DeftectNet: Joint Loss Structured Deep Adversarial Network for Thermography Defect Detecting System

Lingfeng Ruan[1], Bin Gao[1*], Shichun Wu[1], Wai Lok Woo[2]

[1]School of Automation, University of Electronic Science and Technology of China, China

[2]Faculty of Engineering and Environment, Northumbria University, England, UK

*Corresponding author: bin_gao@uestc.edu.cn

Abstract: In this paper, a novel joint loss Generative Adversarial Networks (GAN) framework is proposed for thermography nondestructive testing named Defect-Detection Network (DeftectNet). A new joint loss function that incorporates both the modified GAN loss and penalty loss is proposed. The strategy enables the training process to be more stable and to significantly improve the detection rate. The obtained result shows that the proposed joint loss can better capture the salient features in order to improve the detection accuracy. In order to verify the effectiveness and robustness of the proposed method, experimental studies have been carried out for inner debond defects on both regular and irregular shaped carbon fiber reinforced polymer/plastic (CFRP) specimens. A comparison experiment has been undertaken to study the proposed method with other current state-of-the-art deep semantic segmentation algorithms. The promising results have been obtained where the performance of the proposed method can achieve end-to-end detection of defects.

Key words: Generative Adversarial Network, Loss Function, CFRP, Thermography Nondestructive testing

1. Introduction

In recent years, composite material of CFRP has been extensively used in the aerospace industry with its high strength and low weight. Defect detection of CFRP materials is particularly important and several conventional signal processing algorithms or machine learning algorithms are applied.

In [1], a machine learning method has been proposed by Ahmed *et al.* based on sparse dictionary matrix decomposition, which incorporate the low rank information into the sparse matrix and can extract weaker defects. In [2], Feng *et al.* proposed an automatic seeded region growing with thermographic signal reconstruction algorithms for CFRP defect detection. In [3], Liang *et al.* proposed an algorithm combining wavelet with principal component analysis for defects detection in CFPR. Although these methods have been proved effective in experiments, they are limited by resolution and suffered from the influence of noise.

Deep learning has a profound influence on various old and new fields, in which multiple high-dimensional nonlinear transformations are used to interpret data. Due to the excellent effects on image processing, deep learning algorithms have been applied to Infrared Non-Destructive Testing (IRNDT). Xu [4] *et al.* presented a method that mainly used the Stacked-VAE to denoise the original image for a clear defect location image. Yousefi *et al.* [5] uses a pre-trained VGG model as a feature extractor along with a spectral angler mapper to analyze defects. Olivier *et al.* [6] used the Convolutional Neural Networks (CNN) network to detect the defect of the composite materials. Hu *et al.* [7] uses the Fast-RCNN algorithm to detect the weld defect location and achieved a good result.

However, due to the limitation of the small-scale data and low semantic information as well as high noise interference in the infrared thermal images, it is difficult to train an existing classic network to detect inner defects with a large sufficient capacity. In particular, these methods limit their performance

for detecting weaker detects on a complex and irregular surface.

In order to deal with these limitations, we proposed a novel GAN model with joint loss to detect the defects by using Thermography Nondestructive testing system. The challenge of thermography defect detection is that the defects signal is mixed with background and noise due to irregular shape of sample, non-uniform emissivity and etc. The semantic information is not obvious and the conventional deep learning algorithms are difficult to detect the defects because of the complex structure of the sample. The proposed model can efficiently accomplish the defect detection and realize the defect semantic segmentation.

The contributions of this work are described in the following:

1) By leveraging the characteristics of the data, the proposed method improves the semantic segmentation network as it enhances the ability of feature extraction as well as suppressing the noise interference.

2) The GAN architecture is introduced to adapt with different specimens. The joint loss function incorporates the modified GAN loss and the penalty loss where the new structure makes the training procedure more stable and enables significant improvement on the detection rate.

3) In particular, the model can adapt to different data without parameter adjustments instead of training multiple networks for different samples. Different CFRP samples with various sizes of debond defects at different levels are used to validate the accuracy and robustness of the proposed algorithm.

The new structured GAN model makes the training procedure more stable and enables significant improvement on the detection rate. The comparable analysis has been undertaken with the state-of-the-art deep semantic segmentation algorithms. In addition, the result will be quantitatively validated by using events-based F-Score. Finally, the proposed method is an end-to-end detection system.

The remaining of the paper has been organized as follows: The Section 2 describes the work related to CNN and GAN. The details of the proposed model and the quantitative detectability assessment indicators are described in Section 3. Experiments and result analysis will be elucidated in the Section 4. Finally, Section 5 draws the conclusion of the work and highlights the future work.

2. Related work

CNN has been toned of the most widely used techniques for feature extraction and this has achieved unparalleled progress in the areas of image processing and computer vision. In the field of image semantic segmentation, the goal is to implement the global feature extraction on images to achieve pixel level classification. On the one hand, based on Fully Convolutional Networks (FCN) [9], some additional modules have been used to make the segmentation more precise in these networks [10]. On the other hand, Ronneberger *et al.* [13] proposed UNet based on the architecture of encoder-decoder with skip-connection. It has shown an excellent segmentation results for the small-scale data[15, 16]. Motivated by these ideas, many novel architectures [14] have been proposed for semantic segmentation to improve encoder and decoder performance.

With the rapid development of the deep learning, GAN is firstly proposed by Goodfellow *et al*. [19] to accomplish images generation that cannot be implemented by CNN. Since the training process of GAN is unstable and the quality of the generated image is not ideal, more variants of GAN have been proposed to improve the performance from the modification of the loss function. These GANs [21-23] improve the stability of the GAN by modifying the distance measurement of the original GAN. On the other hand, researchers have been working on the revision of the GAN internal network and the overall

framework. Radford *et al*. [20] put forward to DCGAN by using a CNN architecture, increasing the stability of the generation and the quality of the generated image. Hong *et al.* [27] proposed the Conditional GAN(CGAN) to realize structure domain adaption, making the GAN model better controlled. Shrivastava *et al.* [41] proposed a method of GAN training with additional losses to improve GAN performance for specific task. Relying on the changes of architecture and the improvement of generated image quality, an increasing number of applications of GAN has come out [24-26]. In particular, GAN can be used for the semantic segmentation task [28, 29].

GAN is usually used to generate samples and CNN can be used as the segmentation detection model. DeftectNet used CNN as the feature extractor and prediction network, while using the modified GAN model to enhance performance and adapt to different type of data. Compared to the existing deep learning segmentation models, the DeftectNet obtain more accurate in segmenting the background and defects.

3. Methodology

3.1. Framework of the proposed model

This section interprets the framework of the proposed model and describes the composition of the entire thermography nondestructive testing system. Fig 1 shows the framework of the proposed Joint Loss GAN model. The whole strategy can be summarized as the three parts: data preprocessing and augmentation (Part I), the entire training process (Part II), the prediction and evaluation (Part III).

a. Data Preprocessing and Augmentation

As shown in Fig 1-Part I, for the thermographic sequences $D \in \mathbb{R}^{M \times N \times F}$, where $(M \times N)$ is the size of the frame and F represents the number of frame of the sequences. The sequences D is downsampled as data $E \in \mathbb{R}^{M \times N \times S}$ of size S in which the procedure consists of heating and cooling stages. In order to apply the proposed network for image processing, these sequences need

to be converted into RGB images $x \in \square^{S \times M \times N \times 3}$. Due to limited dataset, the trained model is susceptible to overfitting. Previous study [30, 35] has shown that an appropriate data augmentation can enhance the recognition ability of the network. Therefore, for the infrared thermal images, it is important to use the data augmentation methodology to increase the data and therefore avoid overfitting. Thus, several data augmentation measures have been considered: shift or flip, random crop, color jittering, Principal Component Analysis (PCA) jittering and etc. In the experiment, shift, flip and random crop operation have been added to generate larger training examples. In addition, the whole images are resized into $(256 \times 256 \times 3)$ to fit the input requirements of the network and obtain the final dataset $L \in \square^{A \times 256 \times 256 \times 3}$ for training.

b. Proposed model for IRNDT

As shown in Fig 1-Part II, the joint loss GAN model consists of one generator and two discriminators. The $Dataset\_A$ means the regular sample dataset and the $Dataset\_B$ means the irregular sample dataset. For the regular and irregular shaped samples, the generator with one discriminator is hard to detect defects of the irregular samples. Therefore, two discriminators have been used to distinguish different types of samples. After preprocessing, the different type of sample dataset is sent to the generator for feature extraction and obtain the semantic segmentation images. The revised GAN loss and the penalty loss will be combined to train the whole network. To accomplish the detection task by way of semantic segmentation, the generator network has been structured based on the modified UNet framework. For the generator, the VGG-Net is conducted as an encoder as well as creating the block including Convolutional layer, RELU activation and dropout as the decoder. For the choice of the Encoder, the several traditional frameworks are used for experiments, like original UNet [13], VGG-Net [32] , ResNet [33] , DenseNet [34] and etc. According to the previous works [7, 36] and the results

from some in-house attempts, the VGG16-Net has been chosen as the encoder network. For the decoder, it has been found that it is beneficial to remove the RELU activation on the basis of the original design and use the Batch Normalization(BN) [39] for the infrared thermal images as shown in Fig 2-(b). Therefore, the final generator network is shown in Fig 2-(a). When using the BN layer, the effect of dropout module is not obvious [31]. When the dropout layer is out of use, it will get a slightly better network at last. Therefore, the dropout layer has been replaced by BN layer. The discriminator network takes the images and distinguish the source of images. The generator can benefit from the gradient through the adversarial loss. Traditional GAN discriminator is the pixel level classifier, mapping the images into a single scalar output. The Patch-GAN[24], however, uses the convolutional layer to map the images to many $N \times N$ patches to distinguish the source of patch. From previous work [24], the Patch-GAN discriminator has been introduced to adjust the generator. Finally, the trained generator can be used to predict defects as shown in Fig 1-Part III.

The flowchart of the proposed method is shown in Fig 2-(c). In the flowchart, D, L and P stand for dataset, label and perdition. A and B means the regular sample dataset and irregular sample dataset.

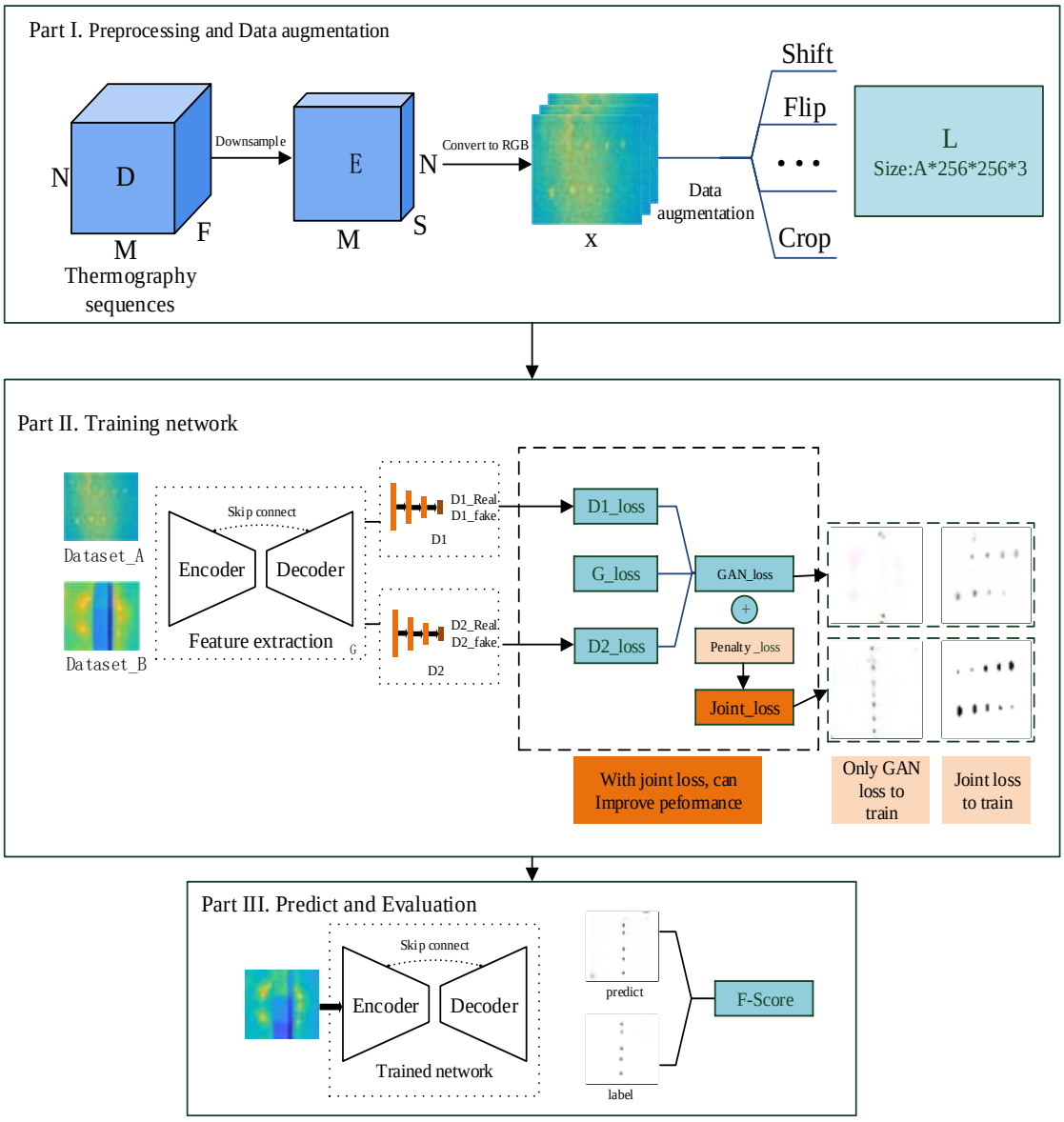
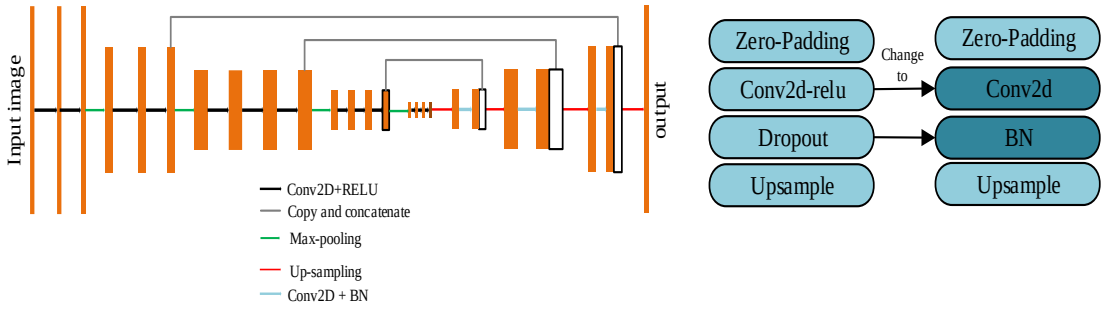
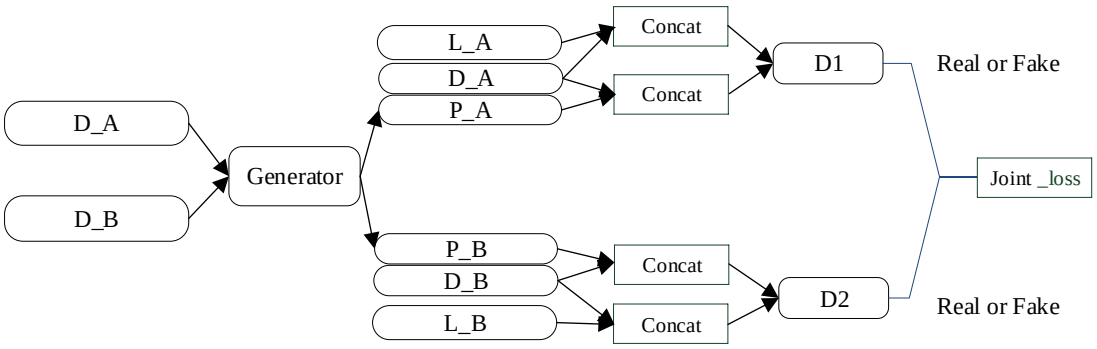


Fig 1. The framework of joint loss GAN model



(a) (b)



(c)

Fig 2. The algorithm framework details. (a) generator network. (b) the decoder designs. (c) the flowchart of the proposed method

3.2. Proposed Model

In the training stage, the joint loss is proposed in which contains the Adversarial loss function and the Penalty loss function.

Let the generator G to be the mapping of infrared thermal images x to defect label image y. The discriminator D is a classifier to discriminate the source of the image. z is defined as noise. Thus, the original object function of GAN can be formulated as:

$$L_{GAN}(D, G) = E_{x\sim p_{data(x)}}\left[log\ D(x,y)\right] + E_{z\sim p_{z(z)}}\left[log\left(1 - D(G(z))\right)\right] \tag{1}$$

The optimization problem is given, namely

$$G_{GAN}^{*} = \arg\min_{G}\max_{D} L_{GAN}(G, D) \tag{2}$$

Since the original GAN is unsupervised, it is not suitable to use it for defect detection. If the defect detection task needs to be done, the method requires the addition of the labeling of infrared thermal images. Therefore, the CGAN is introduce and the problem becomes

$$\min_G \max_D L_{CGAN}(D, G) = \min_G \max_D E_{x\sim p_{data(x)}}\left[log\ D(x\mid y)\right] + E_{z\sim p_{z(z)}}\left[log\left(1 - D(x, G(z\mid x))\right)\right] \quad (3)$$

In particular, the shape of the sample is different, which lead to the inconsistency of heat conduction. In the experiments, for the regular and irregular shaped samples, two discriminators have been used to distinguish different types of samples. One discriminator is responsible for training the regular shaped specimens, while another one train the irregular shaped specimens. During training, two types of images are sequentially fed into the network. The problem of the proposed model can be formulated as:

$$L_{GAN}(D, G) = \left\langle \begin{matrix} E_{x_1\sim p_{data(x_1)}}\left[log\ D(x_1\mid y_1)\right] + E_{x_2\sim p_{data(x_2)}}\left[log\ D(x_2\mid y_2)\right] \\ +E_{z\sim p_{z(z)}}\left[log\left(1 - D(x_{1,2}, G(z\mid x_{1,2}))\right)\right] \end{matrix} \right\rangle \quad (4)$$

where the $x_{1,2}$ means the regular shaped samples and the irregular shaped samples, and $y_{1,2}$ means the label of the $x_{1,2}$. In the experiment, it is found that using original GAN loss can detect defects roughly. However, its performances are not consistent due to the model collapse of the GAN training procedure or overfitting. Particular approaches [21] have been useful for solving these problems. In the proposed method, the log loss, which is cause of the instability of the model, has been removed as suggested in [21, 37].Thus, the final GAN loss function can be formulated as

$$L_{FGAN}(D, G) = E_{x_1\sim p_{data(x_1)}}\left[\ D(x_1\mid y_1)\right] - E_{x_2\sim p_{data(x_2)}}\left[\left(D(x_2\mid y_2)\right)\right] - E_{z\sim p_{z(z)}}\left[\left(1 - D(x_{1,2}, G(z\mid x_{1,2}))\right)\right] \quad (5)$$

In fact, it is not sufficiently satisfactory to rely solely on the GAN loss without applying additional penalty items to the detection of the defect. Thus, the extra loss function has been added to the specific task during training. Therefore, based on the original architecture, the penalty function has been added as part of the training process. For the choice of penalty function, the characteristics of the data should be considered. For the thermal images, the background information of the data is significant more than the defect information and it exits strong noise interference. We assume that defect information is more

abnormal during thermal diffusion after excitation. Therefore, the penalty function can be expressed as

$$L_p(G) = E_{x,y}[\| y - G(x) \|_p] \tag{6}$$

where the p value means the sensitivity of the model to outliers and it imposes constrains on the model. When the value of p takes 1 or 2, it is the common regularization in machine learning. In the experiment, it has been found that the combination of penalty function can make the experimental results more stable and the detection effect better. Therefore, the final penalty function can be formulated as:

$$L_{PEN}(G) = \lambda_1 L_{p1} + \lambda_2 L_{p2} \tag{7}$$

Therefore, by combining the GAN objective equation (5) with the penalty function (7), the final joint loss function is

$$G_{Final} = L_{FGAN}(G, D) + L_{PEN}(G) \tag{8}$$

Thus, it is much better than only a segmentation network that is applied to implement the detection of defect. The mapping from x to y can be learned and the deterministic output can be obtained by the generator without adding the noise z, which is exactly what the detection of defect requires. The discriminator guides the generator to generate the results which is similar to prior label images. In particular, it is beneficial to use the joint loss GAN to train a larger capacity network for different specimens.

3.3. Quantitative detectability assessment

Although the GAN is usually used as a model to generate new images, it is used here as a defect detect model. The proposed method actually classifies defects and non-defects in the form of semantic segmentation. The Intersection Over Union (IOU) has been used to evaluate the algorithms for almost

every semantic segmentation task. However, the IOU is not suitable for evaluating the defect detection task. Because the label reacts to the thermal diffusion of the defect after heating rather than the defect itself. Therefore, according to previous infrared Non-destructive Testing work [8], the F-Score is used to estimate the detection ability of the proposed algorithms. The events based on F-Score is expressed as

$$F = \left(\beta^2 + 1\right)\frac{Precision \times Recall}{\left(\beta^2 \times Precision\right) + Recall} \tag{9}$$

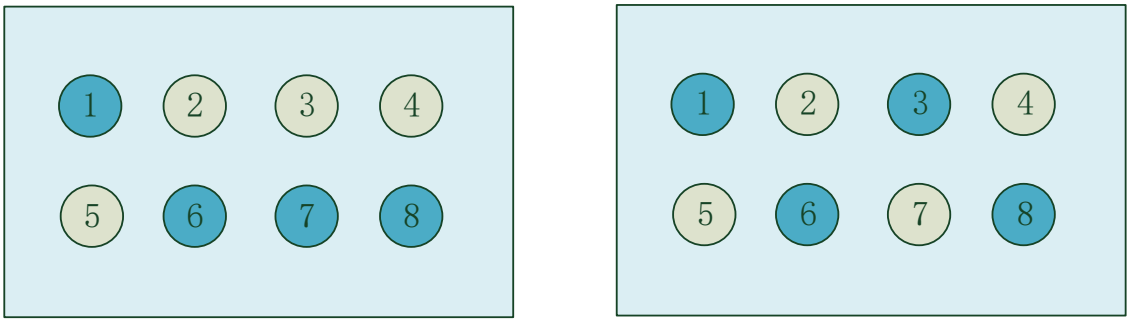
The Precision and Recall are formulated as:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

where the TP is true positive, which means that the defect is existed and is detected; FP is false positive, meaning no defect exists but is detected; FN is false negative, which denotes a defect exists but is not detected; TN is true negative, which denotes no defect exists and none is detected. The β is the weight of the Precision and Recall. For the thermal imaging debond diagnosis, the value of β is set to 2, which is mean that Recall is more important than Precision.

In order to interpret the F-Score, an example will be set up in Fig 3. As is shown in Fig 3-(a), the actual defect area is the 1,6,7,8 and the predicted defect area is 1,3,6,8. According to the definition, the result of TP, FP, FN and TN are 3,1,1 and 3, respectively. And by equation (10-11) we can calculate the Precision and Recall, both of which are 3/4. So, the F-Score is 3/4 depending on the equation (9).



(a) (b)

Fig 3. F-Score objectively evaluates the performance of the algorithms. (a) actual defect area. (b) predicted defect area.

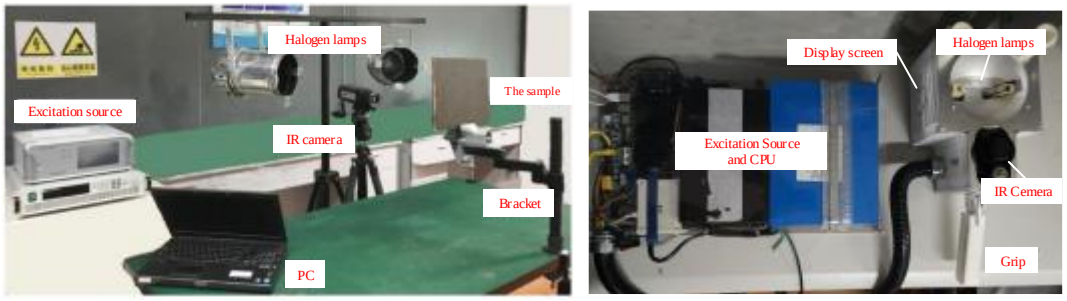
4. Experiment and result analysis

*4.1. Ex*periment setup and Sample preparation

The experiment will be carried out in high-precision Optical Pulse Thermography (OPT) system and in Portable OPT(POPT) system. As shown in Fig 4-(a), the first OPT system has higher precision than portable devices. In the experiment, IR camera(A655sc) is used to collect thermal image sequences and the thermal sensitivity is 0.05℃. The 480×640 size is used in thermal images. The halogen lamps with a power of 2kW is applied as an excitation source and the excitation time can be controlled by the ZY-B type excitation source with a maximum power of 3kW. The Bracket can fix the test sample. As shown in Fig 4-(b), the POPT system includes an integrated computer processer and power control systems. The halogen lamps (ST-PS04) is the excitation source with a power of 800W. The IR camera(MAG62) is used to collect data with thermal sensitivity of 2℃,which can produce 480×640 size thermal images. During the testing, the sample will be excited and the heating and cooling process will be recorded when the control bottom on the grip is pressed.

In addition, several different samples are used to evaluate the effectiveness and robustness of the proposed algorithms. The information of these samples can be found in Table 1. For the former two

samples, they are CFRP sample with flat shape and they have sub-surface debond defects, namely Flat sample. The third sample is a curve surface sample with a rectangular area in the middle that is not defect but affects the prediction results. For the sample 3 to 6, they are CFRP sample with curved shape, named R-area sample. It is difficult to detect the defect of R-area materials because the defect is at elbow. For the sample 7 to 9, these are the images obtained by our Portable OPT system. The sample 7 is a piece of flat sample that is punched in the back. The sample 8 to 9 is the R-area, but it is more difficult to detect due to the lower sensitivity of the IR camera.



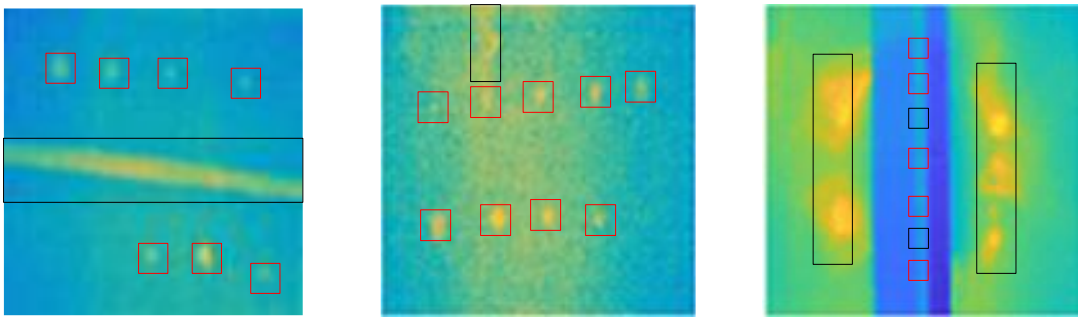
(a) (b)

Fig 4. the experiment systems (a) OPT system (b) POPT system

The proposed method is implemented based on the Keras library with Tensorflow backend and runs in NVIDIA 1080Ti. The experiment will be carried out on two different datasets from the two OPT systems, where the dataset has different types of samples. There are 400 images in the first dataset, which can reach 2000 images after do data augmentation. The second dataset only provide 200 images and it can reach 1000 images after data augmentation. During training, the Adam optimizer is set up with learning rate of 0.0001 and set the exponential decay rate of first moment estimation $\beta_1 = 0.5$. The batch size is set in 4 and the iteration is set at 400. The p_1 and p_2 in the function (7) is set 1&3 in the final test. And the function coefficient λ_1 / λ_2 is set to 4/5.

The positive and negative sample in non-destructive testing coexist in different areas on the same

image. The difficulty of the defect detection task is that the image contains defect, noise and background. Several samples have been selected for illustration. As shown in Fig 5. The red box represents the defect location and the black box represents the noise. For the flat sample, the noise information is consistent with the defect information in temperature changes. As shown in Fig 5-(b) the noise and defect information are mixed. For the R-zone sample, the defect information is weak and its difference from non-defective information is small. As shown in Fig 5-(c), the position of black box information is similar to the defect information, which is difficult to be distinguished.



(a) (b) (c)

Fig 5. Negative and Positive sample of the specimens (a) Flat sample A (b) Flat sample B (c) R-zone sample

4.2. Result and analysis

In this section, in order to evaluate the proposed algorithm, four common semantic segmentation deep learning algorithms were selected for comparison. These methods consist of UNet [13], TernausNet [38], ResNet-UNet [40] and Dense-UNet [14]. The same training set will be used to train each network. The effectiveness and robustness of the proposal will be proved form different perspective of different materials and datasets. In addition, the segmentation result of each network for testing set and the final comparative quantitative results are given in Table 2 and Table 3, respectively.

Table 2 shows the visual results of above algorithms. Firstly, for the sample 1 to 3, when the defect information of the sample is more obvious, these deep learning methods show excellent defect detection

ability. However, it is obvious that the comparison methods have a considerable shortcoming. These methods results are affected by the non-defect area in sample 3, whereas the proposed model and ResNet-UNet can be done without false detection. Compared with ResNet-UNet and Dense-UNet, the original UNet and the TernausNet are more sensitive to noise because of higher false detection rates in the result of sample 1 to 3. However, ResNet-UNet and Dense-UNet cannot detect the R-area data very well because of being too insensitive to defect information. In addition, the proposed model can ensure correct detection while effectively preventing noise interference.

In terms of R-area sample, it still exists challenge to detect the defects because the background information drowns the defect information and the defects is in the elbow of the sample. The result of sample 4 to 6 shows that these networks have failed detection for this type of specimens. For the model of ResNet and DenseUNet, although it introduces the connection of the previous layer to enhance the interaction of the semantic information, they are failure to detect defects in the R-area sample. On the other hand, the TernausNet, a modified UNet with VGG16 as encoder, is a slightly effective one in these networks. This is why the proposed model choose VGG network as the encoder. Therefore, it is not advisable to only consider semantic information purely, but more importantly to extract low semantic information under the high noise interference. In the proposed method, besides of the better segmentation performance for the Flat samples, it also shows excellent detection capability to the R-area sample.

To prove the robustness of the proposed, sample 7 to 9 are used to have further validation. For the sample 7, the semantic information is clear and the defects are more obvious. Compared with the proposed method, the fail detection of the comparison methods is more serious. For the sample 8 to 9, limited by the accuracy of IR camera, the detection results of all methods are disappointed, while the

proposed method is far superior to comparison method. For the regular and irregular shaped specimens, the overall performance of the proposed method is significantly better than all the other methods.

Table 3 shows the Precision, Recall and F-score of all the visual result. The Pr means Precision and the Re means Recall. Although the F-score of comparison on sample 1 to 3 is already high, it is exceeded by the proposed method. In particular, for the R-area shaped sample, the F-score of ResNet-UNet and Dense-UNet is 0%, which means that these algorithms fail to detect a defect. While the proposed can reach 100%, 76.92% and 95.24%. For the sample 8, the F-score of proposed method is only 50%, as there are three comparison methods giving 0% and the UNet only giving 17.24%. On average the comparison methods give the 51.12%, 49.77%, 30.08% and 31.70% defect detection capability. The proposed method gives the highest capability on average that is 84.25%. Therefore, the proposed method is better than other methods in terms of detection ability.

In the experiment, CFRP materials are divided into regular sample and irregular sample. The shape and structure of the regular sample and the irregular sample are different, and therefore they will lead to inconsistent thermal diffusion. This inconsistent thermal diffusion means that the data distribution is quite different. It can be shown from the comparison algorithms that a single model is difficult to detect all defects. Thus, we use the adversarial feature of GAN to train the modified generator separately with different discriminator to adapt to different type of samples.

4.3 The ablation study experiment

According to the algorithm structure, we designed the ablation experiment. Firstly, it cancels the GAN structure and only retain the semantic segmentation network. Secondly, it cancels the description of different discriminator and only retains one interpretation to train generator. Thirdly, it studies the

impact of loss structure.

For the first part, the generator network is compared separately with the proposed GAN model. The result is shown in Fig 6. For the flat sample, the generator network can effectively detect defect. However, for the R-zone sample, the generator network fails to detect defect. The shape and structure of the flat sample and the R-zone sample are different whereas these lead to inconsistent thermal diffusion. This inconsistent thermal diffusion means that the data distribution is quite different. A single network is difficult to detect defects in different specimens. For the proposed method, due to the introduction of GAN, it is compatible to detect defects in both flat sample and R-zone sample. Therefore, the architecture of GAN is validated to beneficial to the detection result.

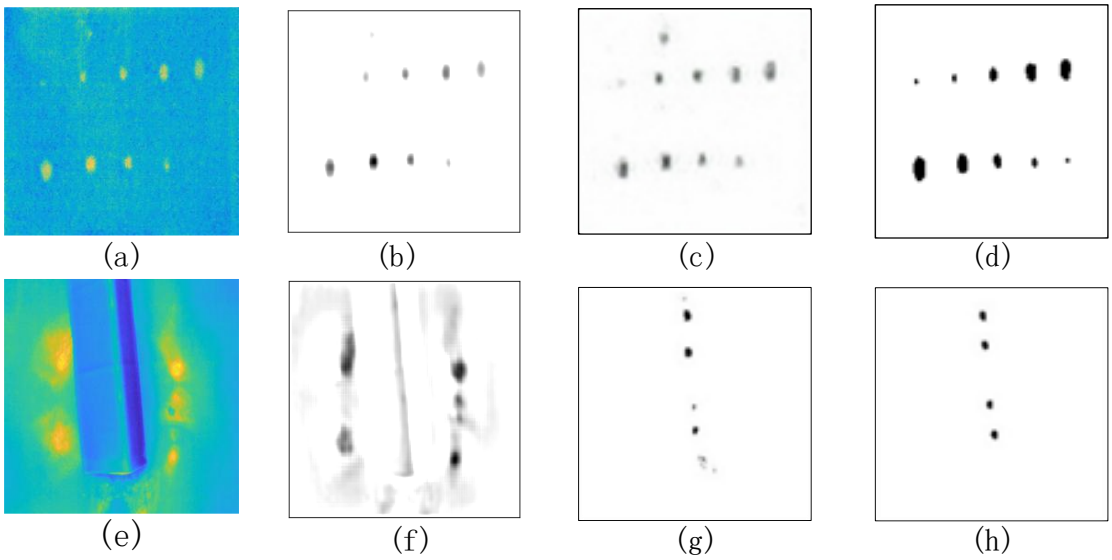


Fig 6. Segmentation ablation experiment (a)Raw image of flat sample. (b)Segmentation result. (c)Proposed. (d)Label of the flat sample. (e)Raw image of R-zone sample. (f) Segmentation result. (g)Proposed. (h)Label of R-zone

For the second part, the result is shown in Fig 7. When one discriminator is constructed to train the generator, the network has missed detection in the R-zone sample. As shown in the result, the network detects most detects of the flat sample. However, there exists missed detections on the R-zone sample.

This shows that such one discriminator model can be compatible with different specimens. However, in industrial applications, missed detection indicates potential safety issues. Therefore, the proposed two discriminators have been used to guide the training of generator to improve the detection rate. Thus, the missed detection rate is reduced and the single detection network can better be compatible with different specimens.

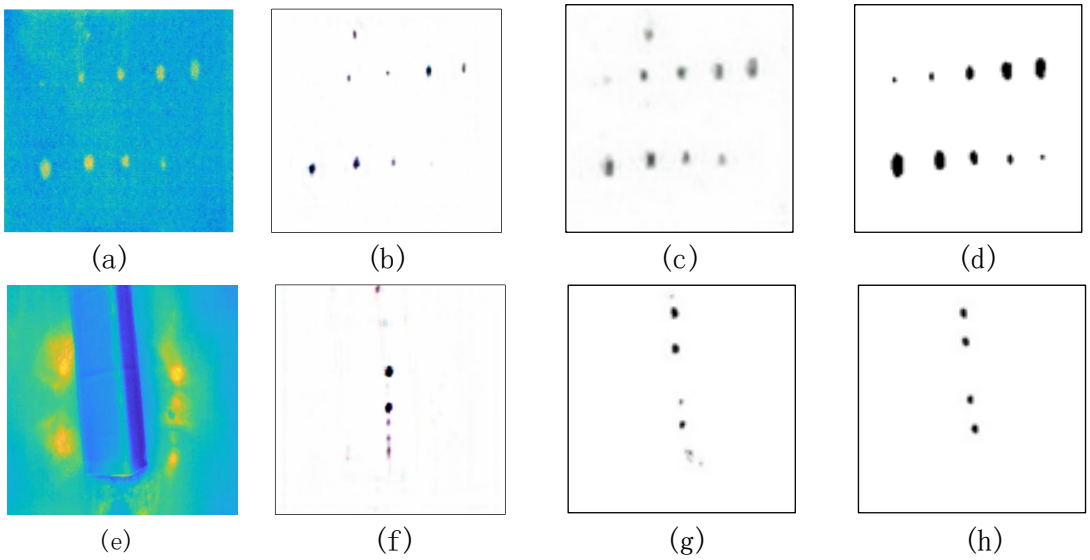
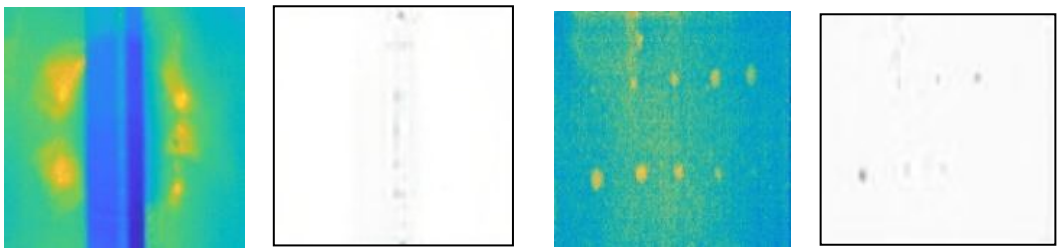


Fig 7. Discriminator ablation experiment (a)Raw image of flat sample. (b)One discriminator result. (c)Proposed. (d)Label of the flat sample. (e)Raw image of R-zone sample. (f)One discriminator result. (g)Proposed. (h)Label of R-zone.

For the third part, the penalty loss ablation experiment has been conducted. In this work, the analysis will focus on the impact of the loss function on the results of the analysis. The main hyper-parameter of penalty loss is the $p_{i(i=1,2)}$ from equation (7). Through the in-house experiment, the p_1 was fixed with 1 and the p_2 varied from 0 to 4. In this experiment, different regularization items are selected to modify the result of the network. Two different samples are used to test the influence of various combination. Table 4 shows the visual results of the experiment and Table 5 shows the F-score. As shown in Table 4, when the L1-norm is used, namely $p_1 = 1 \& p_2 = 0$, the network almost impossible

to detect defects. In fact, only the loss of GAN can be obtained the result without back-end binarization. However, we found using Monte-Carlo analysis that the predicted defect value is greater than 0 and less than the threshold value of 0.5. The main purpose of the algorithm is to build an end-to-end defect detection system. The threshold setting of binarization is not adjusted for each sample. Under the same threshold condition, when p1 and p2 equal to zero, the test results of these specimens are considered as failed prediction. Therefore, the term "Can't detect" has been used. The visual result is shown in Fig 8.



(a) (b) (c) (d)

Fig 8. Using GAN loss to test(a)Sample 1; (b)Sample1 prediction; (c)Sample 2; (d)Sample 2 prediction

Therefore, the combined penalty loss needs to be investigated. When setting $p_1 = 1 \& p_2 = 2$, it can be found that the network trend to find the defect features. It means that the increased sensitivity to outliers through penalty function contributes on features capturing during the adversarial training. However, when setting $p_1 = 1 \& p_2 = 4$, due to the excessive amplification of the outliers, the detection accuracy of the network for obvious defect is reduced, which makes it difficult to detect the R-area defects. When setting $p_1 = 1 \& p_2 = 3$, the result shows that the defects are basically detected. As shown in Table 5, the F-score is highest when setting $p_1 = 1 \& p_2 = 3$. Therefore, in the final predict network, the $p_1 = 1 \& p_2 = 3$ is chosen at the end. Thus, by adding the penalty loss function, the ability of defect feature extraction has been promisingly improved.

Table 1. Sample information

Number	Specimen	Dimension(mm)	Defect Diameter(mm)	Images

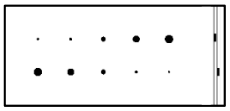
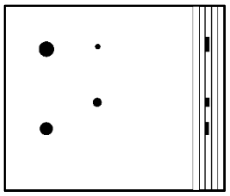
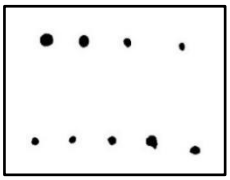
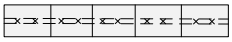
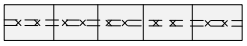
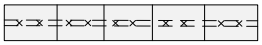
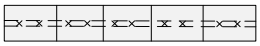
1		150×300×2	Depth: 1 or 1.2 Diameter: 3,5,7,10,12	
2		150×150×2	Depth:1.2 Diameter: 3,5,7,10	
3		250×150×2	Depth:1.2 Diameter: 3,5,7,10,12	
4		100×100×8	Depth:1 to 1.5 Diameter: 9,10	
5		100×100×8	Depth:1 to 1.5 Diameter: 9,10	
6		100×100×8	Depth:1 to 1.5 Diameter: 6,8	
7		530×180×5.5	Depth: 0.5 to 5 Diameter: 10,20,5	
8		100×100×8	Depth:1 to 1.5 Diameter: 9,10	
9		100×100×8	Depth:1 to 1.5 Diameter: 6,8	

Table 2. Visual result of comparison

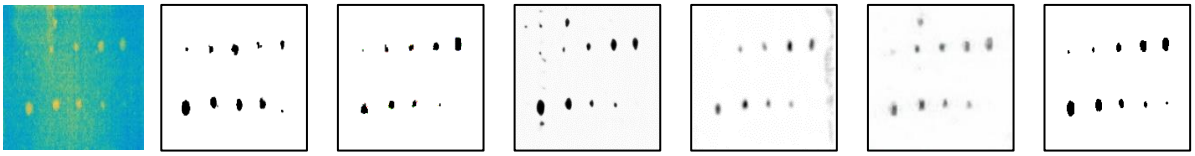
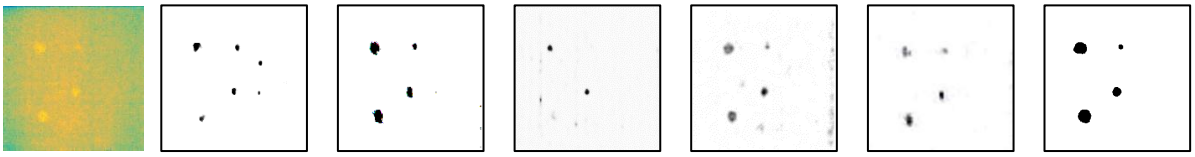
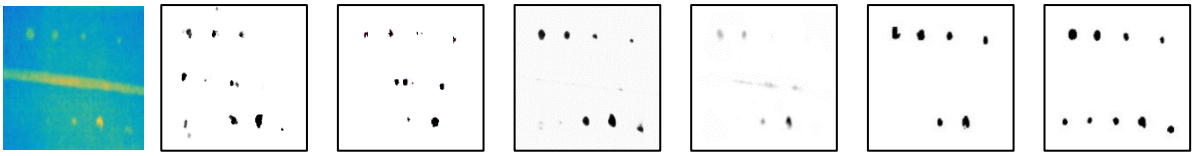
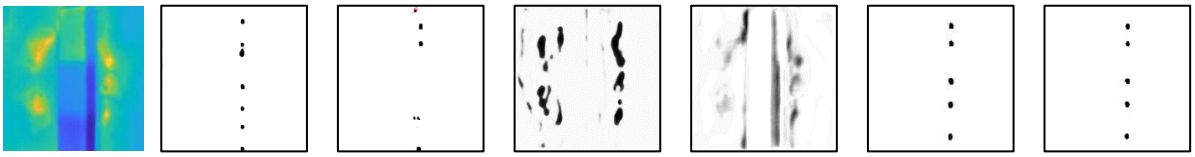
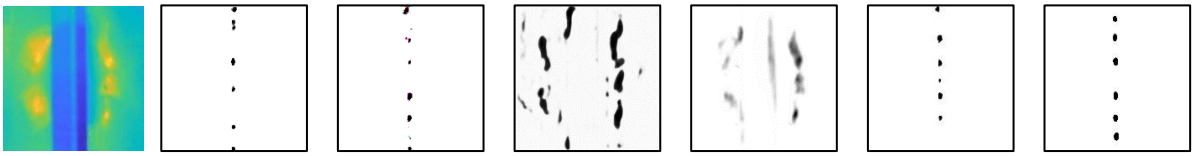
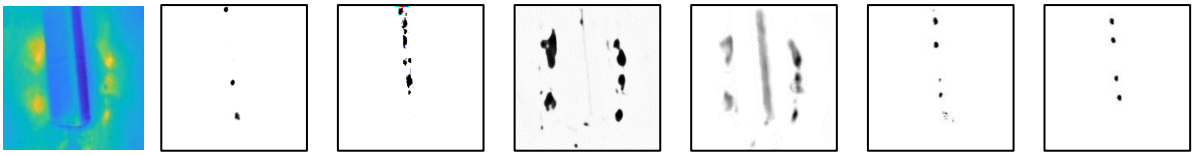
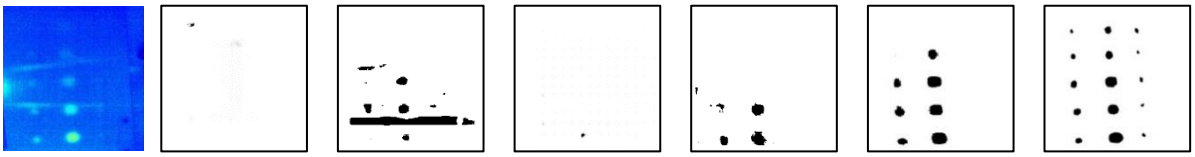
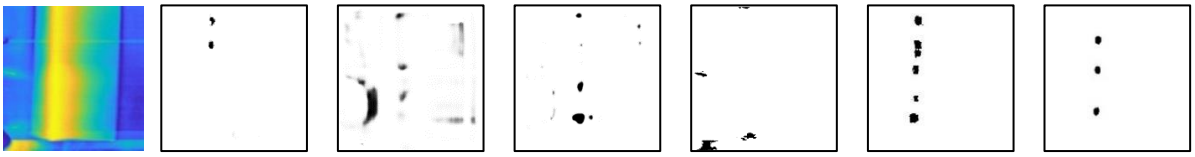
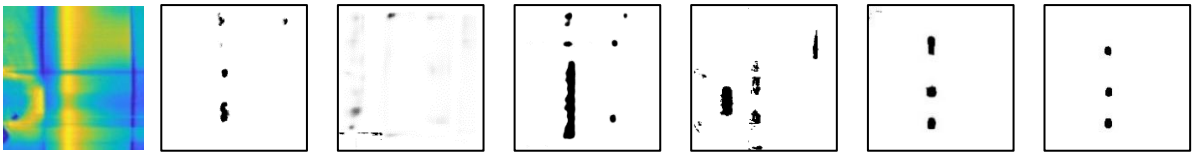
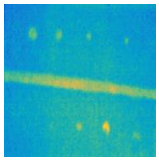
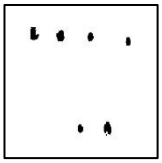
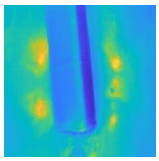
	Original image	UNet	TernausNet	ResNet-UNet	Dense-UNet	Proposed	Label
1							
2							
3							
4							
5							
6							
7							
8							
9							

Table 3. Precision, Recall and F-score of comparison result

Sample		UNet	Ternausnet	ResNet-UNet	DenseNet-UNet	Proposed
	Pr:	1.00	1.00	0.69	1.00	0.90
1	Re:	1.00	0.90	0.90	0.80	0.90
	F-socre	**100%**	91.84%	84.91%	83.33%	90.00%
	Pr:	0.67	1.00	0.50	0.80	1.00
2	Re:	1.00	1.00	1.00	1.00	1.00
	F-socre	90.91%	**100.00%**	83.33%	95.24%	**100.00%**
	Pr:	0.67	0.50	1.00	1.00	1.00
3	Re:	0.67	0.33	0.78	0.44	0.78
	F-socre	66.67%	35.71%	81.40%	50.00%	**81.40%**
	Pr:	0.44	0.70	0.00	0.00	1.00
4	Re:	0.67	0.58	0.00	0.00	1.00
	F-socre	60.61%	60.34%	0.00%	0.00%	**100.00%**
	Pr:	0.375	0.67	0.00	0.00	0.86
5	Re:	0.55	0.67	0.00	0.00	1.00
	F-socre	50.00%	66.67%	0%	0%	**76.92%**
	Pr:	0.33	0.33	0.00	0.00	0.80
6	Re:	0.25	0.75	0.00	0.00	1.00
	F-socre	26.32%	60.00%	0.00%	0%	**95.24%**
	Pr:	0.00	0.33	1.00	0.50	0.96
7	Re:	0.00	0.33	0.05	0.21	0.78
	F-socre	0.00%	33.33%	5.43%	23.43%	**81.40%**
	Pr:	0.50	0.00	0.00	0.00	0.50
8	Re:	0.15	0.00	0.00	0.00	0.50
	F-socre	17.24%	0.00%	0.00%	0.00%	**50.00%**
	Pr:	0.43	0.00	0.125	0.33	0.50
9	Re:	0.50	0.00	0.17	0.33	1.00
	F-socre	48.39%	0%	15.63%	33.33%	**83.33%**
	Pr:	0.49	0.50	0.37	0.40	0.72
Average	Re:	0.53	0.51	0.32	0.31	0.88
	F-socre	51.17%	49.77%	30.08%	31.70%	**84.25%**

Table 4. Visual result for different $p_{i(i=1,2)}$

Sample	$p_1 = 0 \& p_2 = 0$	$p_1 = 1 \& p_2 = 0$	$p_1 = 1 \& p_2 = 2$	$p_1 = 1 \& p_2 = 3$	$p_1 = 1 \& p_2 = 4$
	Can't detect				



Can' detect



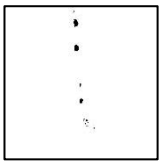



Table 5. F-score of hyper-parameter adjustment result

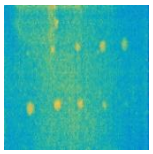
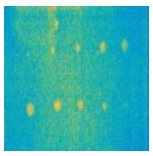
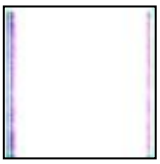
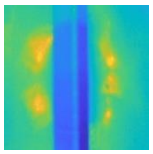
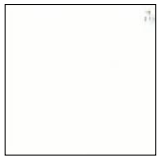
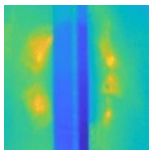
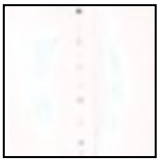
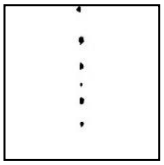
Sample	$p_1 = 0 \& p_2 = 0$	$p_1 = 1 \& p_2 = 0$	$p_1 = 1 \& p_2 = 2$	$p_1 = 1 \& p_2 = 3$	$p_1 = 1 \& p_2 = 4$
1	0%	34.48%	48.78%	**81.40%**	12.5%
2	0%	27.78%	48.78%	**95.24%**	0%

4.4 The impaction discussion of GAN

This section will explore the impact of the different GANs on the result. The first part is the stability of the GAN experiment. The second part will explore the detection effect of different GAN-based models.

For one thing, regarding the stability of GAN, we have carried out objective evaluation on the experimental visual results. The experiments have conducted comparison in the loss of the original GAN with that of proposed where the predicted result of the test sample in different epochs are considered. When using the original GAN loss under the training process, the predicted value is too small. In order to better display the results, the output without binarization is shown. The result is tabulated in Table 6. From the table, when using original GAN, the result will worsen as the numbers of iteration increase. The intensities of the predicted defects become progressively lighter and weaker, that is, the pixel intensity value is deviating farther from label value. Such a situation is easy to occur but it is difficult to avoid through adjusting the parameters. This is the model collapse due to the drawbacks of original GAN loss. However, the detection effect of the proposed method gradually gets better when the number of iteration increases. Therefore, the proposed method improves the stability of the training.

Table 6. Different GAN loss stability experiment

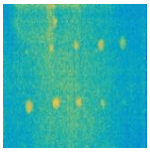
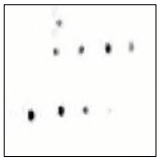
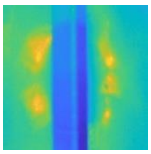
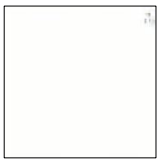
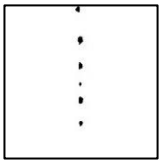
	Sample	Epoch=25	Epoch=100	Epoch=150	Epoch=200
GAN					
proposed					
GAN					
proposed					

In addition, conventional GAN is predominantly used to generate new sample data rather than a model for detect detection. However, the proposed method is designed to accommodate data of different types of specimens and to improve the detection rate through joint-loss. GAN-based semantic segmentation method [29] usually use generator as a segmentation network or use image conversion or image transformation. Therefore, GAN has been used as an image transformation model [24] to perform semantic segmentation with different GAN loss.

In order to better reflect the generalization, the segmentation model of W-GAN[21] and Ls-GAN [22] are used for comparison. Both WGAN and Ls-GAN make GAN training more stable and efficient by modifying the distance measurement of GAN. In Table 7, when using the original GAN loss, we found that the predicted defect value is greater than 0 and less than the threshold value of 0.5. Therefore, the term “Can’t Detect” has been marked. When the objective functions of WGAN and Ls-GAN are used to

train the network, due to the relatively stable training, several acceptable results for the flat sample have been attained. However, for the R zone specimens with small defect targets and large background noise, the model is unable to detect the defects. Therefore, none of the three GAN methods can perform with good results in a single network.

Table 7 The GAN-segmentation experiment result

Samples	GAN	WGAN	Ls-GAN	Proposal
	Can't Detect			
	Can't Detect			

4.5 The public dataset experiment

In order to verify the generalization ability of the proposed method, we have added comparison works using public dataset. The Airbus ship semantic segmentation dataset [43] has been added for analysis. This is a dataset with unbalanced positive and negative samples, which shares similarity with the non-destructive testing for infrared thermal image. The entire dataset contains more than 190000 images with a size of 768×768. In order to compare the effectiveness of the algorithms, the UNet, UNet++[42] and DeeplabV3+[11] are chosen as the comparison algorithms. The Recall takes more important role in the task. Since the label of test set cannot be obtained, one-tenth of the training set is randomly obtained as the test sets. All algorithms are controlled under the same condition, in both training and testing procedure. Part of the representative visual result is shown in Fig 9. According to the visual

result, except for the DeeplabV3+, the remaining algorithms can effectively segment the ship and background.

The Table 8 shows the average quantitative detectability assessment for all test images. As shown in Table 8, the DeeplabV3+ get the highest average recall value. However, its precision value performs the lowest in which case it results in a low F2-score. From the visual results, the recall rate is due to a large number of false detections. UNet++ has the better performance, with higher recall and precision. Although the proposed algorithm is not good as UNet++ in precision and F2-score, it is better than UNet++ in recall value. The experiment results validate that the proposed method is effective in public dataset without any additional tricks.

In particular, the proposed method is specially developed for the field of non-destructive testing. For the better comparison, the UNet, UNet++ and DeeplabV3+ have been trained and tested on the thermography dataset. The experimental visual results are shown in Fig 10. UNet++ builds up more semantic context connections while DeeplabV3+ provides more receptive field to extract features. However, these two algorithms perform poorly result in detecting R-zone defects. This is due to the structure of the specimens which have significant influence to the data distribution. Based on the data characteristics, the proposed build up a network with sufficient feature extraction capabilities. The architecture of joint-loss GAN enables a single network to effectively detect different types of specimens. Compared with the label, the proposed method performs superior to the UNet, UNet++ and DeeplabV3+.

The Table 9 shows the quantitative detectability assessment of the infrared thermal image. The test results of DeeplabV3+ in the R-zone are failed to segment defect and background. Thus, the relevant evaluation value of R-zone result is not calculated and we use INVALID to refers to this situation in Table 9. From the visual result and quantitative detectability assessment, the performance of the proposed

method is found to be superior to other comparison algorithms.

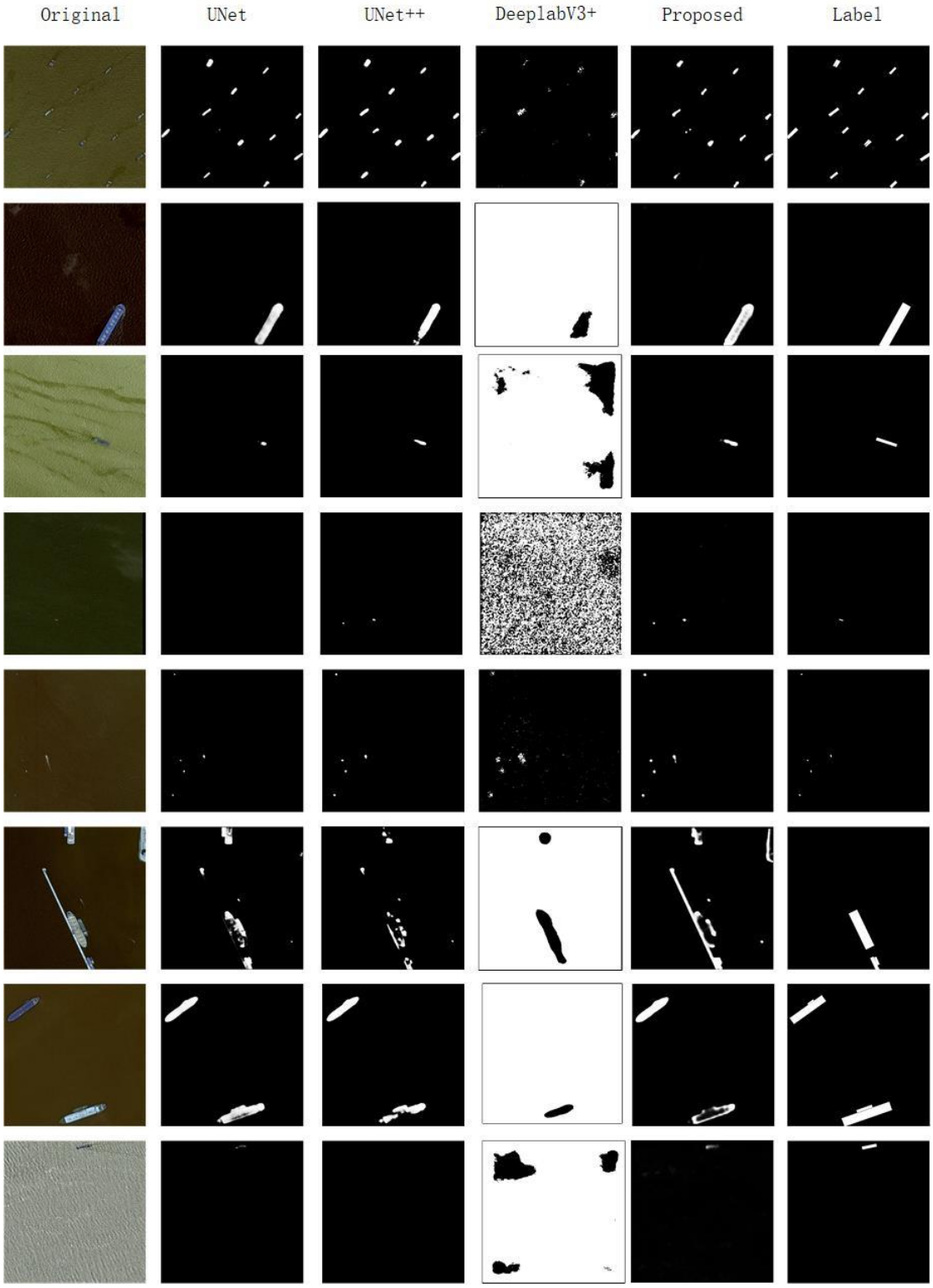


Fig 9. Visual result of the dataset

Table 8. Average Precision, Recall and F2-score

		UNet	UNet++(ResNet34)	DeeplabV3+	Proposed
Average	Precision:	0.40	**0.54**	0.11	0.39

Recall:	0.46	0.50	**0.64**	0.54
F2-score:	0.36	**0.46**	0.01	0.39

UNet ++ and other algorithms are algorithms with strong generalization ability for many types of data. However, such algorithms perform poorly in the thermography dataset. Based on the characteristics of data, the proposed method is specially designed for thermography dataset. The proposed method can also be effectively applied to public dataset and is superior to UNet and DeeplabV3+. The above proves the effectiveness and generalization ability of the algorithm

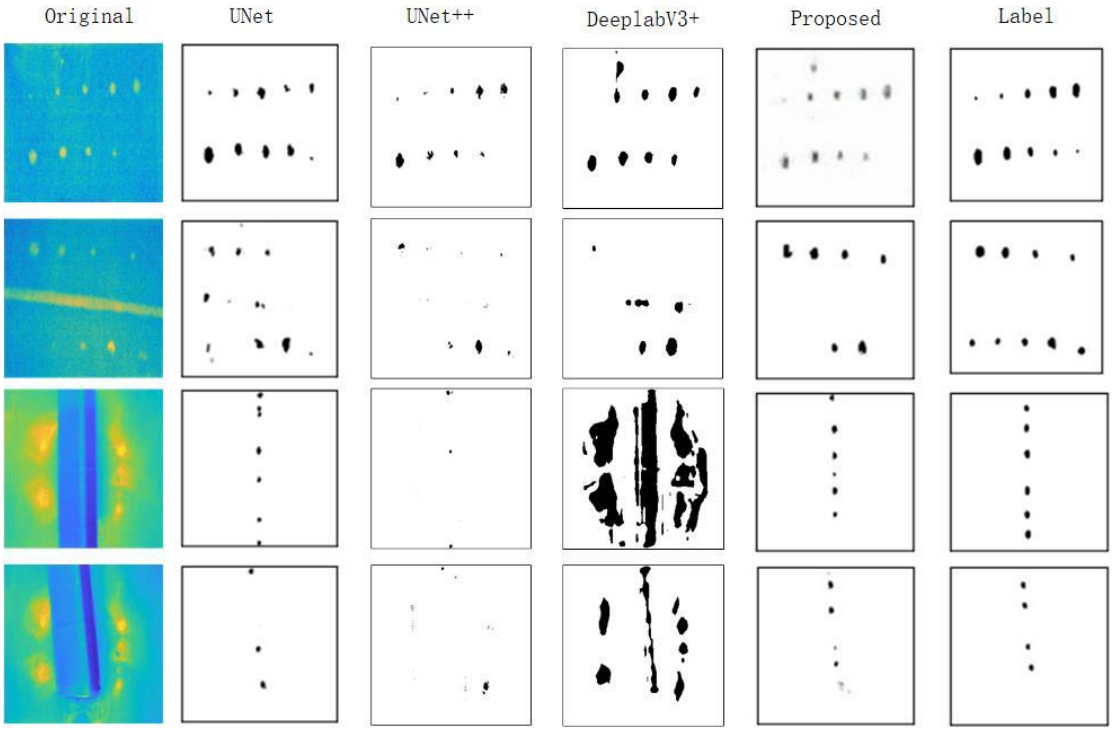


Fig 10. Test result of infrared thermal image

Table 9. Precision, Recall and F2-score

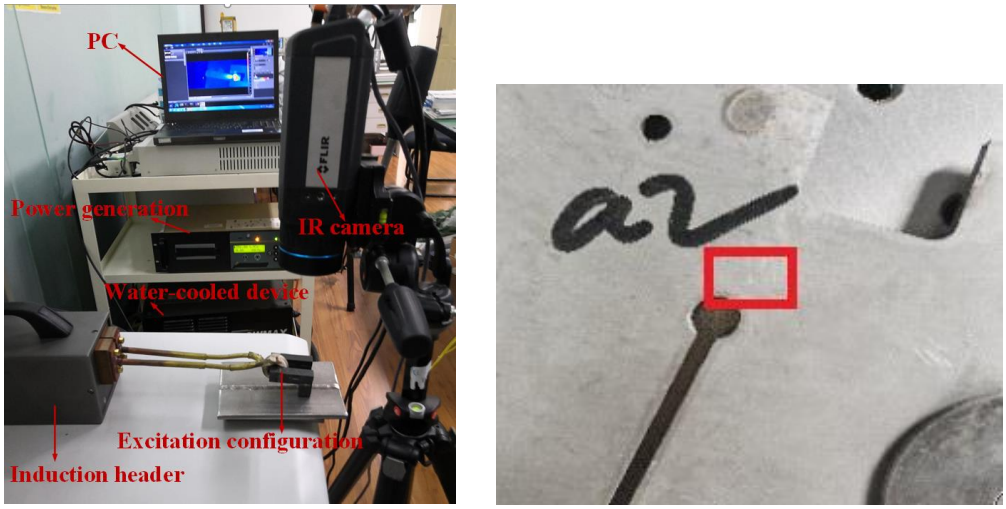
Sample		UNet	UNet++	DeeplabV3+	Proposed
	Pr:	1.00	1.00	0.89	0.90
1	Re:	1.00	0.90	0.80	0.90
	F2-score	**100%**	91.83%	65.32%	90.00%
	Pr:	0.67	0.86	0.33	1.00
2	Re:	0.67	0.43	0.43	0.78
	F2-score	66.67%	71.42%	40.54%	**81.40%**
	Pr:	0.375	0.33	INVALID	0.86
3	Re:	0.55	0.17	INVALID	1.00
	F2-score	50.00%	18.52%	INVALID	**76.92%**
4	Pr:	0.33	0.00	INVALID	0.80

	Re:	0.25	0.00	INVALID	1.00
	F2-score	26.32%	0.00	INVALID	**95.24%**
	Pr:	0.59	0.55	INVALID	0.89
Average	Re:	0.62	0.38	INVALID	0.92
	F2-score	61.00%	45.44%	INVALID	**85.89%**

4.6 The eddy current pulsed thermography experiment testing

The proposed method is an end-to-end algorithm for optical pulsed thermography, which is specially designed for the optical pulsed thermographic (OPT) with carbon fiber reinforced polymer/plastic (CFRP). However, in order to better verify the performance of the algorithm, the eddy current pulsed thermography (ECPT) data has been used for further validation. Defect detection of ECPT data is more challenging as crack signal (which represents the defect) is very weak. Compared with the CFRP defect detection task, ECPT is mainly used for ferromagnetic materials and non-ferromagnetic materials. The surface of the specimens is more complex while the noise and background information are more prominent.

The configuration of ECPT system is shown in Fig 11-(a). With 0.08k of temperature resolution and 200Hz of maximum frame rate, the FLIR infrared camera is positioned normal to the surface of the conductive material. In the experiment, the frequency of the excitation current was fixed at 200 kHz. 350A is used to adjust the current to provide enough energy, and the heating time is 200ms. In addition, one example of test crack sample is shown in Fig 11-(b). As can be seen that, the crack is quite small with irregular shape as this brings detection challenge for the algorithms. In overall, seven different test samples have been employed for validation.



(a) (b)

Fig 11 The ECPT experiment (a) The configuration of ECPT system (b) Test sample with natural crack (red marked area)

All algorithms are controlled under the same condition, in both training and testing procedure. Part of the representative visual result is shown in Fig 12. The first four images show the artificial scratch defects, and the last three show the natural defects. When using UNet and Dense-UNet to train and test, these algorithms cannot detect defects. We used Monte-Carlo analysis that the predicted defect value is greater than 0 and far less than the threshold value of 0.5. Due to the characteristics of the data, it is shown that such networks are required to balance feature fusion and avoid information fitting. In particular, it is observed that UNet's performance is not better than TernausNet, which shows that the importance of the feature extraction ability. For ResNetUNet and Dense-UNet, both methods focus on contextual information fusion. However, excessive contextual information fusion of the Dense-UNet directly leads to detection failure. For the proposed method, although the performance is not as good as the OPT experiment, the precision, recall and F-score are superior to others in comparison.

The ECPT experiments have been conducted and shown that the proposed method is effective in

defect detection task. Compared with the different algorithms, it has achieved better performance in all the three metrics of precision, recall and F2-score as shown in Table 10.

Table 10. Average Precision, Recall and F2-score

		UNet	TernausNet	ResNetUNet	Dense-UNet	Proposed
Average	Precision:	-	0.12	0.49	-	**0.55**
	Recall:	-	0.05	0.24	-	**0.43**
	F2-score:	-	0.06	0.26	-	**0.44**

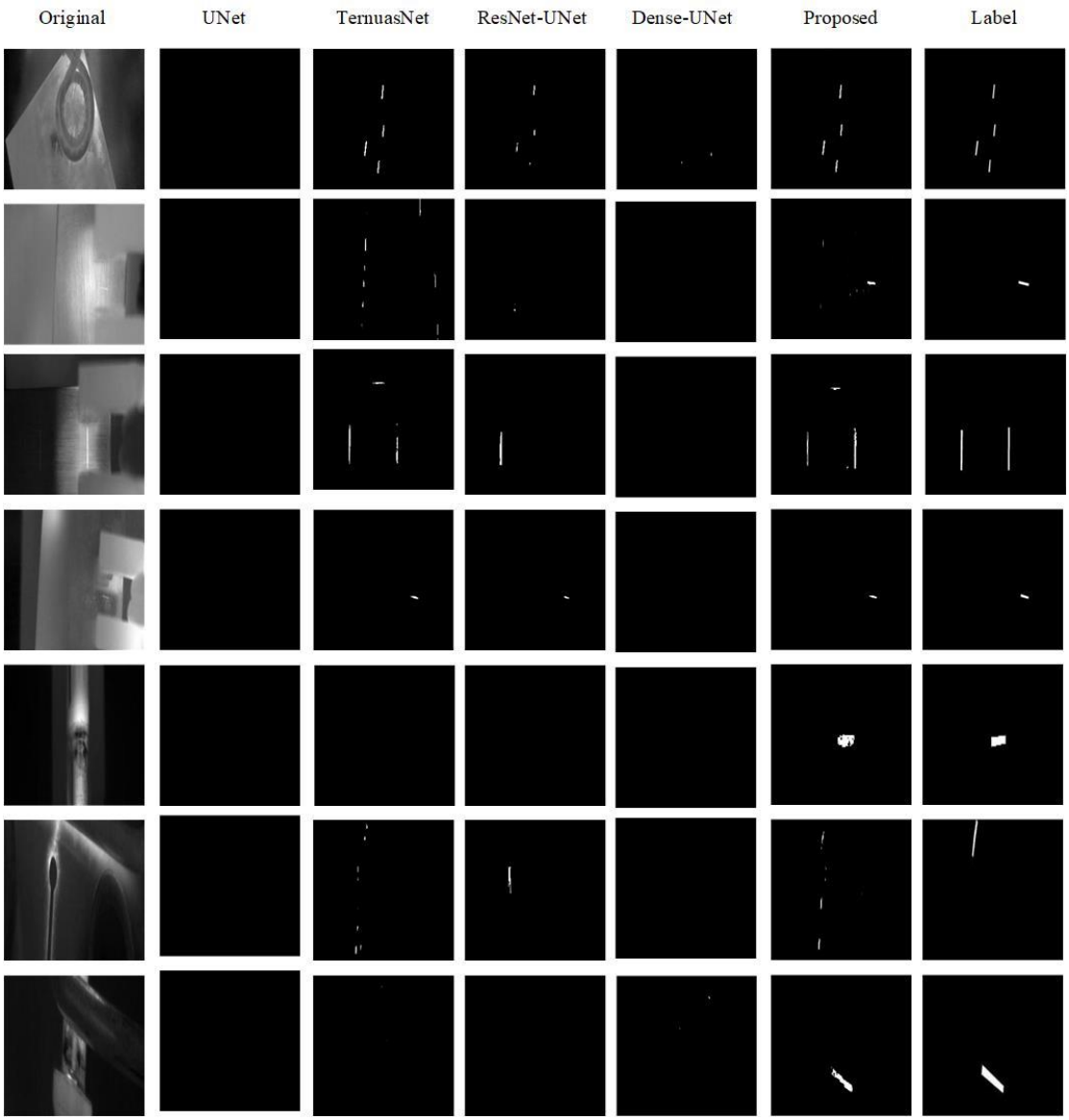


Fig 12 Part of the ECPT experiment visual result

5. Conclusion and Future Work

In this paper, the detection of defects based on the modified GAN algorithms has proposed for the infrared thermal images. Compared with the common deep semantic segmentation CNN model, the modified GAN architecture has been used to improve the detection rate and increase the capacity of the model. In addition, with the participation of the joint loss, GAN training can enhance the performance of CNN while ensuring stability.

The F-score in the proposed has been reached 84.25% on average, which exceeds other comparison methods. The obtained results have shown that the proposed model has attained a high level of defect detection performance. The robustness of the algorithm is proved by the data from different OPT system and the data of different shapes. Future work will focus on high noise infrared thermal images defect detection or separation of background and defects.

Acknowledgement

The work was supported by Science and Technology Department of Sichuan, China (Grant No.2019YJ0208 and 2018JY0655).

Reference:

[1] J. Ahmed, B. Gao, W.L. Woo, Wavelet-Integrated alternating sparse dictionary matrix decomposition in thermal imaging CFRP defect detection, IEEE Trans. Ind. Informatics. (2019). https://doi.org/10.1109/TII.2018.2881341.

[2] Q. Feng, B. Gao, P. Lu, W.L. Woo, Y. Yang, Y. Fan, X. Qiu, L. Gu, Automatic seeded region growing for thermography debonding detection of CFRP, NDT E Int. (2018). https://doi.org/10.1016/j.ndteint.2018.06.001.

[3] T. Liang, W. Ren, G.Y. Tian, M. Elradi, Y. Gao, Low energy impact damage detection in CFRP using eddy current pulsed thermography, Compos. Struct. (2016). https://doi.org/10.1016/j.compstruct.2016.02.039.

[4] C. Xu, J. Xie, C. Wu, L. Gao, G. Chen, G. Song, Enhancing the visibility of delamination during pulsed thermography of carbon fiber-reinforced plates using a stacked autoencoder, Sensors (Switzerland). (2018). https://doi.org/10.3390/s18092809.

[5] B. Yousefi, d. Kalhor, r. Usamentiaga, l. Lei, c. Ibarra-castanedo, x. Maldague, Application of deep learning in infrared non-destructive testing, J. Qirt 2018 proceedings. (2018) https://doi.org/10.21611/qirt.2018.p27.

[6] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, S. Van Hoecke, Convolutional neural network-based fault detection for rotating machinery, J. Sound Vib. (2016). https://doi.org/10.1016/j.jsv.2016.05.027.

[7] J. Hu, W. Xu, B. Gao, G. Tian, Y. Wang, Y. Wu, Y. Yin, J. Chen, Pattern deep region learning for crack detection in thermography diagnosis system, Metals (Basel). (2018). https://doi:10.3390/met8080612.

[8] Y. Wang, B. Gao, W.L. Woo, G. Tian, X. Maldague, L. Zheng, Z. Guo, Y. Zhu, Thermal pattern contrast diagnostic of microcracks with induction thermography for aircraft braking components, IEEE Trans. Ind. Informatics. (2018). https://doi.org/10.1109/TII.2018.2802046.

[9] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2015. https://doi.org/10.1109/CVPR.2015.7298965.

[10] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. (2018). https://doi.org/10.1109/TPAMI.2017.2699184.

[11] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2018. https://doi.org/10.1007/978-3-030-01234-2_49.

[12] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proc. - 30th IEEE Conf.

Comput. Vis. Pattern Recognition, CVPR 2017, 2017. https://doi.org/10.1109/CVPR.2017.195.

[13] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

[14] X. Li, H. Chen, X. Qi, Q. Dou, C.W. Fu, P.A. Heng, H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes, IEEE Trans. Med. Imaging. (2018). https://doi.org/10.1109/TMI.2018.2845918.

[15] H. Chen, Y.F. Li, D. Su, M3Net: Multi-scale multi-path multi-modal fusion network and example application to RGB-D salient object detection, in: IEEE Int. Conf. Intell. Robot. Syst., 2017. https://doi.org/10.1109/IROS.2017.8206370.

[16] W. Xie, J.A. Noble, A. Zisserman, Microscopy cell counting and detection with fully convolutional regression networks, Comput. Methods Biomech. Biomed. Eng. Imaging Vis. (2018). https://doi.org/10.1080/21681163.2016.1149104.

[17] L. Zhou, X. Kong, C. Gong, F. Zhang, X. Zhang, FC-RCCN: Fully convolutional residual continuous CRF network for semantic segmentation, Pattern Recognit. Lett. (2018). https://doi:10.1016/j.patrec.2018.08.030.

[18] R. Zhang, W. Yang, Z. Peng, P. Wei, X. Wang, L. Lin, Progressively diffused networks for semantic visual parsing, Pattern Recognition. (2019). https://doi.org/10.1016/j.patcog.2019.01.011.

[19] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Adv. Neural Inf. Process. Syst., NIPS 2014, 2014.

[20] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional GANs, Int. Conf. Learn. Represent. (2016). https://doi.org/10.1051/0004-6361/201527329.

[21] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, Clin. Toxicol. (2017).

http://doi: 10.2507/daaam.scibook.2010.27.

[22] X. Mao, Q. Li, H. Xie, R.Y.K. Lau, Z. Wang, S.P. Smolley, Least squares generative adversarial networks, in: Proc. IEEE Int. Conf. Comput. Vis., ICCV 2017, 2017. https://doi.org/10.1109/ICCV.2017.304.

[23] S. Nowozin, B. Cseke, R. Tomioka, f-GAN: Training generative neural samplers using variational divergence minimization, in: Adv. Neural Inf. Process. Syst., NIPS 2016, 2016.

[24] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, 2017. https://doi.org/10.1109/CVPR.2017.632.

[25] L. Zhao, H. Bai, J. Liang, B. Zeng, A. Wang, Y. Zhao, Simultaneous color-depth super-resolution with conditional generative adversarial networks, Pattern Recognition. (2019). https://doi.org/10.1016/j.patcog.2018.11.028.

[26] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, Neurocomputing. (2018). https://doi.org/10.1016/j.neucom.2018.09.013.

[27] W. Hong, Z. Wang, M. Yang, J. Yuan, Conditional Generative Adversarial Network for Structured Domain Adaptation, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition., CVPR 2018, 2018. https://doi.org/10.1109/CVPR.2018.00145.

[28] W.C. Hung, Y.H. Tsai, Y.T. Liou, Y.Y. Lin, M.H. Yang, Adversarial learning for semi-supervised semantic segmentation, in: Br. Mach. Vis. Conf. 2018, BMVC 2018, 2019.

[29] N. Souly, C. Spampinato, M. Shah, Semi supervised semantic segmentation using generative adversarial network, in: Proc. IEEE Int. Conf. Comput. Vis., ICCV 2017, 2017. https://doi.org/10.1109/ICCV.2017.606.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, SSD: Single shot multibox

detector, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2016. https://doi.org/10.1007/978-3-319-46448-0_2.

[31] X. Li, S. Chen, X. Hu, J. Yang, Understanding the disharmony between dropout and batch normalization by variance shift, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition., CVPR 2019, 2019. https://doi.org/10.1109/CVPR.2019.00279.

[32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 2015.

[33] S. Wu, S. Zhong, Y. Liu, Deep residual learning for image steganalysis, Multimed. Tools Appl. (2017). https://doi.org/10.1007/s11042-017-4440-4.

[34] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, 2017. https://doi.org/10.1109/CVPR.2017.243.

[35] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM. (2017). https://doi.org/10.1145/3065386.

[36] Liu L, Zhou Y. A closer look at U-net for road detection[C]//Tenth International Conference on Digital Image Processing (ICDIP 2018). International Society for Optics and Photonics, 2018, 10806: 108061I.

[37] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of Wasserstein GANs montreal institute for learning algorithms, Adv. Neural Inf. Process, NIPS 2017, Syst. 2017.

[38] V. Iglovikov, S. Seferbekov, A. Buslaev, A. Shvets, TernausNetV2: Fully convolutional network for instance segmentation, in: IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2018, 2018. https://doi.org/10.1109/CVPRW.2018.00042.

[39] S. Ioffe, Christian Szegedy, Batch normalization: Accelerating deep network training by reducing, J. Mol.

Struct. (2015). https://doi.org/10.1016/j.molstruc.2016.12.061.

[40] M.P. Heinrich, M. Stille, T.M. Buzug, Residual U-Net convolutional neural network architecture for low-dose CT denoising, Curr. Dir. Biomed. Eng. (2018). https://doi.org/10.1515/cdbme-2018-0072.

[41] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, Learning from simulated and unsupervised images through adversarial training, in: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, 2017. https://doi.org/10.1109/CVPR.2017.241.

[42] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging. (2020). https://doi.org/10.1109/TMI.2019.2959609.

[43] [dataset] Airbus ship detection dataset, 2018. https://www.kaggle.com/c/airbus-ship-detection/data