{norm} and abnormal images I_{abnl} are unpaired in content. The two cycles add more constraints to the image transformation process, such as generating local anomalies while ensuring the consistency of the background textures between the normal and generated anomalous images. If constrained conditions do not exist, then the mapping from the "Norm" to "Abnl" domains can freely change the background content. The motivation behind these two cycles is that, if we transform a normal image into an abnormal image and then convert this image back to normal, then we should go back to the original normal input. Therefore, local anomalies can be naturally generated in normal images while keeping the texture background unchanged.

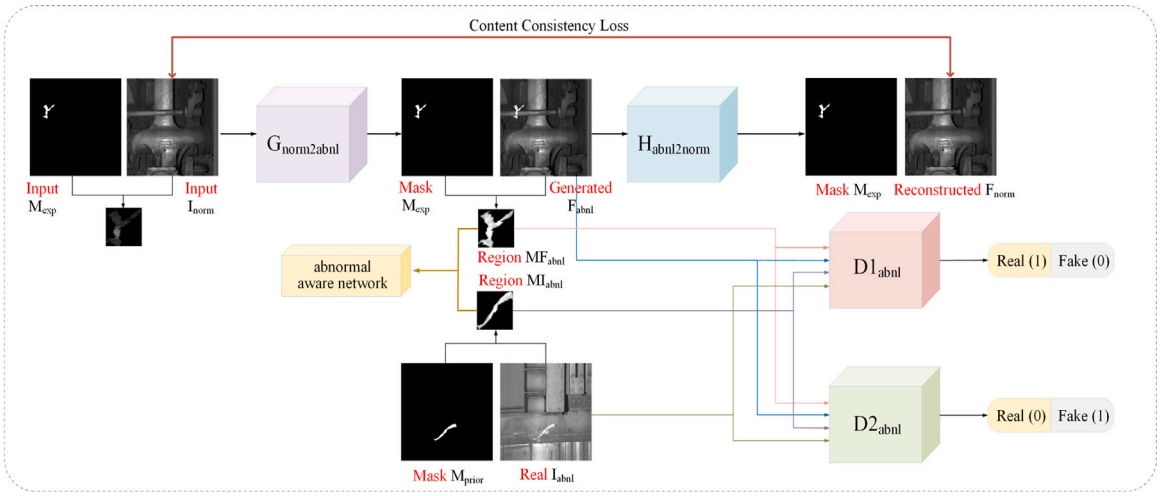
As shown in Fig. 4, all generator and discriminator networks are trained under two unpaired training sets, namely, one set of images from the "Norm" domain and another set of images from the "Abnl" domain. The images from the "Norm" and "Abnl" domains are defined as $\{I_{norm}, I_{norm} \in \text{"Norm"}\}$ and $\{I_{abnl}, I_{abnl} \in \text{"Abnl"}\}$, respectively. Generator $G_{norm2abnl}$ transforms the real normal images into abnormal images, whereas generator $H_{abnl2norm}$ converts the abnormal images into normal images. The discriminators $D1_{abnl}$ and $D2_{abnl}$ aim to classify the real anomaly image $\{Real\ \ I_{abnl}\}$ and generated anomaly image $\{Generated\ \ F_{abnl}\}$. Meanwhile, the discriminators $D1_{norm}$ and $D2_{norm}$ aim to distinguish the real normal image $\{Real\ I_{norm}\}$ from the generated normal image $\{Generated\ F_{norm}\}$. $D1_{abnl}$ and $D1_{norm}$ give high score on real images and low score on generated images. By contrast, $D2_{abnl}$ and $D2_{norm}$ assign high scores to the generated images and low scores to the real images. The network structures of the generator and discriminator are shown in Fig. 5 and Fig. 7, respectively. Generators $G_{norm2abnl}$ and $H_{abnl2norm}$ as well as discriminators $D1_{abnl}, D2_{abnl}, D1_{norm}, D2_{norm}$ have the same network structures. The workflows of the generator and discriminator are described in detail below.

4.2.1. Structure of generators

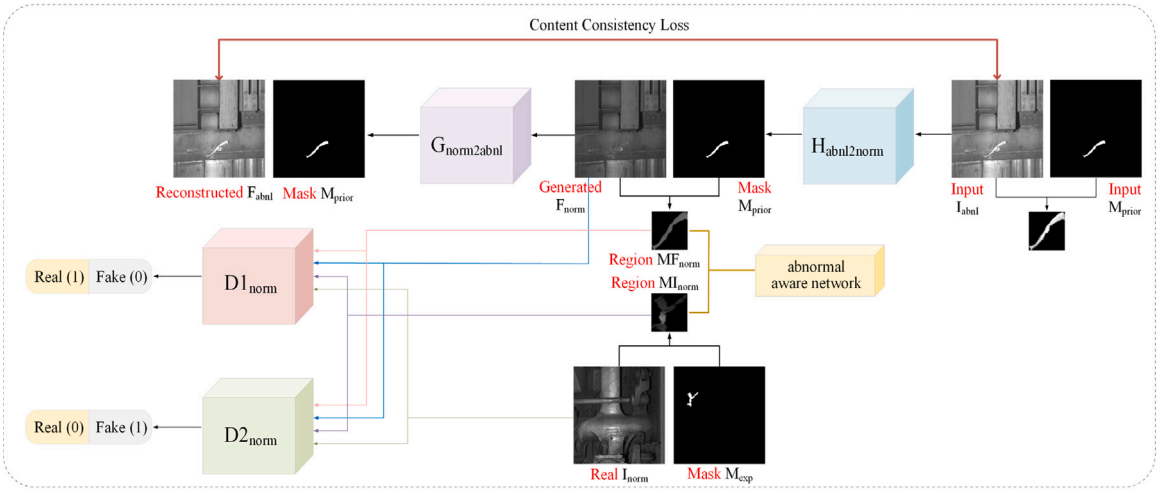
A U-Net (Ronneberger et al., 2015)-based encoder–decoder network is used as the backbone of the generator to synthesize images. The structure of the generator shown in Fig. 5 uses the concatenated 4-channel tensor as input and outputs the 3-channel fake image. The 4-channel tensor ($4 \times 512 \times 512$) is concatenated by the image ($3 \times 512 \times 512$) from the "Norm" or "Abnl" domain and the binary anomaly mask ($1 \times 512 \times 512$) from the mask pool. The expert-experience-based anomaly mask $\{Mask\ M_{exp}\}$ and prior-knowledge-based anomaly mask $\{Mask\ M_{prior}\}$ are available in the two-cycle process. The role of the anomaly mask is to guide the generation of anomalies with various features, as stated in Section 4.1.2. Skip connections are used for the feature transfer among different layers of the generator, which ensures an accurate reconstruction of background textures that would otherwise be difficult to reconstruct the image details.

We expect the generator to generate diverse anomalies in normal images based on the properties of the binary anomaly mask. However, existing image-to-image methods generally focus on the texture conversion of the entire image and hardly pay attention to small-area anomalous features. Furthermore, the location, size, shape, and other attributes of the anomalies in M_{exp} and M_{prior} are unpaired, thereby introducing huge challenges in local anomaly generation. To solve these problems, we design an abnormal aware network at the image generation stage, which makes the Anomaly-GAN perform well in small-area anomalous semantic feature generation. The configuration of the abnormal aware network is shown in Fig. 6.

We take the abnormal aware network in the "$Normal \rightarrow Abnormal \rightarrow Normal$" process as an example. The small-area anomalies in the generated anomaly image $\{Generated\ F_{abnl}\}$ and real anomaly image $\{Real\ \ I_{abnl}\}$ are segmented according to the binary masks M_{exp} and M_{prior}, respectively. Afterward, the cropped regions MF_{abnl} and MI_{abnl}



(a) The "Normal-Abnormal-Normal" Process



(b) The "Abnormal-Normal-Abnormal" Process

Fig. 4. Structure of Anomaly-GAN. (a) "*Normal* → *Abnormal* → *Normal*" process; and (b) "*Abnormal* → *Normal* → *Abnormal*" process.

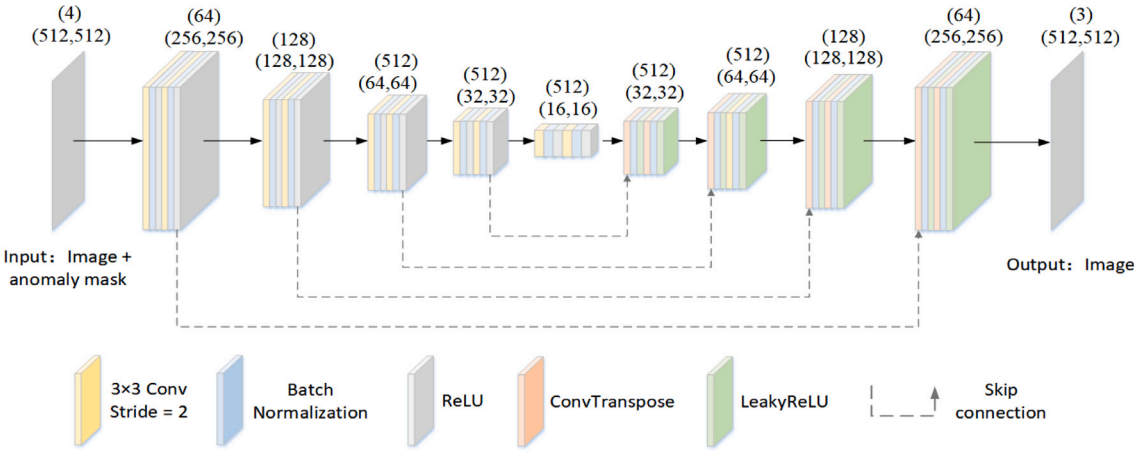


Fig. 5. Structure of generators $G_{abnl2norm}$ and $H_{norm2abnl}$.

are fed into the abnormal aware network for semantic feature extraction. The purpose of the abnormal aware network is to make the generated local anomalies approximate the real anomalies in terms of semantic features. Specifically, we initially perform a dot-multiplication operation $(F_{abnl} \odot M_{exp}, I_{abnl} \odot M_{prior})$ on the image and mask to obtain the foreground objects (anomalies). $\odot$ represents the dot-product operator. Afterward, by calculating the minimum bounding box of the anomaly region in the binary mask, we obtain the upper-left and lower-right coordinates of the anomaly region. The small-area anomalies MF_{abnl} and MI_{abnl} are then cropped from $\{F_{abnl} \odot M_{exp}\}$ and $\{I_{abnl} \odot$

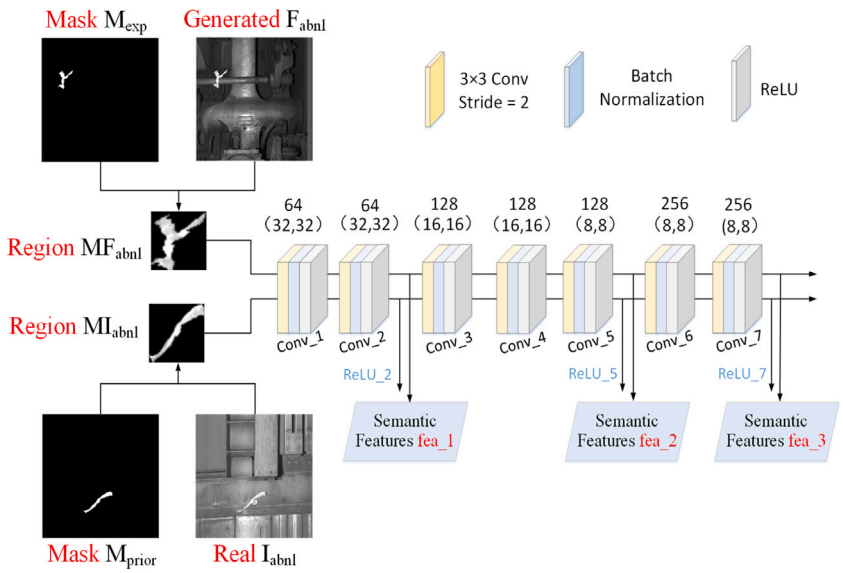


Fig. 6. Configuration of the abnormal aware network.

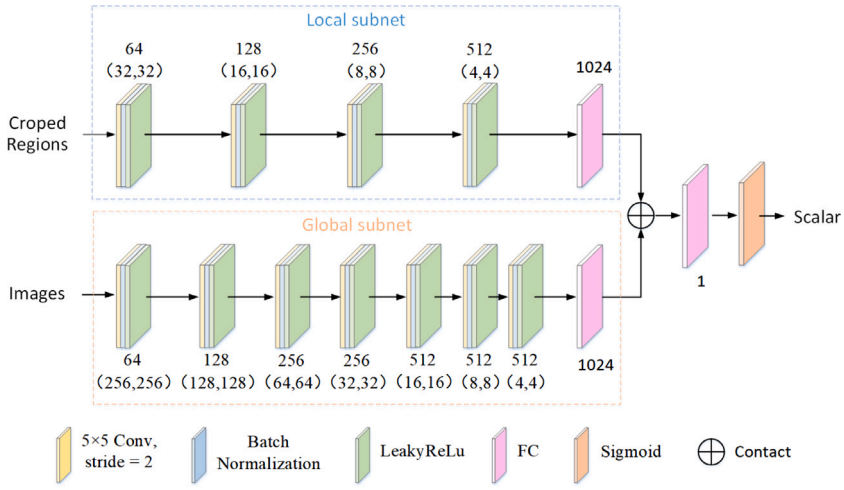


Fig. 7. Configuration of the local versus global consistency discriminator.

$M_{prior}\}$ according to these coordinates. These cropped regions only contain the foreground, hence preventing the unnecessary influence of background texture during the generation of local anomalies.

All cropped regions inputted into the abnormal aware network are resized to 64 × 64 pixels. As shown in Fig. 6, the model consists of seven convolution layers. Two max pooling layers are inserted after the Conv_2 and Conv_4 layers to obtain high-level semantic information in the lower dimension. Moreover, each convolutional layer is followed by a $ReLU$ layer. The output of $ReLU\_2$, $ReLU\_5$ and $ReLU\_7$ are adopted as the final extracted semantic features and used to measure the semantic differences in the perceptual texture. The semantic difference $L_{semantic}$ written as follows:

$$L_{semantic}(MF_{abnl}, MI_{abnl}) = \sum_{i=1}^{3} \frac{fea\_i(MF_{abnl}) - fea\_i(MI_{abnl})}{D_i \cdot H_i \cdot W_i}, \quad (1)$$

where $\{fea\_i, i \in (1,3)\}$ represents the semantic features in the different layers of abnormal aware network, and D_i, H_i and W_i represent the depth, height, and width of the semantic features outputted by different layers, respectively.

The semantic difference $L_{semantic}$ is also regarded as an abnormal aware loss that promotes the generated anomalies at the image generation stage to be more similar to the real anomalies in terms of semantics and category features. Therefore, this abnormal aware loss alleviates the impact of the abnormal attribute mismatch in the unpaired images. If the pixel-wise losses (L1 or L2 loss) are directly applied between MF_{abnl} and MI_{abnl}, then the generator will be forced to generate anomalies with the exact same features as the real anomaly MI_{abnl} due to per-pixel mapping. Therefore, the feature-controlling effect of anomaly masks is broken, and the diversity of generated images is greatly compromised.

4.2.2. Proposed local versus global consistency discriminators

The local versus global consistency discriminators aim to discern whether the global image and cropped local region area real. Each discriminator has two sub-networks, namely, the local and global subnets. This discriminator compresses the images into small feature vectors. The outputs of the two sub-networks are fused together through a concatenated layer to predict the real probability of the image, thereby improving the quality of anomaly generation in small regions. The structure of proposed discriminator is shown in Fig. 7.

The four discriminators ($D1_{abnl}$, $D2_{abnl}$, $D1_{norm}$, $D2_{norm}$) in our Anomaly-GAN have the same structure. Concretely, the local subnet takes as input the cropped region rescaled to 64 × 64 pixels. This subnet consists of four convolutional layers (Conv) and a single fully-connected layer (FC) that outputs a single 1024-dimensional vector. This vector represents the local semantic feature within the cropped anomaly region. The global subnet follows the same pattern and uses the whole image (512 × 512 pixels) as input. Given that the initial input resolution is 8 times that of the local discriminator, the global subnet has 3 more convolutional layers than the local discriminator. The output is a 1024-dimensional vector representing the global context. All convolutions use 5 × 5 kernels with a stride of 2. Each convolution is followed by a batch normalization (BN) layer and LeakyReLU.

The two 1024-dimensional vectors outputted by the local and global subnets are concatenated together into a 2048-dimensional single vector. This aggregated feature is then processed by the FC and sigmoid layers to output a scalar for identifying real and fake images. The scalar value is distributed within the range [0, 1] and represents the real probability of images in terms of local and global.

4.2.3. Objective functions

Our objective function contains three terms, namely, locally aware D2 adversarial loss, abnormal aware loss, and cycle consistency loss. Locally aware D2 adversarial loss helps Anomaly-GAN synthesize the anomalies with high image quality and rich diversity. Abnormal aware loss can make the generated anomalies more realistic in terms of their details and semantic features. The cycle consistency loss enables Anomaly-GAN to generate anomaly images from a small number of anomalous images and a rich set of auxiliary anomaly-free images.

(1) Locally aware D2 adversarial loss

As stated in Section 1, original GAN usually suffers from model collapse and generates images with poor quality and diversity. To address this issue, a D2 adversarial loss (Nguyen et al., 2017) is introduced to stabilize the model training and enrich the diversity features of the generated images. Given that our discriminator needs to judge the "real" local anomalies and global textures at the same time, we design a locally aware D2 adversarial loss on the basic of D2 adversarial loss. The improved adversarial loss helps improve the image quality of the generated small-area anomalies.

For generator $G_{norm2abnl}$ and discriminator $D1_{abnl}$, the objective function is written as

$$\begin{aligned} &L_{adv}(G_{norm2abnl}; D1_{abnl}; I_{norm}; I_{abnl}; MI_{abnl}; MF_{abnl}) \\ &= E_{(I_{norm}\in p_{data}(I_{norm}))}[\log D1_{abnl}(I_{abnl}, MI_{abnl})] \\ &+ E_{(I_{abnl}\in p_{data}(I_{abnl}))}[\log(1 - D1_{abnl}(F_{abnl}, MF_{abnl})]. \end{aligned} \tag{2}$$

Eq.(2) is essentially a process of maximizing $D1_{abnl}$ and minimizing $G_{norm2abnl}$. The closer $D1_{abnl}(I_{abnl}, MI_{abnl})$ is to 1, the more accurately can the discriminator identify the real samples. Meanwhile, the closer $D1_{abnl}(F_{abnl}, MF_{abnl})$ is to 0, the more accurately can the discriminator distinguish the generated samples. This optimization is summarized as $Max_{D1_{abnl}} L_{adv}(\cdot)$. At the stage of generator training, the discriminator is fixed. If $G_{norm2abnl}$ is stronger, then the discriminator will make a misjudgment, thereby increasing $D1_{abnl}(F_{abnl}, MF_{abnl})$ and making $E_{(I_{abnl}\in p_{data}(I_{abnl}))}[\log(1 - D1_{abnl}(F_{abnl}, MF_{abnl})]$ closer to 0, that is, the value of the whole formula is reduced. Therefore, the optimization can be summarized as $Min_{G_{norm2abnl}} L_{adv}(\cdot)$. The overall optimization makes the generator $G_{norm2abnl}$ generate an anomaly image F_{abnl} that is similar to the members of the "Abnl" domain in terms of global texture and local anomaly region.

Following the D2 adversarial loss, $D2_{abnl}$ and $D2_{norm}$ are added to the original GAN to improve the diversity and quality of the generated images. Compared with $D1_{abnl}$ and $D1_{norm}$, $D2_{abnl}$ and $D2_{norm}$ assign higher scores (output 1) to the generated images and smaller scores (output 0) to the real images. The objective function is written as

$$\begin{aligned} &L_{adv\_2}(G_{norm2abnl}; D2_{abnl}; I_{norm}; I_{abnl}; MI_{abnl}; MF_{abnl}) \\ &= E_{(I_{norm}\in p_{data}(I_{norm}))}[\log D2_{abnl}(F_{abnl}, MF_{abnl})] \\ &+ E_{(I_{abnl}\in p_{data}(I_{abnl}))}[\log(1 - D2_{abnl}(I_{abnl}, MI_{abnl}))]. \end{aligned} \tag{3}$$

Next, the objective of locally aware D2 adversarial loss for the generator $G_{norm2abnl}$ and discriminators $D1_{abnl}$ and $D2_{abnl}$ is written as

$$\begin{aligned} &L_{GAN}(G_{norm2abnl}; D1_{abnl}; D2_{abnl}; I_{norm}; I_{abnl}; MI_{abnl}; MF_{abnl}) \\ &= L_{adv}(G_{norm2abnl}; D1_{abnl}; I_{norm}; I_{abnl}; MI_{abnl}; MF_{abnl}) \\ &+ \lambda_1 L_{adv\_2}(G_{norm2abnl}; D2_{abnl}; I_{norm}; I_{abnl}; MI_{abnl}; MF_{abnl}). \end{aligned} \tag{4}$$

where λ_1 controls the relative importance between sample similarity and diversity, $L_{adv}(\cdot)$ guarantees the quality of the generated images, and $L_{adv\_2}(\cdot)$ guarantees the rich diversity of the generated images.

Similarly, the objective of locally aware D2 adversarial loss for generator $H_{abnl2norm}$ and discriminators $D1_{norm}$ and $D2_{norm}$ is written as

$$\begin{aligned} &L_{GAN\_2}(H_{abnl2norm}; D1_{norm}; D2_{norm}; I_{norm}; I_{abnl}; MI_{norm}; MF_{norm}) \\ &= L_{adv}(H_{abnl2norm}; D1_{norm}; I_{norm}; I_{abnl}; MI_{norm}; MF_{norm}) \\ &+ \lambda_1 L_{adv\_2}(H_{abnl2norm}; D2_{norm}; I_{norm}; I_{abnl}; MI_{norm}; MF_{norm}). \end{aligned} \tag{5}$$

(2) Cycle consistency loss.

If the number of anomalous training samples is sufficient, then the locally aware D2 adversarial loss can guarantee the quality and diversity of the generated train surface anomaly images. However, in actual railway operations, the number of abnormal images on the train surface is too small, hence preventing the generation model from directly using real abnormal images to synthesize anomaly images. Meanwhile, when normal train surface images are sufficient and readily available, the anomaly images are ideally generated based on anomaly-free images. Therefore, we introduce a cycle consistency loss (Zhu et al., 2017) to achieve the goal of utilizing normal images to generate anomaly images. This loss can be expressed as

$$\begin{aligned} &L_{cyc}(G_{norm2abnl}, H_{abnl2norm}) \\ &= E_{(I_{norm}\in p_{data}(I_{norm}))}\|F_{norm} - I_{norm}\|_1 \\ &+ E_{(I_{abnl}\in p_{data}(I_{abnl}))}\|F_{abnl} - I_{abnl}\|_1. \end{aligned} \tag{6}$$

For the "$Normal \rightarrow Abnormal \rightarrow Normal$" process, we treat the generated abnormal image $\{Generated\ \ F_{abnl}\}$ from $G_{norm2abnl}$ as the input of $H_{abnl2norm}$ order to generate the reconstructed image $\{Reconstructed\ \ F_{norm}\}$, which is similar to $\{Input\ \ I_{norm}\}$. For the "$Abnormal \rightarrow Normal \rightarrow Abnormal$" process, we also want the reconstructed image $\{Reconstructed\ \ F_{abnl}\}$ to be close to the abnormal image $\{Input\ \ I_{abnl}\}$. The measure of similarity is based on L1 distance. Through the constraint of $L_{cyc}(\cdot)$, the generated abnormal image $\{Generated\ \ F_{abnl}\}$ maintains similarity with the input anomaly-free image $\{Input\ I_{norm}\}$ in the non-anomalous region.

(3) Abnormal aware loss

Given that the anomaly on the train surface only occupies a small part of the whole image, we need to pay more attention to the generation of small-area anomalies. For this purpose, we introduce the following abnormal aware loss designed in the abnormal aware network (Section 4.2.1) to our model:

$$\begin{aligned} &L_{aware}(MF_{norm}, MI_{norm}, MF_{abnl}, MI_{abnl}) \\ &= L_{semantic}(MF_{norm}, MI_{norm}) \\ &+ L_{semantic}(MF_{abnl}, MI_{abnl}). \end{aligned} \tag{7}$$

where $L_{semantic}(MF_{norm}, MI_{norm})$ aims to make the generated small-area normal region MF_{norm} (corresponding to M_{prior}) close to the real normal region MI_{norm} (corresponding to M_{exp}) in terms of semantic features, and $L_{semantic}(MF_{abnl}, MI_{abnl})$ aims to make the generated small-area anomalous region MF_{abnl} maintain a similar feature distribution to the real abnormal region MI_{abnl}. Therefore, even if asymmetric binary masks are included in the two-cycle process, the model can focus on locally feature similarity to ensure that small-area anomaly is generated with high quality.

(4) Full objective loss function

The full objective loss function $L_{Anomaly-GAN}$ consists of locally aware D2 adversarial loss, abnormal aware loss, and cycle consistency loss.

$$\begin{aligned} & L_{Anomaly-GAN} \\ & = L_{GAN}(G_{norm2abnl}; D1_{abnl}; D2_{abnl}; I_{norm}; I_{abnl}; MI_{abnl}; MF_{abnl}) \\ & + L_{GAN\_2}(H_{abnl2norm}; D1_{norm}; D2_{norm}; I_{norm}; I_{abnl}; MI_{norm}; MF_{norm}) \\ & + L_{cyc}(G_{norm2abnl}, H_{abnl2norm}) \\ & + L_{aware}(MF_{norm}, MI_{norm}, MF_{abnl}, MI_{abnl}). \end{aligned} \tag{8}$$

5. Experiments

5.1. Dataset and setups

This study mainly aims to use normal images to synthesize various anomalies and verify the effectiveness of the proposed data augmentation method. The experiments involve two aspects, namely, analysis of the proposed Anomaly-GAN algorithm and validation of the performance improvement brought by data augmentation to the DL-based anomaly detection model.

Dataset. We collected a dataset from real-world railway safety monitoring scenarios called train surface anomaly dataset (TSAD). In this dataset, 638 abnormal images containing paper scraps (196 images), screws_A (138 images), screws_B (140 images), and cloth strips (164 images) are collected from the TEDS platform. Each image has a size of 2048 × 1400 pixels and are in grayscale with various foregrounds (anomalies) and backgrounds. To ensure the smooth training of Anomaly-GAN, the original images are resized to 512 × 512 pixels. Given that the model utilizes normal images to generate abnormal images, we randomly select the corresponding normal images for each abnormal category for model training. In addition, the prior-knowledge-based and expert-experience-based binary masks in the two-cycle process are taken from the mask pool. A pair of normal and abnormal images is taken as model input along with the corresponding binary anomaly mask.

Anomaly detection baseline. To verify the effectiveness of the generated anomaly images in the anomaly detection task, we adopt the popular DL model Mask R-CNN (He et al., 2017) as the baseline to detect and segment train surface anomalies. We train this baseline model with an augmented dataset consisting of a combination of generated and real anomaly images. By comparing the anomaly detection performance of the baseline model before and after data enhancement, the image quality of the anomaly images generated by our proposed Anomaly-GAN can be verified.

Evaluation metrics. To statistically assess the diversity and quality of the anomaly images generated by Anomaly-GAN, we use two popular metrics, namely, the Frechet inception distance (FID) (Heusel et al., 2017b) and learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018). FID can be computed as

$$FID(x, y) = \|\mu_x - \mu_y\|^2 + Tr(\Sigma_x + \Sigma_y - 2(\Sigma_x\Sigma_y)^{\frac{1}{2}}). \tag{9}$$

where μ_x, μ_y and Σ_x, Σ_y denote the feature means and covariance matrices of the real image x and generated image y, respectively.

We also measure the quality of the generated anomaly images and the effectiveness of the proposed data augmentation method by evaluating the anomaly detection accuracy of the baseline model. The average precision (AP) (Tulbure et al., 2022; Zhang et al., 2021b) and intersection over union (IoU) metrics (Bergmann et al., 2021; Xia et al., 2020) are widely used in anomaly object detection. AP refers to the area under the Precision–Recall curve. AP and IoU are defined as

$$Precision = \frac{TP}{(TP + FP)}, \quad Recall = \frac{TP}{TP + FN}. \tag{10}$$

$$IOU = \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} P_{ji} - \sum_{i=0}^{k} P_{ii}}. \tag{11}$$

where TP is the true positive sample, FP is the false positive sample, and FN is the false negative sample. In Eq. (11), p_{ij} denotes the number of pixels whose true category i is predicted to be category j, and p_{ji} and p_{ii} are defined in a similar way.

Implementation details. For Anomaly-GAN, the sizes of the input images and masks are unified to 512 × 512 pixels. The model is optimized by AdamW (Loshchilov & Hutter, 2017) with the learning rate of $2 \times e^{-3}$ and batch size of 16. The epoch is set to 1000 at the training stage. The weight parameters in the object function are set to $\lambda_1 = 0.3$ to ensure a smooth training. For the anomaly detection baseline model, the training parameters are set as follows: input image size 512 × 512, base learning rate (0.001), and batch size (24). After every 2000 epochs, the learning rate is reduced by a factor of 10. The training process is stopped after 5000 epochs. The experimental environment includes the Pytorch platform, an Ubuntu 16.04 operating system, an Intel Core i9-9920X CPU@3.50 GHz, and three NVIDIA RTX 2080Ti GPUs with 33 GB memory.

5.2. Ablation study on each part of anomaly-GAN

We then analyze the efficacy and effects of each component of Anomaly-GAN. The main components of this model include the proposed mask pool, abnormal aware loss, and local versus global consistency discriminators. We visualize the images generated with different components and quantitatively evaluate the image quality improvement brought about by different components.

The mask pool. Some synthetic images are illustrated in Fig. 8 to demonstrate the control ability of Anomaly-GAN. The first, second, third, and fourth blocks illustrate the generated cloth strips, paper scraps, screws_A, and screws_B, respectively. Unlike traditional data augmentation methods, our Anomaly-GAN can control the image generation process through expert-experience-based masks and prior-knowledge-based masks from the mask pools. The prior information from the anomaly masks offers four benefits. First, different types of anomalies with varied shapes can be generated. Second, anomalies with different rotation angles ($0° \sim 360°$) can be generated and seamlessly connected to the background texture. Third, the spatial locations of the anomalies in the generated image can be changed randomly. Fourth, single images with multiple anomalies can also be generated. Unlike the classical geometric-transformation-based data augmentation methods, when scaling, rotation, and translation operations are performed, the anomaly features remain unchanged, and no new anomalies appear. Anomaly-GAN can adaptively change the abnormal textures during the transformation process and naturally blend into the image background, thereby enriching the feature representation. Leveraging these new anomalies with more diverse features to train the anomaly detection baseline model will improve its generalization ability. The performance improvement brought about by the data augmentation method to the baseline model is reported in Section 5.5. Furthermore, given that screws_A and screws_B have fixed physical properties, their shape and angle do not need to be changed.

Abnormal aware loss. Fig. 9 verifies the effect of abnormal aware loss on the image quality of small-area anomaly generation. Unlike the commonly used pixel-level loss (e.g., L2 loss) that performs pixel-by-pixel mapping, the abnormal aware loss performs loss computation at different semantic feature layers of the proposed abnormal aware network. As shown in the second and third rows of Figs. 9(a) and 9(b), although the background textures of the generated images are clear, the images generated by abnormal aware loss have more accurate and diverse small-area anomalies compared with those generated with L2 loss. L2 loss only calculates the distance between the average values of pixels in the abnormal area and ignores the difference between the abnormal texture and semantic features, thereby failing to effectively synthesize abnormal detailed textures. For the quantitative analysis, we also calculate the FID and LPIPS scores between 600 randomly selected generated abnormal images and the real anomaly images. As reported

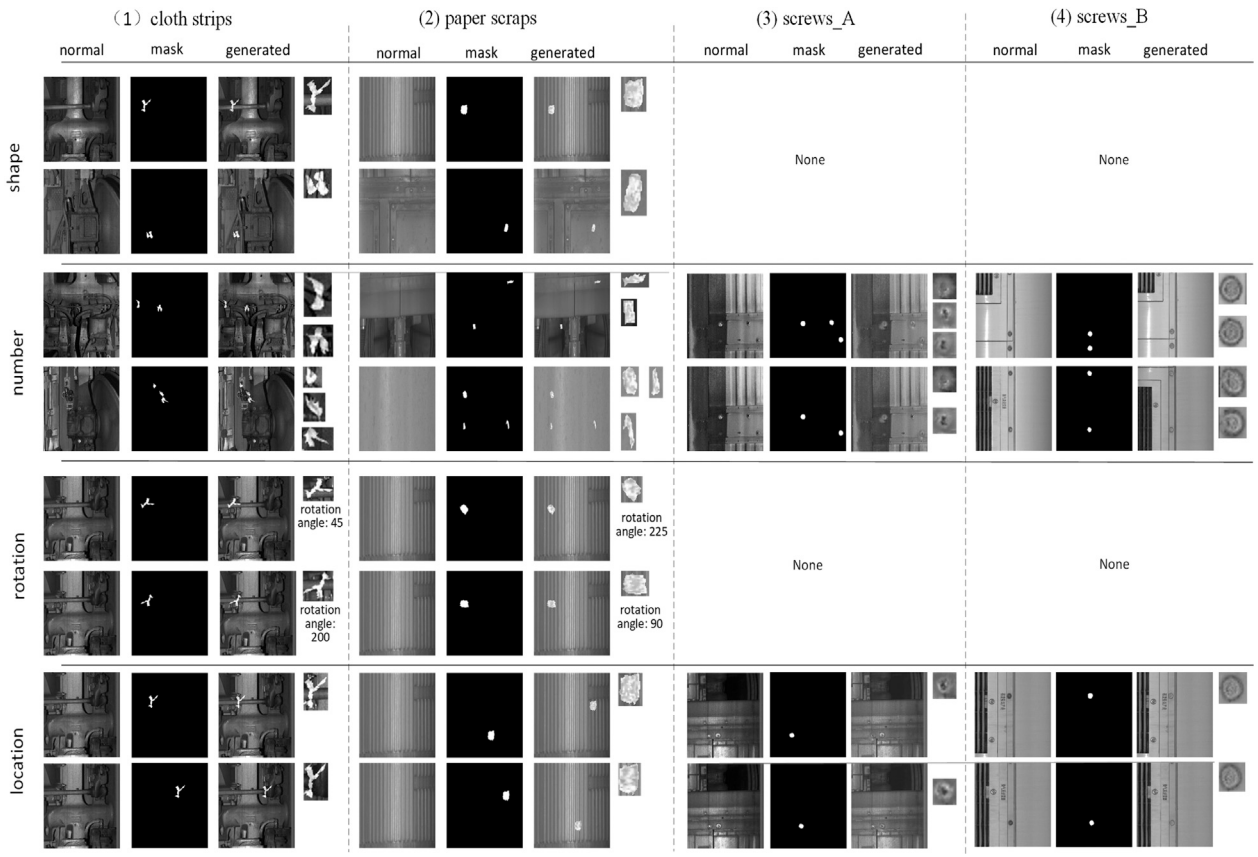


Fig. 8. The effectiveness of the anomaly mask in the controllability of image generation. The diversity of different anomaly attributes such as shape, part number, rotation and position is displayed.

Table 1
Ablation results of different components.

Components	FID score				LPIPS score			
	Cloth strips	Paper scrips	screws_A	screws_B	Cloth strips	Paper scrips	screws_A	screws_B
L2 loss	212	158	93	101	0.587	0.453	0.263	0.282
Global discriminator	178	123	81	83	0.526	0.385	0.177	0.193
Full approach	**94**	**68**	**40**	**44**	**0.428**	**0.321**	**0.143**	**0.16**

in Table 1, if the abnormal aware loss is replaced with L2 loss, the FID and LPIPS scores of the generated images increase significantly, that is, the image quality becomes worse.

Local versus global consistency discriminators. We then investigate the influence of local versus global consistency discriminators by training a model that only uses the global discriminator and comparing its results with those of the full approach. The experimental results are shown in Fig. 9. When the local discriminator is not used, the results of anomaly region are filled with blurred feature. By using both global and local discriminators, we can obtain local and global high-quality generative results. As shown in Table 1, compared with those images generated by a single global discriminator, the images generated by the local versus global consistency discriminators have significantly lower FID and LPIPS, respectively, thereby implying more realistic anomalies with sharper details.

Visualization of the data distribution. To further verify the image generation ability, we use t-SNE (Wang et al., 2018) to visualize the data distribution changes during the Anomaly-GAN training. We take the training process of cloth strips as an example, and the generation process of other anomaly types show similar trends. As illustrated in Fig. 10, the distribution of abnormalities is highly concentrated at the initial stage, and a significant boundary can be observed between the generated (indicated by red dot "0") and real anomalies (indicated by blue dot "1"). When the training epoch goes deep into 600 (Fig. 10(b)), the distribution of the generated samples gradually spans to the boundary of the real abnormal image domain, illustrating an obvious diffusion trend. At the end of the training stage(Fig. 10(c)), the generated anomalies are closely distributed around the real anomalies, and some of the generated sample spreads to places not touched by the real anomaly images, thereby proving that Anomaly-GAN not only generates images that are similar to real anomalies but also generates new samples that have never been seen before. The shape, angle, spatial location, and other geometric properties in the neighboring anomalies have also changed, thereby resulting in new anomaly features. To facilitate comparison with the distribution of real data, we equally divide the real cloth strip images into two datasets and show their distributions in Fig. 10(d). The red and blue dots in Fig. 10(c) and Fig. 10(d) are confused together, thereby suggesting that Anomaly-GAN can generate sufficiently realistic abnormal images.

5.3. Comparative experiment on the number of images required for anomaly-GAN training

Compared with the original GAN (CGAN (Mirza & Osindero, 2014) and LSGAN (Mao et al., 2017)), to directly learn the distribution of real abnormal samples, our network uses the prior information brought by the mask to efficiently learn the features of abnormal regions and

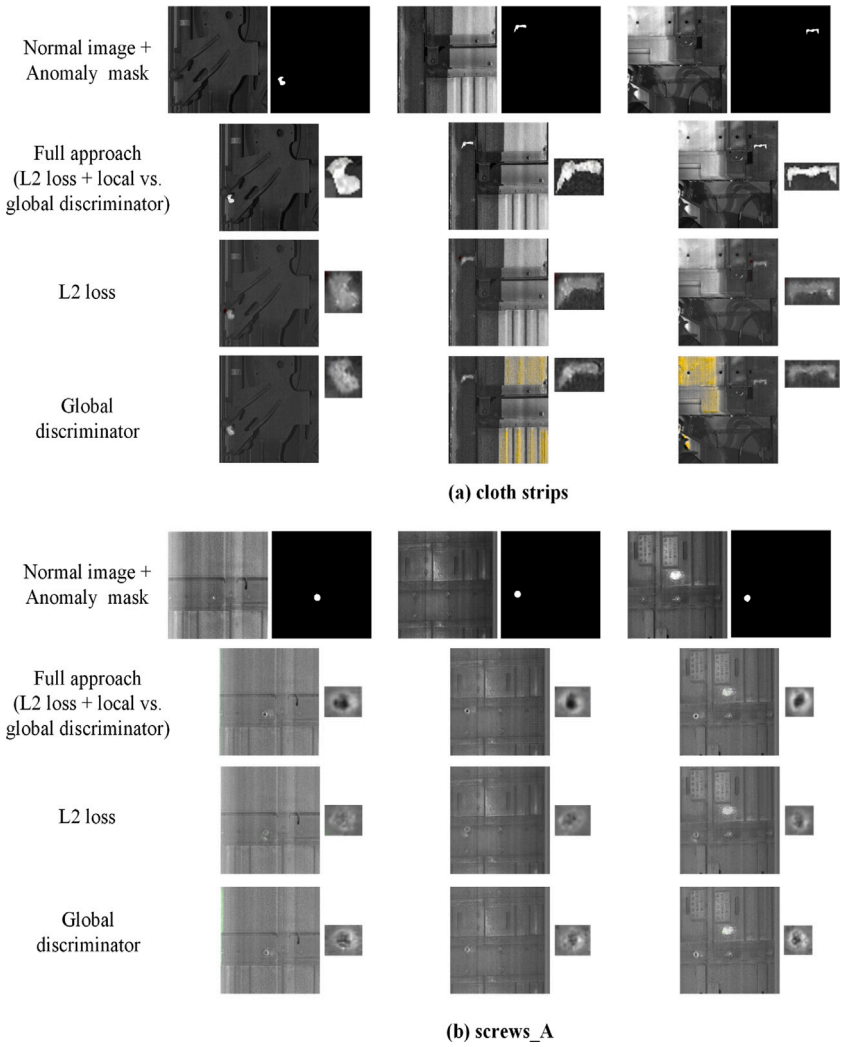


Fig. 9. Comparison of different losses and discriminators. The generation tasks of cloth strips and screws_A are chosen for exhibition.

enhance the abnormal sample generation ability. These masks containing diverse anomaly attributes can guide Anomaly-GAN to directly generate new anomalies from high probability anomaly distribution domains. Instead of generating anomaly images from scratch, Anomaly-GAN directly generates anomalies on different normal images according to the prior distribution information in masks. Therefore, our model can be trained with less images. We conduct an experiment to prove our argument, and the results are shown in Fig. 11, where the horizontal axis shows the number of image pairs (real abnormal and normal images) used for the training, and the vertical axis lists the FID scores of the generated anomalous images outputted by Anomaly-GAN trained with the corresponding number of image pairs. For the horizontal axis, the amount of real abnormal images and prior-knowledge-based anomaly masks is increased through traditional geometric transformations, such as rotation $(90°, 135°, 180°, 270°, etc.)$ and flips (up vs. down, left vs. right). The increase of normal images and corresponding masks are ascribed to the sufficiency of normal data and hand-drawn masks. The change trend of the vertical axis shows that the quality of the generated abnormal images improves along with increasing data volume. For comparison, we randomly divide each category of real anomalies into two sets to obtain the average FID scores. The average FID is denoted by the dashed line. Therefore, the FID scores between the generated and real anomalies close to or below 93 (cloth strips), 42 (screws_A), 46 (screws_B), and 69 (paper scraps) suggest that the generated images are sufficiently realistic and have high quality. In other words, in the train surface anomaly generation task, the proposed Anomaly-GAN can be trained with a small number of samples (such as 400 to 1200 training image pairs) to generate sufficiently realistic anomaly images.

5.4. Comparison with other image generative methods

We then compare Anomaly-GAN with several classical generative methods, such as CGAN (Mirza & Osindero, 2014), LSGAN (Mao et al., 2017), CycleGAN (Zhu et al., 2017), FD-Cycle-GAN (Su et al., 2022), and Pix2Pix (Isola et al., 2017), and the results are shown in Fig. 12. As illustrated, the images generated by CGAN and LSGAN have the worst quality, and the details of their background texture and small-area anomalies are difficult to synthesize because the training of these two models is not stable enough. Specifically, the generator would rather generate simpler yet "safe" samples than high-quality and diverse samples. Although LSGAN uses the least squares loss as the objective function to stabilize the training, the quality of the generated images has not improved by much compared with CGAN. The anomaly images generated by Pix2Pix have better quality than those generated by CGAN and LSGAN. Although the shape of generated anomalies in Pix2Pix is more accurate, the details of small anomalies still need to be improved, such as screws_A and screws_B. Furthermore, when the background texture is complex, the normal regions reconstructed by Pix2Pix is not ideal. The images generated by CycleGAN and FD-Cycle-GAN have a more stable background than those generated by Pix2Pix. However, the connection between the small-area anomalies and background is

Table 2
The FID and LPIPS scores of different methods.

Method	Param (M)	Times (s)	FID score				LPIPS score			
			Cloth strips	screws_A	screws_B	paper strips	Cloth strips	screws_A	screws_B	paper strips
CGAN (Mirza & Osindero, 2014)	–	–	219	114	124	229	0.754	0.387	0.425	0.712
LSGAN (Mao et al., 2017)	–	–	213	109	117	217	0.765	0.349	0.372	0.696
CycleGAN (Zhu et al., 2017)	40.3	0.132	175	91	98	173	0.611	0.266	0.286	0.522
FD-Cycle-GAN (Su et al., 2022)	42.8	0.146	153	79	87	152	0.594	0.256	0.274	0.498
Pix2Pix (Isola et al., 2017)	217.4	0.143	116	53	58	103	0.487	0.189	0.213	0.373
proposed method	43.5	0.152	**94**	**40**	**44**	**68**	**0.368**	**0.143**	**0.16**	**0.291**

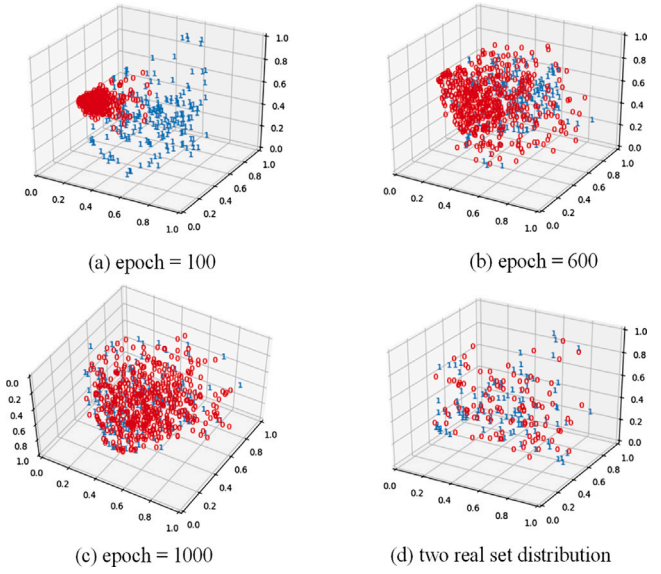


Fig. 10. Visualization of abnormal data distribution in different training epochs. The red dot "0" represents the generated abnormal image, whereas the blue dot "1" indicates the real abnormal image.

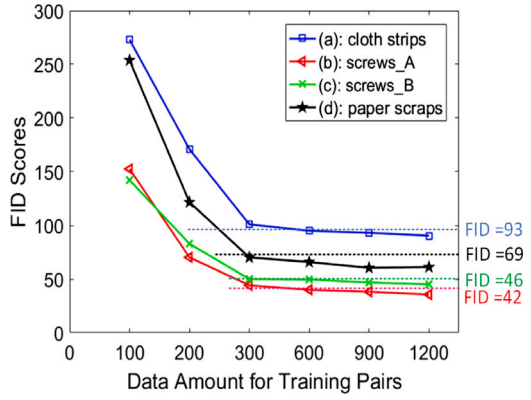


Fig. 11. Image quality of the model trained on different amounts of data.

unnatural, and some of the anomaly regions are not realistic enough. Compared with these generative methods, the proposed Anomaly-GAN synthesizes more realistic images with sharper details of small-area anomalies and better background texture. Moreover, the generated images have controllable shapes and rich diversity.

For a quantitative analysis, we use the FID and LPIPS scores to evaluate the quality of the generated anomalies. The results are reported in Table 2. Anomaly-GAN obtains the lowest FID score, which indicates a high similarity between the generated anomalies and the real image. Anomaly-GAN also obtains the lowest LPIPS score, which indicates that its generated images have the best quality and have a high perceptual similarity with the real anomalous images. Table 2 also lists the comparison of model parameters and time consumption between our algorithm and other generation algorithms. Anomaly-GAN obtained a competitive result, which achieves a speed of 0.152 s with 43.5M parameter for single image generation.

Table 3
Division of the real abnormal dataset.

	Cloth strips	screw_A	screw_B	Paper scraps
Train set	123	103	105	147
Test set	41	35	35	49

5.5. Data augmentation applied in train surface anomaly detection

To demonstrate the applications of Anomaly-GAN in train surface anomaly detection, we conduct further experiments using the anomaly detection baseline model. Given the insufficient number of anomaly images in the train surface anomaly detection scene, we use data augmentation based on Anomaly-GAN to improve the generalization ability of the baseline model.

Train surface anomaly detection. For comparison, we set up nine control groups up to validate the improvement of our data augmentation method over the baseline model. The TEDS system collects 638 train surface anomaly images, of which 25% are used for model testing. The amount of training and testing data for each anomaly category is shown in Table 3, whereas Table 4 presents the settings of different control groups. To verify that Anomaly-GAN can use a limited number of anomaly samples for model training and generate diverse anomaly images to improve the accuracy of the anomaly detection baseline model, we consider the case of few-shot training data. In group 1 (no augmentation group), the baseline model is directly trained with 486 anomaly images to measure the anomaly detection performance in the case of limited samples. In group 2 (normal augmentation), traditional data augmentation methods, such as rotation and flips, are used to increase the number of original training samples from 486 to 4800. In group 3, a simple "copy-paste" method is used to paste the abnormal objects into the normal images and maintain the same number of training samples as group 2. For data augmentation in the other control groups, we use a limited number of anomaly samples to train different image generative models to synthesize new abnormal samples. The prior-knowledge-based and expert-experience-based anomaly masks from the mask pool are embedded onto the generative model to increase the diversity of the generated images. After data augmentation, the number of different anomaly categories in each comparative experiment is balanced. The number of each anomaly category has been expanded to 1200. We use Mask R-CNN (He et al., 2017) as the baseline model for each control group to detect train surface anomaly. The data volume of enhancement training dataset in each control group is displayed in Table 4. The test dataset in each group is the same as that in Table 3.

As reported in Table 4, the mAP and mIoU significantly increase along with the data volume. As for the baseline, the experimental group with data augmentation is significantly better than the experimental group without data augmentation, thereby highlighting the importance of big data volume for improving the generalization ability of the deep

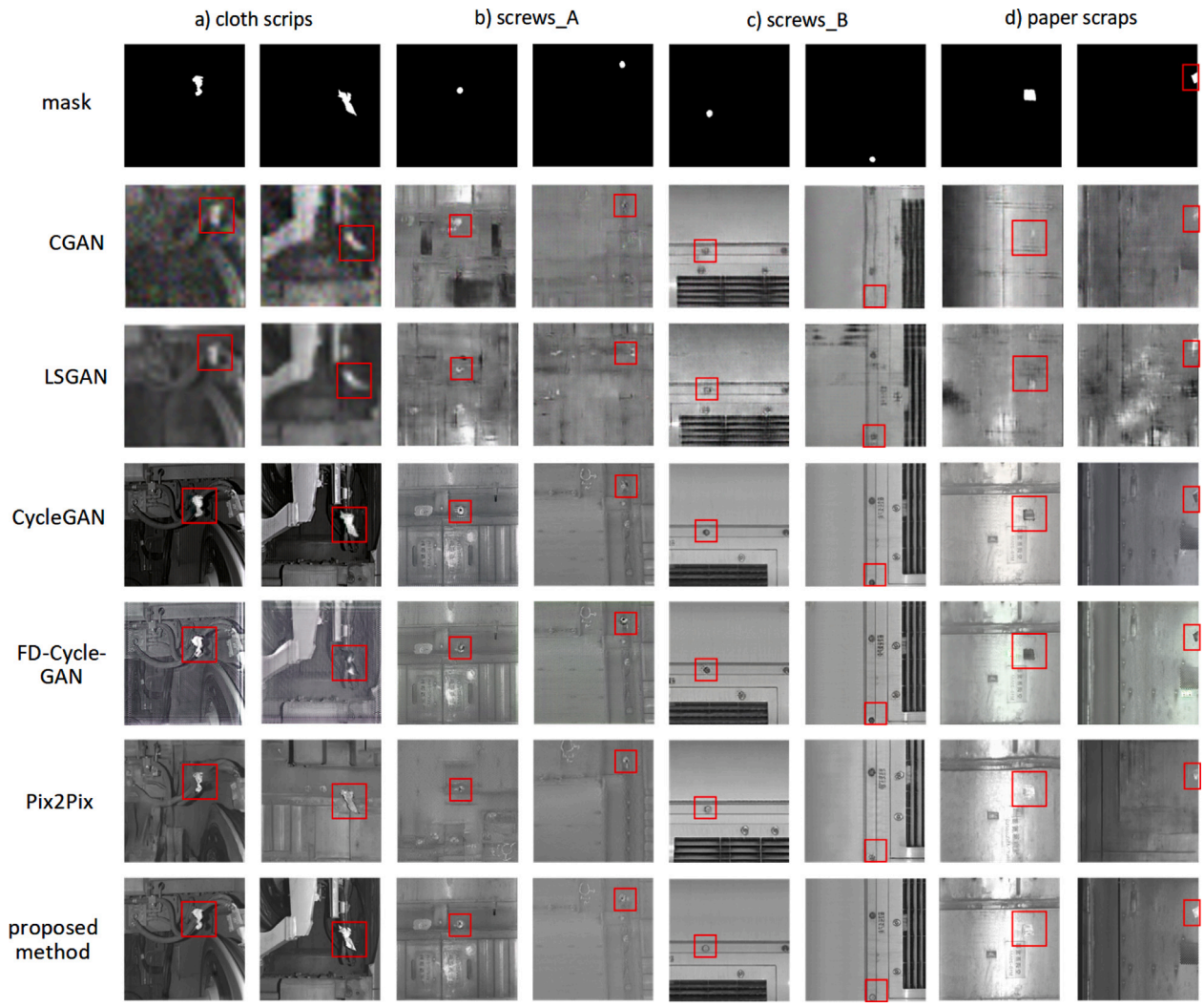


Fig. 12. Qualitative comparison between different anomaly image generation algorithms.

Table 4
Performance of inspection model with different augmentation methods.

Method	Mask pool	Data number	Detection					Segmentation				
			Cloth strips	screws _A	screws _B	Paper scraps	**mAP**	Cloth strips	screws _A	screws _B	Paper scraps	**mIoU**
No aug		478	72.2	50.5	61.1	66.4	62.5	69.3	49.6	61	66.1	61.5
Normal aug		4800	80.2	57.3	70.7	74.5	70.6	77.5	55.6	68	73.5	68.6
Copy-paste		4800	81.3	58.5	72.1	75.8	71.9	79.8	58.2	69.3	75.7	70.7
CGAN (Mirza & Osindero, 2014)	✓	4800	77.4	53.3	65.4	70.8	66.7	74.8	52.5	64.2	70.2	65.4
LSGAN (Mao et al., 2017)		4800	78.9	55.7	68.2	73.4	69.1	75.4	54.3	66.4	72.7	67.2
CycleGAN (Zhu et al., 2017)	✓	4800	87.8	60.6	75.7	81.9	76.5	83.5	60.2	73.1	80.9	74.7
FD-Cycle-GAN (Su et al., 2022)	✓	4800	90.4	62.5	77.3	84.6	78.7	89.3	61.1	75.2	83.4	77.2
Pix2Pix (Isola et al., 2017)	✓	4800	93.2	65.2	81.2	88.7	82	91.8	64.2	80.5	86.9	80.8
proposed method	✓	4800	**97.7**	**71.2**	**87.9**	**95.5**	**88.1**	**95.5**	**69.7**	**85.3**	**92.6**	**85.7**

learning model. In addition, the baseline model trained with images synthesized by the mask pool based generative methods generally have better detection accuracy, thereby suggesting that the mask pool can further improve the ability of the baseline model to detect new anomalies that do not appear in the training set. By contrast, the data augmentation algorithms based on traditional geometric transformations (normal augmentation and "copy-paste" method) show limited improvements (70.6% of mAP and 68.6% of mIoU for the normal augmentation method, 71.9% of mAP and 70.7% of mIoU for "copy-paste" method) in their detection accuracy and cannot make the baseline model learn more new anomaly features. Pix2Pix shows better image generation quality than CycleGAN and FD-Cycle-GAN, hence contributing to a higher accuracy improvement for the baseline. This result can be ascribed to the fact that the small-area anomalies generated by CycleGAN and FD-Cycle-GAN are inferior to those generated by Pix2Pix in terms of detail and realness. CGAN and LSGAN also show a slightly inferior performance compared with normal augmentation because the mode collapse leads to an unstable image generation, and the generated low-quality images may adversely affect the training of the baseline model. With the guidance of the anomaly masks, Anomaly-GAN can generate new anomalies with various shape and texture features in different backgrounds. The baseline model trained with these highly diverse and high-quality images have a stronger feature expression

Table 5
Performance of our data augmentation on state-of-the-art DL-based anomaly detection model.

Model	Training data	Detection (mAP)	Segmentation (mIoU)	Time (s)
queryinst (Fang et al., 2021)	real	72.9	71.7	0.062
	generate	80.4	76.2	
	real + generate	84.6	84.2	
Mask R-CNN (He et al., 2017)	real	73.5	72.8	0.041
	generate	82.7	79.5	
	real + generate	88.2	85.8	
MS r-cnn (Huang et al., 2019)	real	74.8	73.3	0.064
	generate	83.8	82.6	
	real + generate	90.1	88.7	
retinanet (Lin et al., 2017)	real	74.9	-	0.054
	generate	86.9		
	real + generate	91.7		

Table 6
Performance on different number of generated images.

Groups	Cloth strips		screws_A		screws_B		Paper scraps		**mAP**	**mIoU**
	Real	Generated	Real	Generated	Real	Generated	Real	Generated		
1	123	0	103	0	105	0	147	0	62.5	61.5
2	0	123	0	103	0	105	0	147	59.3	58.6
		400		400		400		400	63.2	58.3
		800		800		800		800	72.3	70.7
3	123 × 2	400	103 × 2	400	105 × 2	400	147 × 2	400	68.8	67.1
		800		800		800		800	78.4	77.8
		1200		1200		1200		1200	85.0	83.1
4	123 × 4	800	103 × 4	800	105 × 4	800	147 × 4	800	84.4	82.8
		1200		1200		1200		1200	86.2	83.5
		1600		1600		1600		1600	87.7	84.9
5	123 × 6	1200	103 × 6	1200	105 × 6	1200	147 × 6	1200	88.0	85.6
		1600		1600		1600		1600	88.1	85.7
6	123 × 8	1600	103 × 8	1600	105 × 8	1600	147 × 8	1600	88.1	8 5.7
		1800		1800		1800		1800	88.2	85.8

ability so as to obtain the best detection and segmentation results. Compared with the no-augmentation group, Anomaly-GAN has 25.6% and 24.2% higher mAP and mIoU, respectively, whereas compared with Pix2Pix, Anomaly-GAN demonstrates 6.1- and 4.9-point improvements, respectively.

To further verify the universality of the proposed data enhancement method, we conducted experiments on several state-of-the-art detection models in recent years, such as queryinst (Fang et al., 2021), MS r-cnn (Huang et al., 2019) and retinanet (Lin et al., 2017). As shown in Table 5, training data includes three different settings, namely, "real", "generated", and "real+generated". The number of each anomaly category in the "real" set and "generated" set is set to 1800. The number of real and generated exceptions of each category in the "real+ generated " set is 600 and 1200 respectively. All generated images are output by Anomaly-GAN. The results show that training with realistic and diverse images generated by Anomaly-GAN is helpful to improve the performance of DL based anomaly detection models. In addition, the time consumption of CNN detectors applied in Table 5 on a single image is about 0.04–0.06 s. With the image generation time of anomaly-GAN (0.15 s), the time consumption of entire pipeline is about 0.019–0.021 s.

Optimal ratio of real and generated anomalies in the training set. We conduct additional experiments to investigate the effect of increasing the number of generated and real anomaly training images on the final results of the baseline model. We design six experimental groups for analysis as shown in Table 6. Group 1 only uses real images for baseline training, whereas group 2 only uses the generated images. In the other groups, the number of real and generated images grows exponentially. The number of normal images is increased through geometric transformations, such as rotation and flip. The comparison of groups 1 and 2 reveals that those models trained with the generated data can achieve a similar or better detection performance than the models trained with real data. In the within-group comparison, when the number of real images is fixed, the accuracy of the model increases along with the volume of generated images. Meanwhile, in the across-group comparison, if the number of generated images is fixed, then increasing the number of real images also improves the performance of the baseline model. In theory, we can use Anomaly-GAN to generate endless abnormal samples to improve the model accuracy. However, with the increasing number of generated images, the improvement in mAP and mIoU indicators will gradually slow down and stabilize at 88.1% and 85.7%, respectively. After reaching a certain limit, having more images will not improve accuracy. Therefore, the size of the real and generated data for each anomaly category is fixed to about 600 and 1600 images, respectively. In sum, Anomaly-GAN can generate high-quality and diversified fake images as a supplement to the training set and effectively improve the performance of the anomaly detection model.

6. Conclusion

In order to generate diversified anomaly objects on the train surface, we propose a data augmentation method called Anomaly-GAN based on mask pool, abnormal aware loss, and local versus global discriminators. The anomaly masks in the mask pool play an important role in guiding the generation of diverse anomalies. Anomalies with different shapes, rotation angles, spatial locations, and part numbers can be synthesized in a controllable manner using this model. The anomaly awareness loss pays more attention to small-area anomalies, thereby promoting the generation of anomalies with a more detailed texture and richer semantics. Meanwhile, the local versus global consistency discriminators combine local and global feature expressions, thereby promoting the

generators to generate more realistic images and improving the naturalness of the connection between the local anomalies and background. The experiments reveal that the augmented data using Anomaly-GAN obtain the highest mAP and mIoU. Our generated images also obtain the lowest FID and LPIPS scores compared with the other generation methods, thereby proving that our generated images have high quality and rich diversity.

In addition, our method also has some limitations. (1) The initial masks are still created by hand drawing, and there may be a small amount of unreasonable prior information (such as the position and rotation angle of the mask in the background) that is difficult to distinguish. How to generate the mask and evaluate its rationality in an intelligent way is still a challenge for us in the future. (2) The images we generated are not very sensitive to the change of light. How to simulate different light changes is our next research direction.

The expansion of our future work will mainly focus on two aspects. (1) The relationship between image features (such as illumination and texture) may be further analyzed to obtain more realistic and diversified abnormal object synthesis results. (2) This method will consider introducing more prior information in the next research, such as the severity of the anomaly, to further improve the control ability of the anomaly feature generation process. In summary, we hope to carry out more exploration in the future to further optimize the algorithm and make it better applied in actual scenarios.

CRediT authorship contribution statement

Ruikang Liu: Methodology, Software, Validation, Writing – review & editing. **Weiming Liu:** Supervision, Conceptualization, Project administration. **Zhongxing Zheng:** Data curation, Validation, Writing – review & editing. **Liang Wang:** Data curation, Writing – review & editing. **Liang Mao:** Data curation, Writing – review & editing. **Qisheng Qiu:** Writing – review & editing. **Guangzheng Ling:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Key Research and Developmente Program of China under Grant 2016YFB1200402-020, project of research on multi-scale foreign object detection system and related technical standards for ultra-long continuous space of urban rail platform.

References

Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., & Steger, C. (2021). The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, *129*(4), 1038–1059.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*, *29*.

Chen, C., Li, K., Zhongyao, C., Piccialli, F., Hoi, S. C., & Zeng, Z. (2020). A hybrid deep learning based framework for component defect detection of moving trains. *IEEE Transactions on Intelligent Transportation Systems*.

Chen, C., Li, K., Zhongyao, C., Piccialli, F., Hoi, S. C., & Zeng, Z. (2020). A hybrid deep learning based framework for component defect detection of moving trains. *IEEE Transactions on Intelligent Transportation Systems*.

Dai, W., Li, D., Tang, D., Wang, H., & Peng, Y. (2022). Deep learning approach for defective spot welds classification using small and class-imbalanced datasets. *Neurocomputing*, *477*, 46–60.

Dong, B., Li, Q., Wang, J., Huang, W., Dai, P., & Wang, S. (2019). An end-to-end abnormal fastener detection method based on data synthesis. In *2019 IEEE 31st international conference on tools with artificial intelligence* (pp. 149–156). IEEE.

Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., & Liu, W. (2021). Instances as queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6910–6919).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).

He, D., Yao, Z., Jiang, Z., Chen, Y., Deng, J., & Xiang, W. (2019). Detection of foreign matter on high-speed train underbody based on deep learning. *IEEE Access*, *7*, 183838–183846.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, *30*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, *30*.

Huang, Z., Huang, L., Gong, Y., Huang, C., & Wang, X. (2019). Mask scoring r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6409–6418).

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

Isomoto, Y., & Yoshine, K. (1995). Data structure and retrieval method of scenic image database based on fuzzy set theory. In *Proceedings of 1995 IEEE international conference on fuzzy systems, Vol. 2* (pp. 749–756). IEEE.

Kang, G., Gao, S., Yu, L., & Zhang, D. (2018). Deep architecture for high-speed railway insulator surface defect detection: Denoising autoencoder with multitask learning. *IEEE Transactions on Instrumentation and Measurement*, *68*(8), 2679–2690.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., & Cho, K. (2019). Augmentation for small object detection. arXiv preprint arXiv:1902.07296.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Liu, L., Cao, D., Wu, Y., & Wei, T. (2019). Defective samples simulation through adversarial training for automatic surface inspection. *Neurocomputing*, *360*, 230–245.

Liu, J., Ma, Z., Qiu, Y., Ni, X., Shi, B., & Liu, H. (2021). Four discriminator cycle-consistent adversarial network for improving railway defective fastener inspection. *IEEE Transactions on Intelligent Transportation Systems*.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2794–2802).

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Nguyen, T., Le, T., Vu, H., & Phung, D. (2017). Dual discriminator generative adversarial nets. *Advances in Neural Information Processing Systems*, *30*.

Niu, S., Li, B., Wang, X., & Lin, H. (2020). Defect image sample generation with GAN for improving defect recognition. *IEEE Transactions on Automation Science and Engineering*, *17*(3), 1611–1622.

Niu, S., Li, B., Wang, X., & Peng, Y. (2021). Region-and strength-controllable GAN for defect generation and segmentation in industrial images. *IEEE Transactions on Industrial Informatics*, *18*(7), 4531–4541.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1), 157–173.

Su, S., Du, S., & Lu, X. (2022). Geometric constraint and image inpainting-based railway track fastener sample generation for improving defect inspection. *IEEE Transactions on Intelligent Transportation Systems*.

Tulbure, A. A., Tulbure, A. A., & Dulf, E. H. (2022). A review on modern defect detection models using DCNNs–deep convolutional neural networks. *Journal of Advanced Research*, *35*, 33–48.

Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8798–8807).

Xia, Y., Zhang, Y., Liu, F., Shen, W., & Yuille, A. L. (2020). Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *European conference on computer vision* (pp. 145–161). Springer.

Yang, B., Liu, Z., Duan, G., & Tan, J. (2021). Mask2Defect: A prior knowledge based data augmentation method for metal surface defect inspection. *IEEE Transactions on Industrial Informatics*.

Yu, J., & Liu, J. (2021). Multiple granularities generative adversarial network for recognition of wafer map defects. *IEEE Transactions on Industrial Informatics*, *18*(3), 1674–1683.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586–595).

Zhang, Z., Ma, J., Huang, D., Zhou, Z., Wan, Z., & Qin, N. (2021). Fault diagnosis of train clamp based on faster R-CNN and one-class convolutional neural network. In *2021 IEEE 16th conference on industrial electronics and applications* (pp. 1394–1399). IEEE.

Zhang, H., Pan, D., Liu, J., & Jiang, Z. (2022). A novel MAS-GAN-based data synthesis method for object surface defect detection. *Neurocomputing*.

Zhang, J., Su, H., Zou, W., Gong, X., Zhang, Z., & Shen, F. (2021). CADN: A weakly supervised learning-based category-aware object detection network for surface defect detection. *Pattern Recognition*, *109*, Article 107571.

Zhang, Y., Wang, Q., & Hu, B. (2022). MinimalGAN: diverse medical image synthesis for data augmentation using minimal training data. *Applied Intelligence*, 1–18.

Zhang, H. D., Yuan, X., Li, D. Y., You, J., Liu, B., Zhao, X. M., Cai, W. M., & Ju, S. (2022). An effective framework using identification and image reconstruction algorithm for train component defect detection. *Applied Intelligence*, 1–19.

Zhao, B., Dai, M., Li, P., Xue, R., & Ma, X. (2020). Defect detection method for electric multiple units key components based on deep learning. *IEEE Access*, *8*, 136808–136818.

Zheng, Y., & Cui, L. (2022). Defect detection on new samples with siamese defect-aware attention network. *Applied Intelligence*, 1–16.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).