Computers in Industry 151 (2023) 103978

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.sciencedirect.com/journal/computers-in-industry

A generalized well neural network for surface defect segmentation in Optical Communication Devices via Template-Testing comparison

Tongzhi Niu [a], Zhiyu Xie [a], Jie Zhang [a], Lixin Tang [a,*], Bin Li [a,b], Hao Wang [c]

[a] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, 430074, Hubei, China
[b] Wuhan Intelligent Equipment Industrial Institute Co., Ltd., 8 Ligou South Road, Wuhan, 430074, Hubei, China
[c] Department of Mechanical Engineering, College of Design and Engineering, National University of Singapore, Singapore 117575, Singapore

ARTICLE INFO

Keywords:
Flexible manufacturing system
Surface defect segmentation
Template-Testing comparison
Attention mechanism
Siamese networks

ABSTRACT

Surface defect detection is an important task in the field of manufacturing, and dealing with imbalanced data is a challenge that has been addressed using methods such as anomaly detection and data augmentation. However, optical devices pose a particular challenge due to their characteristics of small batches and varying types, resulting in insufficient positive sample data and difficulty in predicting the data distribution of new batches. To address this issue, we propose a neural network that learns to compare the differences between templates and testing samples, rather than directly learning the representations of the samples. By collecting templates, the model can generalize to new batches. The challenge of extracting defect features by comparison is to remove background noise, such as displacements, deformations, and texture changes. We propose a Dual-Attention Mechanism (DAM) in the stage of feature extraction, which extracts the noise-free defect features using the non-position information of self-attention. In the stage of feature fusion, we introduce a Recurrent Residual Attention Mechanism (RRAM) to generate spatial masks that shield noise and enable multi-scale feature fusion. We evaluate our method on three datasets of Optical Communication Devices (OCDs), Printed Circuit Boards (PCBs) and Motor Commutator Surface Defects (MCSD), and demonstrate that it outperforms existing state-of-the-art methods. Our work provides a promising direction for addressing the challenge of surface defect detection in OCDs and can be generalized to other flexible manufacturing system (FMS).

1. Introduction

Optical Communication Devices (OCDs) are devices that convert optical and electrical signals in Gigabit Passive Optical Networks and Optical Network Terminals. They consist of a base, pins, and various Surface Mounted Devices (SMD) components connected by jump wires. Manual inspection of OCDs is costly and inefficient. The advent of deep learning and computer vision has led to the widespread adoption of automatic optical inspection (Božič et al., 2021; Gao et al., 2022; Zhuxi et al., 2022; Luo et al., 2023). This technology has been shown to effectively improve the quality and efficiency of OCDs production. Deep learning has shown remarkable success in various fields, but its performance is heavily dependent on the availability of large and balance dataset. However, many industrial settings often face constraints with limited and imbalanced datasets. To tackle this issue, recent research has proposed solutions such as anomaly detection (Bergmann et al., 2021; Tao et al., 2022; Roth et al., 2022b) and image generation (Yun et al., 2020; Niu et al., 2021; Ren et al., 2022). But, as a classical flexible manufacturing system, OCDs are characterized by small batch sizes and a diversity of designs, resulting in significant variations in component type, placement and quantities from batch to batch, as illustrate in Fig. 1 (a).

Therefore, the surface defect detection of OCDs faces several challenges, including: (1) The wide variability of samples from different batches, making it difficult for a single model to be effective for all batches. (2) Limited sample availability for new batches, making it challenging to train multiple models for different batches. Data augmentation techniques are applied to improve the distribution of data, but the distribution of new batch remains unpredictable. Anomaly detection methods rely on numerous defect-free samples for their success, but this can be difficult to collecting. Therefore, it is imperative to develop a generalized network capable of handling small batches with diverse designs in flexible production lines.

Recently, the concept of one-shot learning (OSL) has been introduced as a solution to generalize to new tasks with limited samples by leveraging prior knowledge. In the defect classification task, Dong et al.

[*] Corresponding author.
E-mail addresses: tzniu@hust.edu.cn (T. Niu), m202070437@hust.edu.cn (Z. Xie), 2434350181@qq.com (J. Zhang), lixintang@hust.edu.cn (L. Tang), libin999@hust.edu.cn (B. Li), mpewhao@nus.edu.sg (H. Wang).

https://doi.org/10.1016/j.compind.2023.103978
Received 17 March 2023; Received in revised form 9 June 2023; Accepted 12 June 2023
0166-3615/

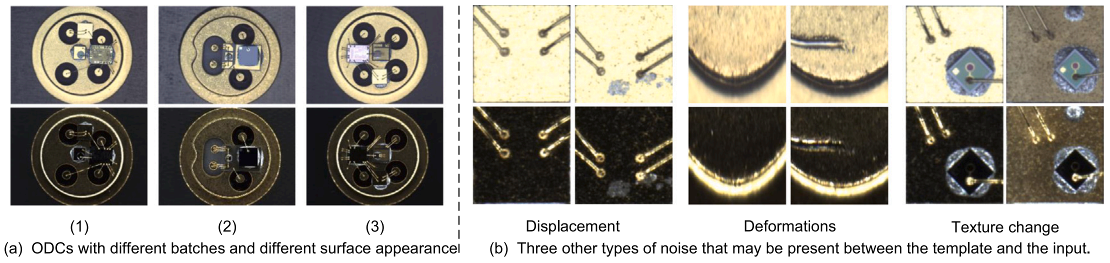

(1)
(2)
(3)
(a) ODCs with different batches and different surface appearance
Displacement
Deformations
Texture change
(b) Three other types of noise that may be present between the template and the input.

Fig. 1. The Optical Communication Devices (OCDs) images. As shown in figure (a), OCDs are significant variations in component type, placement, and quantity across different batches. As depicted in figure (b), in addition to defect features, the difference between templates and input samples contains three types of noise.

(2021) propose a few-shot pavement distress detection method based on metric learning, which can effectively learn new categories from a few labeled samples. The prototypical network of few-shot learning algorithm introduced by Zhan et al. (2022) follows an N-way K-shot paradigm to forces the number of samples within each class to be uniformly distributed. In the defect segmentation task, Bao et al. (2021) introduce a triplet-graph reasoning network to achieve few-shot metal generic surface defect segmentation. To address the limited generalization in scenes with distribution differences, Ma et al. (2023) propose a one-shot unsupervised domain adaption framework. Although these methods achieve impressive performance in metal surface defect segmentation, they are not suitable for OCDs samples. Compared to metal surfaces, the difference between batches of OCDs surfaces is not only manifested in texture and color, but also in structure, component type, location, and other factors.

In OSL, a naive idea is to not directly learn the characteristics of the samples, but to learn how to compare the differences between templates and testing samples, so that the model can generalize to new tasks through the collecting of templates. Lu et al. (2020) present the Co-attention siamese network (COSNet) to tackle the zero-shot video object segmentation task by exploiting the inherent correlation among video frames in a comprehensive manner. Kwon et al. (2019) propose a Siamese U-Net with healthy template to segment the abnormal regions of intracranial hemorrhage more accurately from patients' CT images. In surface defect detection, Ling et al. (2022) developed a Siamese Semantic Segmentation Network (DSSSnet) that combines similarity measurement with an encoder–decoder network for detecting welding defects in Printed Circuit Boards (PCBs).

However, in addition to defect features, the difference between templates and input samples contains some noise. As shown in Fig. 1(b), the noise can be summarized as three points: (1) displacement, including translation and rotation; (2) deformations, such as arcs of jump wires and components, shape of solder, etc; (3) texture change. As we know, it is easy for neural networks to remove the noise of texture change. However, removing displacement and deformation noise poses a challenge, given that convolutional operations are translation equivariant.

To solve the above problems, we propose a novel generalized well network (GWNet) via Template-Testing comparison for surface defects segmentation in optical communication devices, as illustrate in Fig. 2. During the training phase, a paired set consisting of templates, samples, and labels is constructed to train the model. This enables the model to acquire the ability to perform comparisons. In the apply phase, only the templates are acquired and new batches can be detected by comparison. The network is divided into two parts: feature extraction and fusion.

In the stage of feature extraction, the key point is to eliminate noise and obtain defect features. Self-attention (Wang et al., 2018) is commonly used to model dependencies between different positions in a sequence, such as words in a sentence or pixels in an image. And it does not depend on the specific position or order of the input features, and can thus be more robust to displacement and deformation noise. Therefore, we propose a dual-attention mechanism (DAM) that utilizes the non-positional information of self-attention. The self-attention feature maps of test samples and the cross-attention maps of templates and test samples are used to calculate the difference to obtain noise-free defect features.

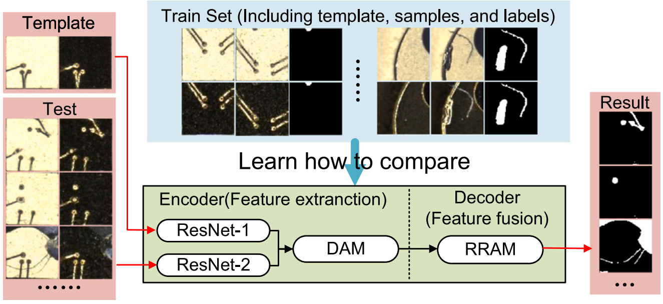


Fig. 2. A novel generalized well network via Template-Testing comparison for surface defects segmentation.

In the stage of feature fusion, to fuse features of multi-scales, we propose recurrent residual attention mechanism, which generalize masks to shield the noise in shallow feature maps. The core concept involves using the noise-free defect feature obtained from DAM in a deep-to-shallow manner, to acquire multi-scale masks. Specifically, we leverage the spatial attention mechanism (Wang et al., 2017) to generate denoising masks and draw inspiration from recurrent neural networks (Mnih et al., 2014) to fully utilize multi-scale feature information of the context.

In summary, our contributions mainly include the following points:

(1) We propose a novel generalized well neural network via Template-Testing comparison for surface defect segmentation, which emphasizes the comparison of differences between templates and test samples, enabling the model to generalize effectively to new batches of unseen data in training by utilizing collected templates.

(2) In order to remove displacement and deformation noise, we introduce a new dual-attention mechanism that utilizes the non-positional information of self-attention to extract noise-free defect features.

(3) We design a recurrent residual attention mechanism to achieve multi-scale feature fusion, which can obtain multi-scale denoising masks in a deep-to-shallow manner.

(4) Finally, we evaluate GWNet on three datasets, which include Optical Devices (OCDs), Printed Circuit Boards (PCBs) and Motor Commutator Surface Defects (MCSD). Our results demonstrate that GWNet outperforms existing state-of-the-art methods across these datasets.

2. Related works

2.1. Surface defect inspection

The advancement of automatic optical inspection has led to an increased focus on deep learning-based surface defect detection methods. However, collecting labeled defect data can be time-consuming and labor-intensive, leading researchers to focus on small sample. In this

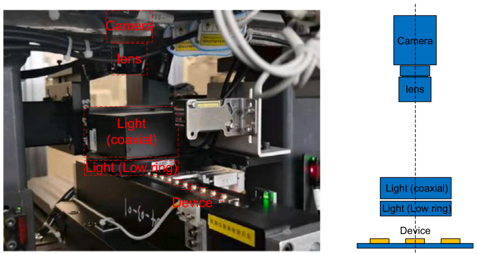


Fig. 3. Schematic and physical diagram of the image acquisition system.

paper, we categorized methods suitable for small datasets into data augmentation and unsupervised methods.

Data augmentation can be utilized to enhance the diversity of samples by transforming original data and incorporating prior knowledge. Niu (Niu et al., 2021) proposed a method Region-and strength-controllable GAN. Ren (Ren et al., 2022) not only regulate the characteristics of generated images, but also reduce distribution differences between training and test sets. The proposed vision-based defect inspection system (Yun et al., 2020) by Yun utilizes a data generation algorithm based on the conditional variational auto-encoder technology. While data augmentation methods are effective, they cannot predicate the distribution of new batches.

In industrial quality inspection, while ample data on the desired product appearance is available during training, significantly fewer defective samples are available. To address the aforementioned challenges, unsupervised anomaly detection (Tao et al., 2022) using only positive samples has been extensively studied. The MVTec datasets, comprehensive real-world anomaly detection datasets (Bergmann et al., 2021), are introduced to train and evaluate new approaches and ideas. Recent works Roth et al. (2022b) combining embeddings from models pre-trained on large external natural image datasets with a memory bank of nominal patch-features have demonstrated state-of-the-art detection performance. However, collecting enriched positive samples for new batches can also be difficult.

In conclusion, the research in this paper focuses on the generalization problem of the model, which is trained on old batch samples and needs to adapt to new batches. Both data augmentation and unsupervised methods are not suitable for this scenario.

2.2. One-shot-learning

To train a model with limited examples but achieve generalization to unfamiliar data without extensive retraining, one-shot learning (OSL) is proposed (Wang et al., 2020). Existing OSL methods can be categorized into three groups: data-based, algorithm-based, and model-based.

Data-based OSL methods leverage prior knowledge to augment the train dataset, thereby enriching the supervised information. These methods can be categorized based on the source of data augmentation, including: transforming samples from the training dataset (Liu et al., 2018), utilizing weakly labeled or unlabeled datasets (Douze et al., 2018), and incorporating similar datasets (Gao et al., 2018). Though data-based method is a good approach to solve small data problem, the distribution of current dataset in OCDs areas skewed with respect to other batches, the augmented data distribution is skewed as well.

Algorithm-based OSL methods alter search strategy in hypothesis space by prior knowledge. According to how the search strategy is affected by prior knowledge, we classify methods into refining existing parameters (Roth et al., 2022b), refining meta-learned parameters (Finn et al., 2017) and learning the optimizer (Andrychowicz et al., 2016). Specifically, the refinement of existing parameters involves loading pre-training parameters and fine-tunes them through techniques such as early-stopping and elective updates. Moreover, the parameters are further adapted to the task-specific data through refinement by a meta-learner. In contrast, instead of relying on gradient descent, the last method trains an optimizer to directly output updates.

Model-based OSL constrain hypothesis space by prior knowledge. In terms of what prior knowledge is used, the model-based OSL can be classified into multitask learning (Zhang and Yang, 2022), embedding learning (Vinyals et al., 2016; Koch et al., 2015), learning with external memory (Gong et al., 2019) and generative modeling (Niu et al., 2021). Above methods are efficient under certain conditions. For an example, When there exist similar tasks or auxiliary tasks, multitask learning can be used to constrain the hypothesis. In this paper, OCD is produced in small batches with various types. It is easy to obtain a template image in each batch. Therefore, based on the embedding method, we further explore and design an attention-based Siamese network.

2.3. Attention mechanism in computer vision

In computer vision, an attention mechanism can be thought of as a dynamic selection process that adaptively weights features based on their importance to the input. According to the dimension of attention, existing attention methods can be divided into channel attention, spatial attention, temporal attention, and branch channel attention (Guo et al., 2022b). In this paper, we focus on the spatial differences between batch templates and input samples.

Recurrent attention model (Mnih et al., 2014) adopts RNNs and reinforcement learning to make the network learn where to pay attention. Residual Attention Network (Wang et al., 2017) is proposed to focus on targeted regions while suppressing feature activations in irrelevant regions. Self-attention mechanism (Wang et al., 2018) is used to capture global information. Self-attention is powerful, but has a quadratic complexity. Most variants (Huang et al., 2020; Guo et al., 2022a; Fu et al., 2019) focus on reducing its computational complexity. Additionally, Vision Transformers (Dosovitskiy et al.), a pure attention-based networks, have been shown to achieve results comparable to modern Convolutional Neural Networks.

In this paper, the challenges in detection are the component displacement and feature deformation. To address these, we extend previous work and propose a dual-attention mechanism encoder and a recurrent residual attention mechanism decoder.

2.4. Siamese networks

Siamese Networks are a type of neural network architecture designed to perform tasks like one-shot learning and similarity comparisons. They consist of two identical subnetworks with shared weights, connected by a layer that computes the similarity between their outputs. Siamese Networks are trained on pairs of input samples, learning to differentiate between similar and dissimilar pairs. Common applications include face recognition, signature verification, and object tracking (Zhang and Peng, 2019).

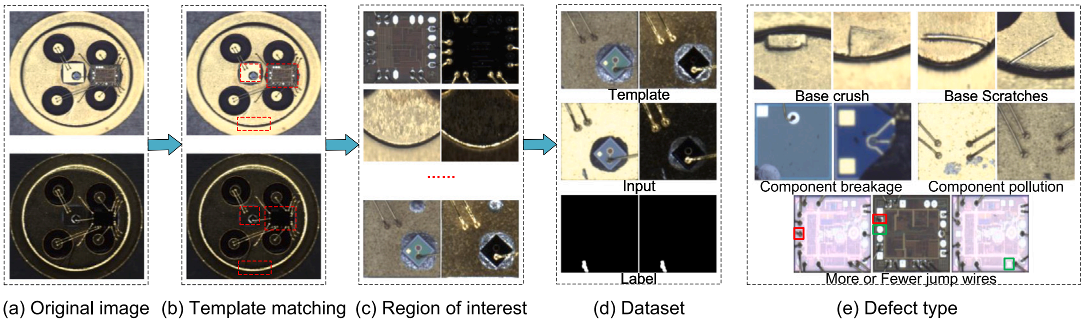
With the development of deep neural networks and datasets, there are many pre-train feature extraction networks. After careful consideration of feature extraction capabilities and real-time detection efficiency, we opted to use Resnet-34 (He et al., 2016) as the fundamental structure of Siamese network.

3. Methodology

3.1. Image acquisition system and datasets

3.1.1. Image acquisition system

The importance of designing an appropriate lighting scheme cannot be overstated when developing a visual acquisition system for OCDs, as it is essential for producing clear edges, complete areas, and highlighted defect characteristics. This, in turn, can help mitigate the difficulty of



(a) Original image (b) Template matching (c) Region of interest (d) Dataset (e) Defect type

Fig. 4. OCDs dataset.

subsequent algorithm processing. This paper utilizes two light sources, namely coaxial light and low-angle ring light, to illuminate the optical communication device, as shown in Fig. 3.

The coaxial light source emits light perpendicular to the surface of the optical communication device, maintaining clear edges of the component, eliminating shadows caused by height differences in the patch, and providing necessary color and grayscale information for detection. The low-angle ring light source emits light at a specific angle to the surface of the optical communication device, allowing for expressive imaging of the welding lines distributed in space and sufficient information regarding concave–convex and texture. Fig. 4(a) displays the images collected under both light sources. In summary, the use of coaxial light and low-angle ring light illumination images combines the advantages of both types of lighting to provide sufficient color, grayscale, bump, and texture information for subsequent segmentation algorithms.

3.1.2. OCDs dataset

To meet the required detection accuracy, the resolution of the collected image is set at 4608 × 3288 pixels. However, this is too large to be directly used as input for the detection network. Since component breakage, component pollution, and more or fewer jump wires can be detected based solely on the image information within the component range. And base crush and base scratches have specific prone areas for different batches. Therefore, only the corresponding Region of Interest (ROI) is extracted through template matching for defect segmentation. A standard template is drawn in the normal image, and the area of the component to be tested in the test image is determined by matching, as shown in Fig. 4(b).

Then, the size and direction of ROIs can vary greatly, and it is possible that ROIs with completely different sizes from the training data will appear in future detection processes. Therefore, after intercepting the ROI, the intercepted image is straightened according to the angle of its smallest circumscribing rectangle and uniformly scaled to 256 to ensure consistency, as shown in Fig. 4(c).

Finally, we created segmentation labels through manual annotation and stored templates, samples, and labels as datasets, as shown in Fig. 4(d). The constructed dataset contains a total of 918 sets of data, including 60 sets of base crushing, 27 sets of base scratches, 375 sets of component pollution, 240 sets of component breakage, and 216 sets of more or fewer jump wires. The dataset is partitioned into training, validation, and test sets at a 6:2:2 ratio. The detailed information regarding this division can be found in Section 3.2. The defect types are shown in Fig. 4(e) The specific number of samples for each type after division is presented in Table 1.

3.2. Problem definition

This paper focuses on the challenge of comparing the differences between template and testing samples. We assume the presence of three sets: a training set D_t, a validation set D_v, and a testing set D_s. Our model is trained on the training dataset D_t and the validation dataset D_v, and evaluated on the testing dataset D_s. Samples and their corresponding templates are collected from different batches. Each batch contains a defect-free template and several defective samples. We divide all samples into C_{seen} and C_{unseen}. It is assumed that the training dataset D_t and the validation dataset D_v contain only C_{seen}, while the testing dataset D_s contains only C_{unseen} in episodes, meaning that $C_{seen} \cap C_{unseen} = \emptyset$. Additionally, the templates are both observed in the seen category.

(1) $D_t = \left\{\left(x_i^t, \hat{x}_i^t, y_i^t\right)\right\}_1^N$, where x_i^t denotes the sample (query image), $\hat{x}_i^t$ represents the template (support image), and both x_i^t and the corresponding $\hat{x}_i^t$ belong to the same batch. The set $\left\{x_i^t\right\}_1^N \in C_{seen}$ contains all types of defects (including base crush, base scratches, component breakage, component pollution, and more or fewer jump wires). y_i^t corresponds to the sample labels, and N indicates the number of training episodes.

(2) $D_v = \left\{\left(x_i^v, \hat{x}_i^v, y_i^v\right)\right\}_1^M$, similar to D_t, x_i^v, $\hat{x}_i^v$, and y_i^v represent the sample (query image), the template (support image), and the label, respectively. To select the most optimal parametric model, we choose a validation set D_v from a different batch than the training set D_t.

(3) $D_s = \left\{\left(x_i^s, \hat{x}_i^s, y_i^s\right)\right\}_1^M$, analogous to D_t, x_i^s, $\hat{x}_i^s$, and y_i^s denote the sample (query image), the template (support image), and the label, respectively. However, the samples $\left\{x_i^s\right\}_1^M \in C_{unseen}$.

In this paper, our focus is on the model's ability to generalize to new batches rather than its ability to generalize to new defect types. Different batches exhibit variations in component type, placement, and quantities. Importantly, these differences primarily affect the defect-free background rather than the defect foreground. As a result, we simplify the problem to a binary classification task, specifically defective foreground (which includes five defects) and defect-free background segmentation problems. Consequently, in our task, to achieve generalization for a new batch, only one defect-free sample needs to be collected as a template, which includes the new background features.

It is worth noting that the differences between samples and templates not only include defect features but also noise. We assume that the defect feature is d, and the noise of displacement, deformations, and texture change is z_p, z_f, and z_t, as shown in Fig. 1. Then the feature map is represented as $F = (f_{i,j}) \in \mathbb{F}^{H \times W}$. For ease of comprehension, we set $H, W = 3, 3$. The feature maps of template $F_{\hat{x}_i^t}$, sample only with defect features $F^d_{x_i^t}$, sample with displacement noise $F^{z_p}_{x_i^t}$, sample with deformation noise $F^{z_f}_{x_i^t}$, and sample with texture change $F^{z_t}_{x_i^t}$ are represented as:

$$F_{\hat{x}_i^t} = \begin{bmatrix} f_{1,1} & f_{1,2} & f_{1,3} \\ f_{2,1} & f_{2,2} & f_{2,3} \\ f_{3,1} & f_{3,2} & f_{3,3} \end{bmatrix} \tag{1}$$

$$F^d_{x_i^t} = \begin{bmatrix} f_{1,1} & f_{1,2} & f_{1,3} \\ f_{2,1} & \mathbf{d_{2,2}} & \mathbf{d_{2,3}} \\ f_{3,1} & f_{3,2} & f_{3,3} \end{bmatrix} \tag{2}$$

Table 1
The number of samples for each type defects in OCDs dataset.

Type	Base crush	Base Scratches	Component breakage	Component pollution	More or fewer jump wires
Training Set	36	17	144	225	130
Validation Set	12	5	48	75	43
Testing Set	12	5	48	75	43

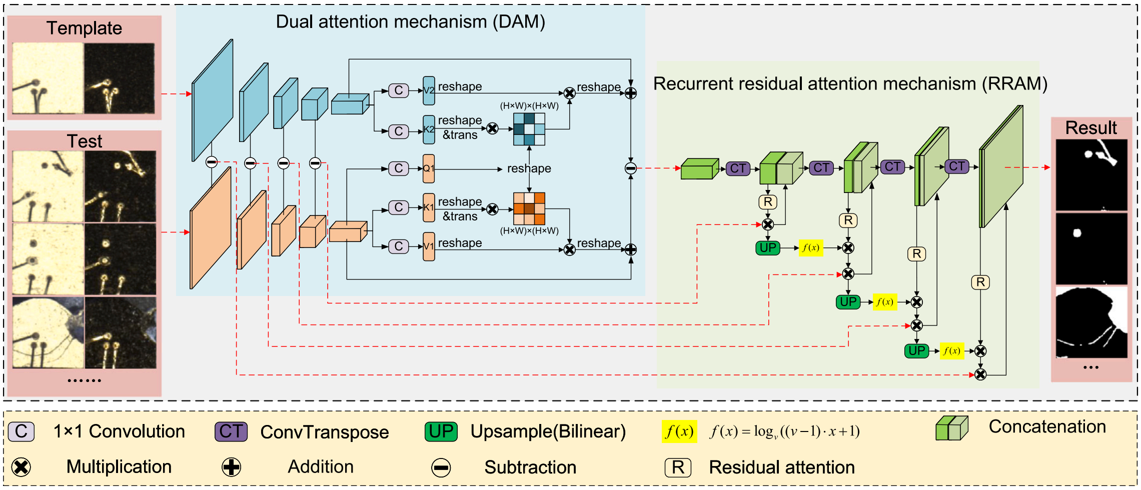


Fig. 5. The overview of generalized well networks (GWNet) for defect segmentation.

$$F_{x_i^t}^{z_p} = \begin{bmatrix} f_{1,1} & f_{1,2} & f_{1,3} \\ \boldsymbol{f}_{\mathbf{3,1}} & \boldsymbol{f}_{\mathbf{3,2}} & f_{2,3} \\ \boldsymbol{f}_{\mathbf{2,1}} & \boldsymbol{f}_{\mathbf{2,2}} & f_{3,3} \end{bmatrix} \tag{3}$$

$$F_{x_i^t}^{z_f} = \begin{bmatrix} f_{1,1} & \boldsymbol{f}_{\mathbf{2,3}} & f_{2,3} \\ \boldsymbol{f}_{\mathbf{1,2}} & f_{2,2} & \boldsymbol{f}_{\mathbf{3,2}} \\ f_{3,1} & \boldsymbol{f}_{\mathbf{2,1}} & f_{3,3} \end{bmatrix} \tag{4}$$

$$F_{x_i^t}^{z_t} = \begin{bmatrix} \lambda(f_{1,1}) & \lambda(f_{1,2}) & \lambda(f_{1,3}) \\ \lambda(f_{2,1}) & \lambda(f_{2,2}) & \lambda(f_{2,3}) \\ \lambda(f_{3,1}) & \lambda(z_{3,2}) & \lambda(f_{3,3}) \end{bmatrix} \tag{5}$$

where d is the defect feature, $\lambda(\cdot)$ is the texture change noise. Neural networks have the ability to represent data strongly, which makes the process of aligning and eliminating noise z_t relatively straightforward. However, the displacement and deformation noise are not as easily removed due to the translation equivariance property of neural networks.

3.3. Framework overview

To address the above challenges, we propose a generalized well networks for defect segmentation called GWNet. As shown in Fig. 5, GWNet are divided into two stages, feature extraction and feature fusion. In the stage of feature extraction, the encoder consist of siamese networks and dual-attention mechanism (DAM). We use Siamese network to extract features from both samples and templates. Then DAM is employed to compare feature differences and eliminate noise, thereby obtaining noise-free defect features. In order to remove the noise of deformations and deformations, inspired by language processing method for long dependency text, we leveraged self-attention feature maps extracted from test samples, as well as cross-attention maps derived from templates and test samples, neither of which are position-independent. In the stage of feature fusion, we employ the recurrent residual attention mechanism (RRAM) to fuse multi-scale features to achieve higher accuracy segmentation. To eliminate noise in multi-scale features extracting by Siamese networks, we employ a deep-to-shallow approach that utilizes the noise-free features obtained from DAM to obtain multi-scale noise masks.

3.4. Stage of feature extraction

In this section, we provide a detailed introduction to Siamese networks and DAM. Additionally, we elucidate the principle behind how DAM is able to shield noise.

The convolutional operation is translationally equivariant, which means that when the features in the image are displaced or deformed, the feature map is also displaced or deformed. So that the convolution operation is difficult to remove the noise of displacement and deformation. In natural language processing, self-attention involves calculating the response as a weighted sum of the input sequence, but this process can cause the loss of positional information. Therefore, we propose a DAM based on self-attention.

The structure of DAM is shown in Fig. 6. The inputs are the deepest feature maps extracting from template and testing sample by Siamese networks. The template feature map is $F_{\hat{x}} = (\hat{f}_{i,j}) \in \mathbb{F}^{C\times H\times W}$, and the testing sample feature map is $F_x = (f_{i,j}) \in \mathbb{F}^{C\times H\times W}$. At first, $Q, K, \hat{K} \in \mathbb{F}^{C_1\times H\times W}(C_1 < C)$ and $V, \hat{V} \in \mathbb{F}^{C\times H\times W}$ are obtained by 1×1 convolution.

$$\begin{aligned} Q &= W^q F_x = (W^q f_{i,j}) \in \mathbb{F}^{C_1\times H\times W} \\ K &= W^k F_x = (W^k f_{i,j}) \in \mathbb{F}^{C_1\times H\times W} \\ \hat{K} &= W^k F_{\hat{x}} = (W^k \hat{f}_{i,j}) \in \mathbb{F}^{C_1\times H\times W} \\ V &= W^v F_x = (W^v f_{i,j}) \in \mathbb{F}^{C\times H\times W} \\ \hat{V} &= W^v F_{\hat{x}} = (W^v \hat{f}_{i,j}) \in \mathbb{F}^{C\times H\times W} \end{aligned} \tag{6}$$

Then, $Q, K, \hat{K}$ are all resized to $Q, K, \hat{K} \in \mathbb{F}^{M\times C_1}$, and $V, \hat{V}$ are resized to $V, \hat{V} \in \mathbb{F}^{M\times C}$, where $M = H \times W$. The self-attention feature map $F_{self} \in \mathbb{F}^{M\times C}$ is

$$F_{self} = (\frac{QK^T}{\sqrt{C_1}})V \tag{7}$$

The cross-attention feature map $F_{cross} \in \mathbb{F}^{M\times C}$ is

$$F_{cross} = (\frac{Q\hat{K}^T}{\sqrt{C}})\hat{V} \tag{8}$$

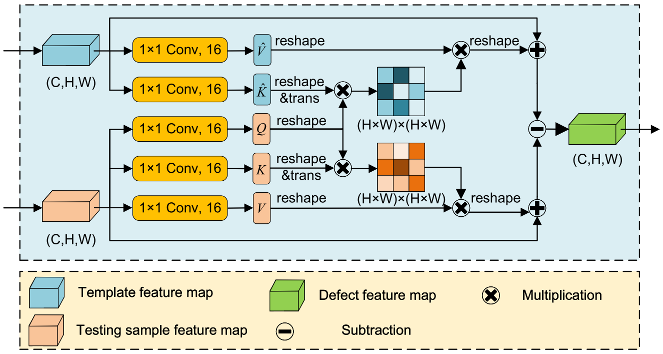


Fig. 6. Dual-attention mechanism.

Finally, the defect feature map $F^d \in \mathbb{F}^{M\times C}$ are obtained by calculating the subtracting of F_{self} and F_{cross}.

$$F^d = F_{self} - F_{cross} \tag{9}$$

In general, there are not only defect features but also some noise between the template and the test sample, including displacement, deformation, and texture variation. The dependencies of noise and defect features are very complex and not the focus of this paper. Therefore, this paper decomposes the proof process into two steps. Firstly, it is proved that DAM can remove the noises. Secondly, it is proved that DAM can preserve defect features.

For the convenience of calculation, the dimensionality reduction operation of Q, K, $\hat{K}$ is also ignored in the calculation. After resize operation, $Q = (q_i) \in \mathbb{F}^M$, $K = (k_i) \in \mathbb{F}^M$, $\hat{K} = (\hat{k}_i) \in \mathbb{F}^M$, $V = (v_i) \in \mathbb{F}^M$, $\hat{V} = (\hat{v}_i) \in \mathbb{F}^M$. Then, according to Eqs. (6) (7) (8) (9), the defect feature map $F^d = (f_i^d) \in \mathbb{F}^M$ can be calculated as follows:

$$\begin{aligned} f_i^d &= \sum_{j=1}^{M}(q_i k_j v_i - q_i \hat{k}_j \hat{v}_i) \\ &= \sum_{j=1}^{M} q_i(k_j v_i - \hat{k}_j \hat{v}_i) \\ &= \sum_{j=1}^{M} W^q f_i(W^k f_j W^v f_i - W^k \hat{f}_j W^v \hat{f}_i) \\ &= \sum_{j=1}^{M} W^q f_i W^k(f_j W^v f_i - \hat{f}_j W^v \hat{f}_i) \end{aligned} \tag{10}$$

where $i = 1, \dots, M$.

As for the first step, there are three types of noises. (1) For the displacement noise, we compare $F_{x_i^t}^{z_p}$ in Eq. (3) with $F_{\hat{x}_i^t}$ in Eq. (1) using Eq. (10). For the calculation of ith feature map f_i^d, the subterm $\sum_{j=1}^{M}(f_j W^v f_i - \hat{f}_j W^v \hat{f}_i)$ is decisive. We find that if $[f_1, f_2 \dots, f_M]$ is merely shuffled, the final result will not change, $f_i^d = 0$. (2) For the deformation noise, as shown in Eq. (4), the $[f_1, f_2 \dots, f_M]$ is just shuffled by other means too. The final result $f_i^d = 0$. (3) For the texture change noise, as shown in Eq. (5), $\lambda(\cdot)$ is easily fitted by the powerful representation ability of the neural network.

As for the second step, we compare $F_{x_i^t}^d$ in Eq. (2) with $F_{\hat{x}_i^t}$ in Eq. (1) using Eq. (10). $f_i^d = W^q f_i W^k(d_5 W^v f_i + d_6 W^v f_i - f_5 W^v f_i - f_6 W^v f_i)$. Therefore, it is proved that DAM can preserve defect features

In summary, DAM uses self-attention non-position information to extracts defect features while removing the displacement noise and deformation noise.

3.5. Stage of feature fusion

Multi-scale feature fusion is the key to achieve accuracy defect segmentation. Deep feature maps are imbued with a wealth of semantic information, while shallow feature maps harbor an abundance of intricate details. While the deepest feature maps that exclusively capture noise-free defect semantics can be obtained through DAM, other feature maps remain prone to noise. Therefore, we introduce a RRAM to generate multi-scale denoising masks in a deep-to-shallow manner.

As shown in Fig. 5, there are five feature maps obtained from Siamese networks. The feature maps of the first four layers are skip-connected to the feature fusion stage. At first, the resolution of the deepest feature map is increased by a factor of two through the use of a deconvolution operation. Secondly, using RRAM, we generate a single-channel mask of the same size as the feature map. The mask assigns values between 0 and 1 to all positions, with 0 indicating complete masking and 1 indicating complete passage. And the skip-connected feature map is multiplied bitwise with the mask to mask the noise. Finally, the masked feature maps and the upsampling feature maps from the last decoder are concatenated to calculate the feature map of the next layer. After five times of upsampling, we can get a feature map of the same size as the input sample. The defect segmentation probability map is obtained by 3 × 3 convolution operation.

The aforementioned operations remain consistent across all layers, and the deepest feature map is devoid of noise. Therefore, based on residual attention (Wang et al., 2017) and recurrent neural networks (Mnih et al., 2014), we propose a deep-to-shallow RRAM approach, in which each spatial mask generated is corrected based on the previous masking results, as shown in Fig. 7.

Each mask is obtained by multiplying all previous masks, which can effectively reduce the interference of occasional noise in subsequent masks. But the low resolution of the previous mask blurs the edge of the upsampling mask. Therefore, to maintain relatively clear edges and prevent the spatial mask from being excessively small when obtained by probability multiplication, we enhance the probability prior to the multiplication process. The mapping function is shown in Eq. (11).

$$f(x) = \log_v((v-1)\cdot x + 1) \tag{11}$$

where v is utilized to adjust the level of enhancement. The larger the value of V, the stronger the tensile strength of the function for the low probability part, as shown in Fig. 8

3.6. Loss function

Defective pixels are typically sparse, and there is an imbalanced distribution between positive and negative class pixels. To address this issue, this paper proposes the use of focal loss (Lin et al., 2017) as the loss function during training, which prioritizes hard-to-segment and mis-segmented pixels, thereby mitigating the problems caused by data imbalance.

$$Loss = -\alpha(1-p_t)^{\gamma}\log(p_t) \tag{12}$$

where $p_t \in [0,1]$ is the probability obtained by GWNet. And $\gamma \geq 0$ is tunable focusing parameter. When $\gamma = 0$, focal loss is equivalent to cross entropy loss, and as γ is increased the effect of the modulating factor is like wise increased (we set $\gamma = 2$ in our model). $\alpha \in [0,1]$ as a weighting factor to address class imbalance (we set $\alpha = 0.25$ in our model).

3.7. Training mechanism

In this paper, the dataset is divided into training, validation, and test sets in a 6:2:2 ratio. We assume that $D_t = \left\{\left(x_i^t, \hat{x}_i^t, y_i^t\right)\right\}_1^N$, $D_v = \left\{\left(x_i^v, \hat{x}_i^v, y_i^v\right)\right\}_1^N$, and $D_s = \left\{\left(x_i^s, \hat{x}_i^s, y_i^s\right)\right\}_1^N$.

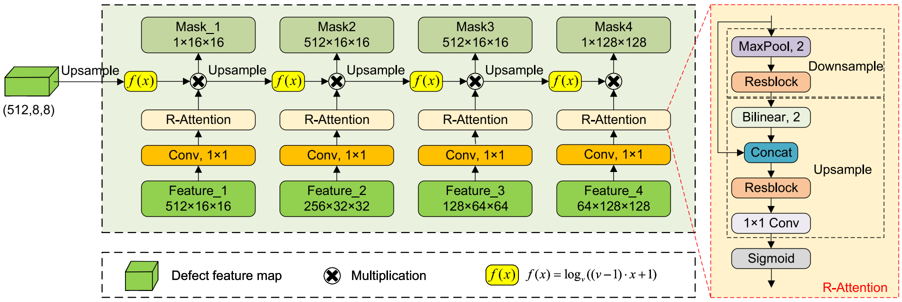


Fig. 7. Recurrent residual attention mechanism.

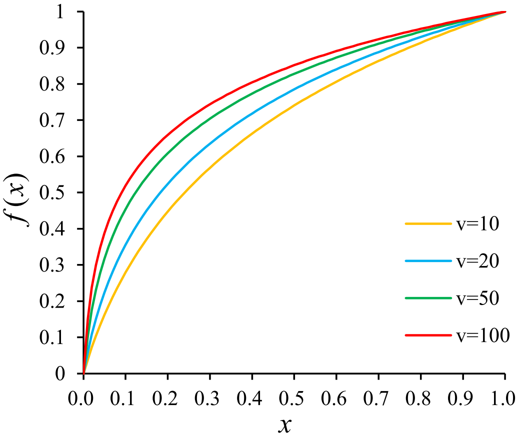


Fig. 8. Mapping function graph.

As shown in algorithm 1, we input the concatenation of sample x_i^t and template $\hat{x}_i$ into GWNet to obtain the probability $f_\theta(x_i, \hat{x}_i)$. Then, according to Eq. (12), we calculate the segmentation loss $Loss(f_\theta(x_i, \hat{x}_i), y_i)$. In training stage, the GWNet is updated by training dataset. And the evaluation dataset is used to evaluate and save the best model.

Algorithm 1 Pseudo code of GWNet

```
Input: Training set $\{(x_i^t, \hat{x}_i^t, y_i^t)\}_{i=1}^N$, validation set $\{(x_i^v, \hat{x}_i^v, y_i^v)\}_{i=1}^M$,
    model $f_\theta$, the maximum number of iterations $T$, break index $T_b$
Initialization: $f_\theta$, index $t_b \leftarrow 0$, training loss $l^t \leftarrow 0$, validation loss
    $l^v \leftarrow 0$, best validation loss $l_{bset}^v \leftarrow 0$;
for $t = 1 \cdots T$ do
    $t_b = t_b + 1$;
    if $t_b > T_b$ then
        break;
    end if
    $l^t \leftarrow \frac{1}{N} \sum_{i=1}^N Loss(f_\theta(x_i^t, \hat{x}_i^t), y_i^t)$;
    Update $f_\theta$ to minimize $l^t$: $\theta \leftarrow \nabla_\theta l^t$;
    $l^v \leftarrow \frac{1}{M} \sum_{i=1}^M Loss(f_\theta(x_i^v, \hat{x}_i^v), y_i^v)$;
    if $l^v < l_{bset}^v$ then
        $t_b \leftarrow 0$, $l_{bset}^v \leftarrow l^v$;
    end if
end for
Output: Updated model $f_\theta$
```

4. Experiments

4.1. Implementation details

Our network is implemented in the Pytorch platform with a single NVIDIA Tesla V100. We utilize the Adams optimizer to train GWNet with a batch size of 16 and a learning rate of 0.00001. The maximum number of iterations $T = 250$, break index $T_b = 25$.

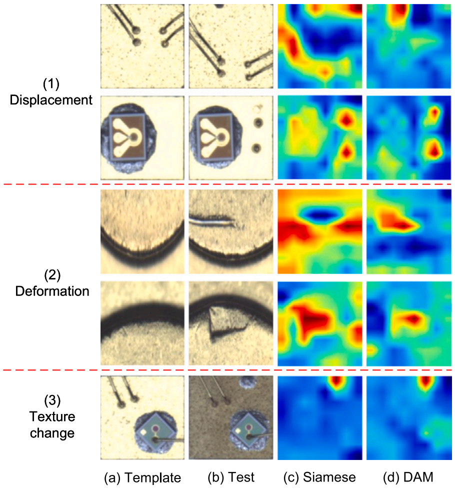


Fig. 9. Visual comparison of DAM with Siamese network in the OCDs dataset.

4.2. Evaluation metrics

We employ five widely used evaluation metrics for semantic segmentation: precision (Pre), recall (Rec), F-measure (F1), mIoU, and mACC to evaluate the performance between different methods. The mIoU is a popular evaluation metric for semantic segmentation that measures the degree of overlap between the predictions and labels. F-measure is the weighted harmonic mean of precision and recall, which comprehensively reflects the performance of binary semantic segmentation. The definitions of this metrics are:

$$Pre = \frac{TP}{TP + FP} \tag{13}$$

$$Rec = \frac{TP}{TP + FN} \tag{14}$$

$$F1 = \frac{2 \times Pre \times Rec}{Pre + Rec} \tag{15}$$

$$mIoU = \frac{TP}{FP + FN + TP} \tag{16}$$

$$mACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{17}$$

where TP, FP, TN, and FN represent the number of true positive, false positive, true negative, and false negative.

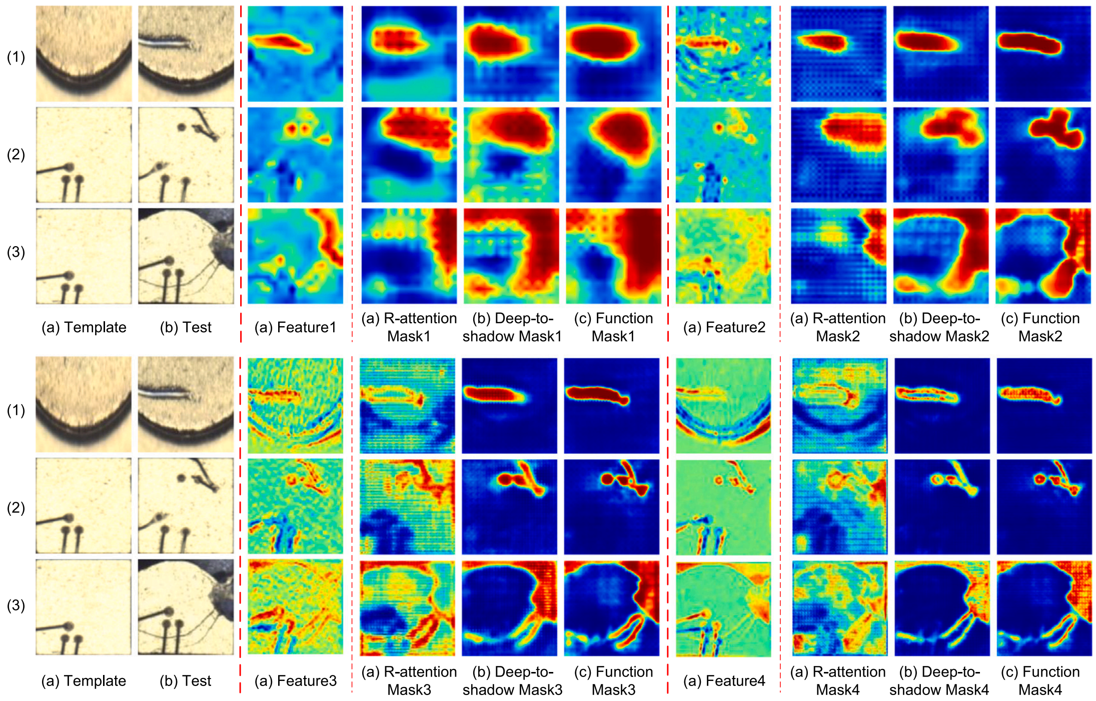


Fig. 10. Visual ablation of RRAM in the OCDs dataset.

Table 2
Results of ablation in the OCDs dataset.

Modules	Baseline	Siamese	DAM	RRAM	mIoU	F1
S1	✓				0.6766	0.8102
S2	✓	✓			0.6975	0.8246
S3	✓	✓	✓		0.7747	0.8772
S4	✓	✓		✓	0.7692	0.8732
S5	✓	✓	✓	✓	0.8074	0.8980

4.3. Ablation studies and discussion

In this section, we perform sufficient ablation experiments to demonstrate the effectiveness of each component. The performance of different variants of GWNet is shown in Table 2. The baseline method utilizes ResNet-34 as the encoder for feature extraction, while the decoder structure is consistent with the one used in our proposed model, with the removal of the RRAM component. The input for the baseline method is formed by concatenating the samples and templates in the channel dimension. Furthermore, utilizing the class activation map, we provide a qualitative analysis of the noise removal effects of DAM and RRAM.

4.3.1. The importance of DAM

We analyzed the effectiveness of DAM by both quantitative and qualitative methods. Quantitatively, as shown in Table 2, simply Siamese networks has an improvement over baseline, but the improvement is limited (IoU increased by 2.09%). With DAM, the method has been significantly improved (compared with baseline, IoU increased by 9.81%).

Qualitatively, we compare the DAM and Siamese networks through visual analysis, as shown in Fig. 9. The feature maps are subtracted to obtain the defect feature map. The size of defect feature map is 512 × 8 × 8, which cannot be analyzed visually. In this paper, we calculate the average activation value of each spatial position by averaging along the channel direction, resulting in a single channel 8 × 8 average activation Map. For ease of observation, the size is upsampled to 256 × 256. After normalization, the resulting image is transformed into a pseudo-color representation and displayed.

As illustrate in Fig. 9, the feature maps of Siamese networks cannot remove the noise of displacement and deformation. Occasionally, the attention paid to noise is even higher than that of semantic defects, which is not conducive to the precise segmentation of defects in subsequent feature fusion decoder. The proposed DAM is effective in suppressing noise during the calculation of the defect feature map, resulting in improved segmentation. And both the Siamese network and DAM can remove the noise of texture changes, which also proves the robustness of convolution operations to pixel value changes.

4.3.2. The effectiveness of RRAM

Qualitatively, as shown in Table 2, compared with the Siamese networks, RRAM also significantly improve the method (compared with baseline, IoU increased by 9.26%). In addition, combined RRAM with RRAM, the method has more improvement over baseline (IoU increased by 13.08%). Comparison of S4 with S5 demonstrates that employing the noise-free defect feature in a deep-to-shadow approach can lead to further improvements in segmentation accuracy.

We obtain the four down-sample pseudo-color representation, as shown in Fig. 10. Deep feature maps are imbued with a wealth of semantic information, while shallow feature maps harbor an abundance of intricate details. Nevertheless, the presence of numerous noises in both deep and shallow features, particularly shadows, poses a significant challenge. Therefore, we propose the RRAM, and analysis the influence each part component of RRAM (residual-attention (R-attention), deep-to-shallow, and the enhanced mapping function).

In the result of R-attention, due to the independence of the four calculations of the spatial mask parameters, the downsampling portion cannot acquire adequate global information when calculating the spatial mask as the size of the decoder feature map increases during the upsampling process. As a result, the masks calculated by the last two stages may contain more noisy regions and be less precise.

In the result of deep-to-shallow, our findings indicate that the noise in the mask was considerably reduced and progressively became more accurate from deep-to-shallow.

Table 3
Quantitative comparison with state-of-arts methods in OCDs datasets.

	Method	Pre	Recall	F1	mIoU
Classical methods	U-Net	0.8597	0.6926	0.7590	0.6325
	FCN	0.8831	0.7376	0.8038	0.6580
	SegNet	0.8949	0.3907	0.5440	0.3662
	DeepLabV3+	0.8295	0.7967	0.8128	0.6702
	PGANet	0.9186	0.4793	0.6299	0.4483
Attention-based methods	CCNet	0.8224	0.3875	0.5268	0.3614
	DUNet	0.8716	0.3100	0.4574	0.2942
	DANet	0.8220	0.5748	0.6765	0.5130
	Swin-U-Net	0.6612	0.2569	0.3700	0.2076
One-shot learning methods	TGRNet	0.2446	0.4770	0.3233	0.1638
	PFENet	0.1929	0.2713	0.2255	0.1233
	Siamese U-Net	0.8913	0.6946	0.7807	0.6243
	DSSSNet	0.8931	0.8148	0.8521	0.7405
Ours	GWNet	**0.9070**	**0.8891**	**0.8980**	**0.8074**

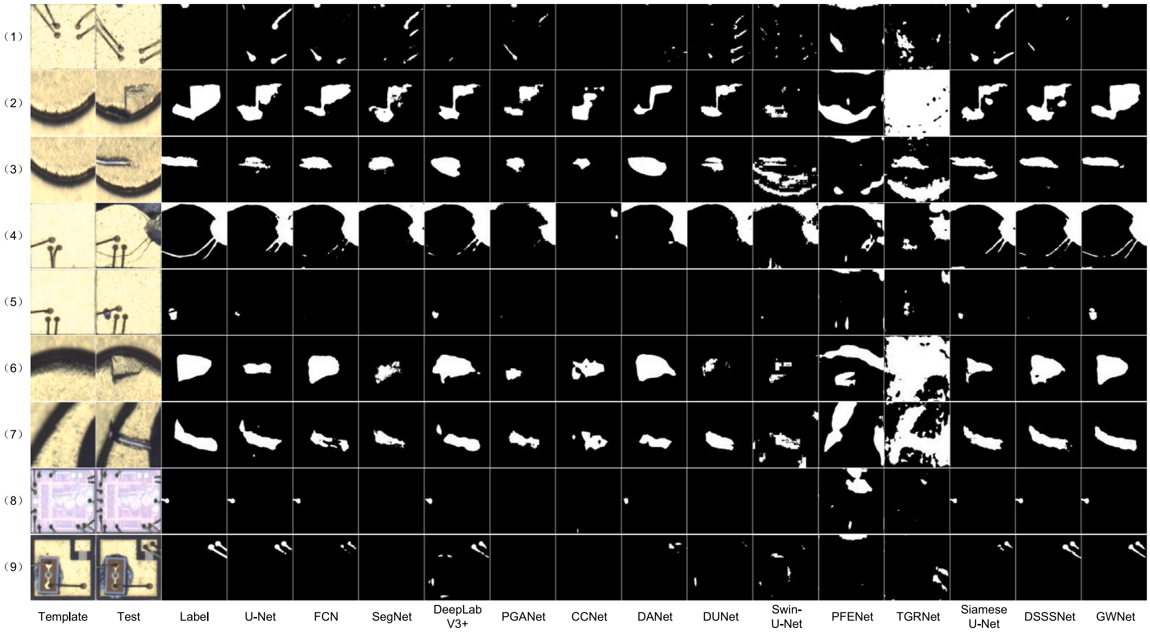


Fig. 11. Visual comparison with state-of-the-arts methods in OCDs.

In the results of the enhanced mapping function, compared to deep-to-shallow approach, the masks undergo a probabilistic stretching before being forwarded to the subsequent layer. This enhances the clarity of mask edges, with no significant transitional regions, and the issue of undersized masks during multiple multiplications is eliminated.

4.4. Comparison with the state-of-the-art model

We compared our method GWNet with eleven state-of-the-art methods, including five classical segmentation methods (U-Net Ronneberger et al., 2015, FCN Long et al., 2015, SegNet Badrinarayanan et al., 2017, DeepLabV3+ Chen et al., 2018), and PGANet (Dong et al., 2019), four attention based segmentation methods (CCNet Huang et al., 2020, DUNet Jin et al., 2019, DANet Fu et al., 2019) and Swin-U-Net (Cao et al., 2023), and four One-shot learning methods (DSSSNet Ling et al., 2022, Siamese U-Net Kwon et al., 2019, TGRNet Bao et al., 2021, PFENet Tian et al., 2020). And these methods are compared on three different datasets, including OCDs dataset (binary segmentation), PCBs dataset (Ling et al., 2022) (multi-class segmentation), and MCSD dataset (Niu et al., 2022)binary segmentation. OCDs dataset and PCBs dataset are both flexible production lines, with obvious characteristics of small batches and multiple types. And there are noises between inputs and templates, including displacement, deformation, and texture change. The MCSD dataset is collected from a flexible production line, and the primary differences between various batches are mainly due to changes in texture.

4.4.1. Training setting

As shown in Section 3.2, there are three sets: a training set $D_t = \left\{\left(x_i^t, \hat{x}_i^t, y_i^t\right)\right\}_1^N$, $D_v = \left\{\left(x_i^v, \hat{x}_i^v, y_i^v\right)\right\}_1^M$, and $D_s = \left\{\left(x_i^s, \hat{x}_i^s, y_i^s\right)\right\}_1^M$, among that, x_i is the sample, $\hat{x}_i$ is the template, y_i is the label.

For strong supervision training methods, which encompass both classical and attention-based approaches, the sample x_i and template $\hat{x}_i$ are concatenated in the channel dimension as the input. As demonstrated in Algorithm 1, the training set D_t is utilized to adjust the model's parameters, weights, and biases to minimize the loss function. The validation set assists in model selection, which involves choosing the optimal hyperparameters for the model to generalize well to new data.

One-shot learning-based methods can be categorized into two types: Siamese-based networks (such as GWNet, DSSSNet, and Siamese U-Net) and other networks (like TGRNet and PFENet). For Siamese networks, the sample x_i and template $\hat{x}_i$ are independently input into the two branches of the Siamese network. The resulting two sets of features are compared using specific operations to obtain defect features. The training mechanism is illustrated in Algorithm 1.

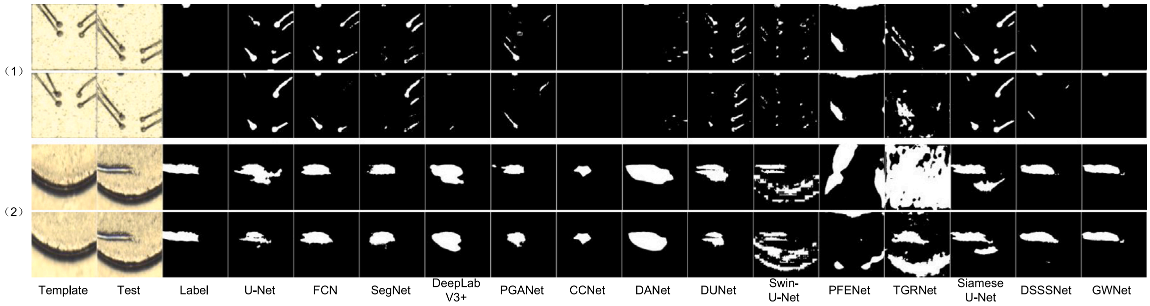


Fig. 12. The influence of changing the template in OCDs.

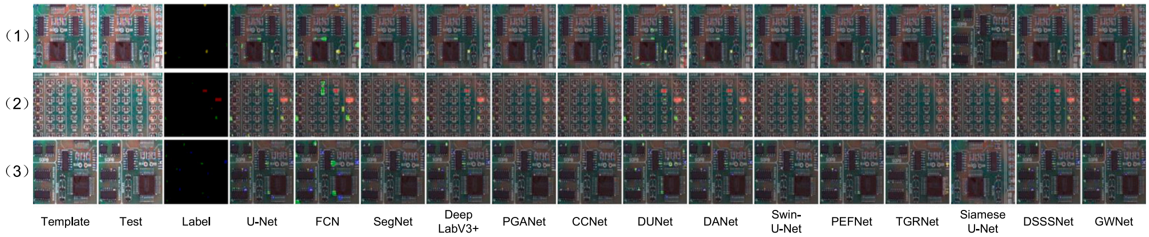


Fig. 13. Visual comparison with state-of-the-arts methods in PCBs.

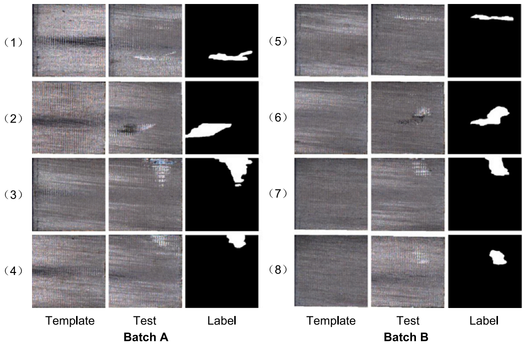


Fig. 14. MCSD dataset visualization. As a result of tooling and tool wear, the background and defect characteristics of different batches exhibit variations in texture.

In the case of PFENet, one sample from each type of defect (one type in OCDs and MCSD, five types in PCB) is selected as support images, while the remaining samples are chosen as query images. If there are K types of defects, the task is referred to as a K-way one-shot task. For TGRNet, not only one sample from each type of defect are selected as support images, but the templates $\hat{x}_i$ are also considered normal images. This is referred to as the K-way One-shot W-normal task in Ling et al. (2022). The training mechanism for TGRNet and PFENet is similar to that of strong supervised training methods and Siamese-based networks, as shown in Algorithm 1.

4.4.2. Comparison in OCDs dataset

In this section, we not only analyzed quantitatively (as shown in Table 3) and qualitatively (as shown in Fig. 11), but also compared the impact of different templates on the experimental results (as shown in Fig. 12).

Quantitatively, in Table 3, we can observe that with Siamese networks, existing one-shot learning based methods show a performance improvement compared to the classical methods, which demonstrates that Siamese networks could effectively improve the generalization ability. In contrast, attention based methods perform worse, probably because it fits the training data well, but performs poorly on the unseen test data. In addition, our method still performs the best results. Compared with the suboptimal method (DSSSNet), we improved the mIoU by 6.69%. Due to the ability to remove noise, our method has more outstanding performance.

Qualitatively, in Fig. 11, we can find that classical method and attention based methods are very sensitive to noises (including displacements and deformations). In contrast, DSSSNet performs well with Siamese networks by utilizing a global pooling operator to obtain the attention map, which makes the features less sensitive to local offsets. However, when the Siamese U-Net compares the difference between the template and the test sample, it cannot shield the noise because it is directly subtracted, and the attention map is obtained through the activation function. Pooling operators are widely recognized for their ability to attain translation and rotation invariance in neural networks, whereas activation functions do not have this capability. Thus, the performance of DSSSNet and Siamese U-Net illustrates that the convolutional neural networks' translation equivariance is a factor that makes comparison features vulnerable to displacement and deformation noise. Our method outperforms other state-of-the-art methods by utilizing a non-position information DAM and multi-scale denoising RRAM.

At last, to further validate the effectiveness of GWNet, we evaluate the influence of changing the template on the experimental outcomes. Fig. 12 illustrates that GWNet is minimally affected by template changes, whereas other methods exhibit high sensitive. Among the other methods, DSSSNet performs better, which further corroborates the conclusion about translation equivariance in the previous paragraph.

4.4.3. Comparison in PCBs dataset

To further validate the method, we conduct experiments on the public dataset PCBs. Since PCBs have multiple classification labels, we added two One-shot learning methods, including TGRNet (Bao et al., 2021) and PFENet (Tian et al., 2020).

Quantitatively, as shown in Table 4, the superior performance of attention-based approaches on PCBs, as opposed to the OCDs dataset, can be attributed to the identical distribution of the training and testing sets. In addition, it is worth noting that TGRNet and PFENet show only marginal improvement in performance, suggesting that solely emphasizing texture changes is insufficient without considering local offsets. In contrast, GWNet achieve the best mIoU and mACC value.

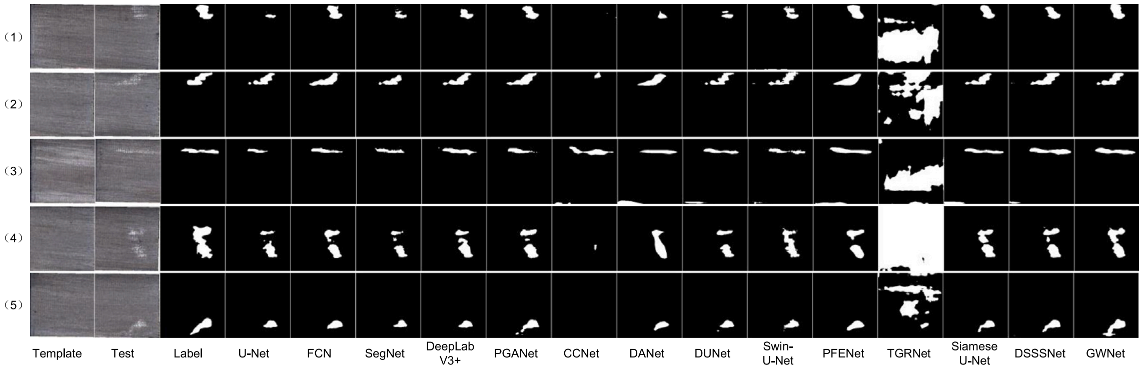


Fig. 15. Visual comparison with state-of-the-arts methods in MCSD.

Table 4
Quantitative comparison with state-of-arts methods in PCBs datasets.

	Method	mIoU	mACC	Params (MB)
Classical methods	U-Net	0.5981	0.9046	148.5
	FCN	0.4985	0.8203	**15.32**
	SegNet	0.7864	0.9974	40.47
	DeepLabV3+	0.7094	0.9351	32.98
	PGANet	0.7894	0.9975	51.41
Attention-based methods	CCNet	0.4786	0.9087	67.70
	DUNet	0.7513	0.9126	31.48
	DANet	0.7243	0.9003	49.63
	Swin-U-Net	0.7719	0.9972	27.16
One-shot learning methods	TGRNet	0.7068	0.8988	32.12
	PFENet	0.7214	0.9105	30.25
	Siamese U-Net	0.7837	0.9954	7.85
	DSSSNet	0.7634	0.9678	33.60
Ours	GWNet	**0.8243**	**0.9978**	26.54

Table 5
Quantitative comparison with state-of-arts methods in MCSD datasets.

	Method	Pre	Recall	F1	mIoU
Classical methods	U-Net	**0.9644**	0.4696	0.6316	0.4603
	FCN	0.9256	0.6085	0.7343	0.5756
	SegNet	0.9473	0.4339	0.5951	0.4227
	DeepLabV3+	0.8949	0.6333	0.7417	0.5857
	PGANet	0.8991	0.7205	0.8000	0.6678
Attention-based methods	CCNet	0.8224	0.3875	0.5268	0.3614
	DUNet	0.8716	0.3100	0.4574	0.2942
	DANet	0.8220	0.5748	0.6765	0.5130
	Swin-U-Net	0.9108	0.6083	0.7296	0.5739
One-shot learning methods	TGRNet	0.2471	0.4000	0.3055	0.0289
	PFENet	0.8096	0.6763	0.7370	0.5721
	Siamese U-Net	0.9174	0.6593	0.7672	0.6237
	DSSSNet	0.8831	0.7232	0.7952	0.6552
Ours	GWNet	0.8790	**0.7451**	**0.8065**	**0.6724**

Furthermore, GWNet has the second-lowest number of parameters among the tested models.

Qualitatively, as depicted in Fig. 13, our method achieves the most precise segmentation results. Specifically, while other methods can identify the defect, they struggle to accurately locate the complex defect shapes. Our approach, thanks to DAM and RRAM, can accurately segment the defect region from the complex background. In row (3), where the defect areas in the image are small and dense, most methods fail to completely segment the defect areas, while our method still accurately identifies the defects.

4.4.4. Comparison in MCSD dataset

This paper focuses on how to distinguish defect features from noise features. Although our focus is on how to remove displacement and deformation noise, it also works on texture change noise. To demonstrate this, we augment the MCSD dataset.

MCSD is a metal surface defect detection task. We selected two batches of data. As a result of tooling and tool wear, the background and defect characteristics of different batches exhibit variations in texture. As shown in Fig. 14, due to tool wear, the defect features in Batch B are shallow, whereas in Batch A, with minimal tool wear, a specular reflection appears in the middle of the image, and the defect features are more pronounced.

Quantitatively, as shown in Table 5, our method achieves the best results in both F1-score and mIoU metrics. As the differences are primarily due to texture changes, it becomes apparent that classical methods and other one-shot learning techniques have also yielded improved results. This finding supports the assertion in this article that neural networks can effectively eliminate the noise caused by texture changes. In addition, compared to U-net, the Siamese U-net has improved F1-score and mIoU by 0.1356 and 0.1634, respectively. This demonstrates that, in the context of texture change noise, learning to compare the differences between templates and testing samples, rather than directly learning the sample representations, is also effective.

Qualitatively, as illustrated in Fig. 15, our method produces the most precise segmentation results. Owing to RRAM's fusion of multi-scale information and noise shielding, our method detects defects with sharper outlines and more accurate identification of weak-feature defects.

5. Conclusion

In this paper, we proposed a novel Generalized Well Neural Network (GWNet) via Template-Testing comparison for surface defect segmentation in Optical Communication Device. Our approach learns to compare the differences between templates and test samples, allowing the model to generalize to new batches by collecting templates. We addressed the challenge of noise removal by introducing the Dual-Attention Mechanism (DAM) and the Recurrent Residual Attention Mechanism (RRAM) in the feature extraction and feature fusion stages, respectively. Our experiments on OCDs, PCBs and MCSD datasets demonstrated that GWNet outperforms state-of-the-art methods. Our work provides a promising direction for addressing the challenge of surface defect detection in flexible manufacturing systems. By developing avia Template-Testing comparison method that can generalize to new batches, we have the potential to reduce the amount of data required to train models for surface defect detection, saving time and resources. This work can have a significant impact on improving the efficiency and reliability of flexible manufacturing systems. In future work, we plan to expand our approach to other FMS applications and explore lightweight networks for real-time detection.

CRediT authorship contribution statement

Tongzhi Niu: Designing the experiments, Implementing the algorithms, Analyzing the results, Writing the manuscript, Participated in the discussion and revision of the manuscript. **Zhiyu Xie:** Implementing the algorithms, Writing the manuscript, Participated in the discussion and revision of the manuscript. **Jie Zhang:** Provided valuable insights and suggestions throughout the research process, Participated in the discussion and revision of the manuscript. **Lixin Tang:** Provided valuable insights and suggestions throughout the research process, Oversaw the project, Provided guidance, Played a significant role in the finalization of the manuscript, Participated in the discussion and revision of the manuscript. **Bin Li:** Provided valuable insights and suggestions throughout the research process, Participated in the discussion and revision of the manuscript. **Hao Wang:** Provided valuable insights and suggestions throughout the research process, Participated in the discussion and revision of the manuscript.

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

Data will be made available on request.

Acknowledgment

All authors approved the version of the manuscript to be published. The authors would like to thank the editors and the anonymous reviewers for their valuable comments and suggestions, which have led to a substantial improvement of this paper. This work was supported by the China Scholarship Council 202206160060.

References

Andrychowicz, M., Denil, M., Gómez, S., Hoffman, M.W., Pfau, D., Schaul, T., Shillingford, B., de Freitas, N., 2016. Learning to learn by gradient descent by gradient descent. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 29. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2016/file/fb87582825f9d28a8d42c5e5e5e8b23d-Paper.pdf.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Bao, Y., Song, K., Liu, J., Wang, Y., Yan, Y., Yu, H., Li, X., 2021. Triplet-graph reasoning network for few-shot metal generic surface defect segmentation. IEEE Trans. Instrum. Meas. 70, 1–11.

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. Int. J. Comput. Vis. 129 (4), 1038–1059.

Božič, J., Tabernik, D., Skočaj, D., 2021. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. Comput. Ind. 129, 103459.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2023. Swin-unet: Unet-like pure transformer for medical image segmentation. In: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III. Springer, pp. 205–218.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Dong, H., Song, K., He, Y., Xu, J., Yan, Y., Meng, Q., 2019. PGA-net: Pyramid feature fusion and global context attention network for automated surface defect detection. IEEE Trans. Ind. Inform. 16 (12), 7448–7458.

Dong, H., Song, K., Wang, Q., Yan, Y., Jiang, P., 2021. Deep metric learning-based for multi-target few-shot pavement distress classification. IEEE Trans. Ind. Inform. 18 (3), 1801–1810.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.

Douze, M., Szlam, A., Hariharan, B., Jégou, H., 2018. Low-shot learning with large-scale diffusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3349–3358.

Finn, C., Abbeel, P., Levine, S., 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 70, PMLR, pp. 1126–1135, URL https://proceedings.mlr.press/v70/finn17a.html.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3146–3154.

Gao, H., Shou, Z., Zareian, A., Zhang, H., Chang, S.-F., 2018. Low-shot learning via covariance-preserving adversarial augmentation networks. Adv. Neural Inf. Process. Syst. 31.

Gao, L., Zhang, J., Yang, C., Zhou, Y., 2022. Cas-VSwin transformer: A variant swin transformer for surface-defect detection. Comput. Ind. 140, 103689.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d., 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV.

Guo, M.-H., Liu, Z.-N., Mu, T.-J., Hu, S.-M., 2022a. Beyond self-attention: External attention using two linear layers for visual tasks. IEEE Trans. Pattern Anal. Mach. Intell.

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R.R., Cheng, M.-M., Hu, S.-M., 2022b. Attention mechanisms in computer vision: A survey. Comput. Vis. Media 8 (3), 331–368.

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.

Huang, Z., Wang, X., Wei, Y., Huang, L., Shi, H., Liu, W., Huang, T., 2020. CCNet: Criss-cross attention for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell.

Jin, Q., Meng, Z., Pham, T.D., Chen, Q., Wei, L., Su, R., 2019. DUNet: A deformable network for retinal vessel segmentation. Knowl.-Based Syst. 178, 149–162.

Koch, G., Zemel, R., Salakhutdinov, R., et al., 2015. Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop, Vol. 2, No. 1. Lille.

Kwon, D., Ahn, J., Kim, J., Choi, I., Jeong, S., Lee, Y.-S., Park, J., Lee, M., 2019. Siamese U-Net with healthy template for accurate segmentation of intracranial hemorrhage. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. Springer, pp. 848–855.

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.

Ling, Z., Zhang, A., Ma, D., Shi, Y., Wen, H., 2022. Deep siamese semantic segmentation network for PCB welding defect detection. IEEE Trans. Instrum. Meas. 71, 1–11.

Liu, B., Wang, X., Dixit, M., Kwitt, R., Vasconcelos, N., 2018. Feature space transfer for data augmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9090–9098.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Lu, X., Wang, W., Shen, J., Crandall, D., Luo, J., 2020. Zero-shot video object segmentation with co-attention siamese networks. IEEE Trans. Pattern Anal. Mach. Intell. 44 (4), 2228–2242.

Luo, X., Li, S., Wang, Y., Zhan, T., Shi, X., Liu, B., 2023. MaMiNet: Memory-attended multi-inference network for surface-defect detection. Comput. Ind. 145, 103834.

Ma, S., Song, K., Niu, M., Tian, H., Wang, Y., Yan, Y., 2023. Shape consistent one-shot unsupervised domain adaptation for rail surface defect segmentation. IEEE Trans. Ind. Inform.

Mnih, V., Heess, N., Graves, A., et al., 2014. Recurrent models of visual attention. Adv. Neural Inf. Process. Syst. 27.

Niu, T., Li, B., Li, W., Qiu, Y., Niu, S., 2022. Positive-sample-based surface defect detection using memory-augmented adversarial autoencoders. IEEE/ASME Trans. Mechatronics 27 (1), 46–57. http://dx.doi.org/10.1109/TMECH.2021.3058147.

Niu, S., Li, B., Wang, X., Peng, Y., 2021. Region-and strength-controllable GAN for defect generation and segmentation in industrial images. IEEE Trans. Ind. Inform. 18 (7), 4531–4541.

Ren, X., Lin, W., Yang, X., Yu, X., Gao, H., 2022. Data augmentation in defect detection of sanitary ceramics in small and Non-i.i.d datasets. IEEE Trans. Neural Netw. Learn. Syst. 1–10. http://dx.doi.org/10.1109/TNNLS.2022.3152245.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P., 2022b. Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 14318–14328.

Tao, X., Gong, X., Zhang, X., Yan, S., Adak, C., 2022. Deep learning for unsupervised anomaly localization in industrial images: A survey. IEEE Trans. Instrum. Meas.

Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., Jia, J., 2020. Prior guided feature enrichment network for few-shot segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 44 (2), 1050–1065.

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., Wierstra, D., 2016. Matching networks for one shot learning. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 29. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017. Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3156–3164.

Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M., 2020. Generalizing from a few examples: A survey on few-shot learning. ACM Comput. Surv. (Csur) 53 (3), 1–34.

Yun, J.P., Shin, W.C., Koo, G., Kim, M.S., Lee, C., Lee, S.J., 2020. Automated defect inspection system for metal surfaces based on deep learning and data augmentation. J. Manuf. Syst. 55, 317–324. http://dx.doi.org/10.1016/j.jmsy.2020.03.009, URL https://www.sciencedirect.com/science/article/pii/S027861252030042X.

Zhan, Z., Zhou, J., Xu, B., 2022. Fabric defect classification using prototypical network of few-shot learning algorithm. Comput. Ind. 138, 103628.

Zhang, Z., Peng, H., 2019. Deeper and wider siamese networks for real-time visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4591–4600.

Zhang, Y., Yang, Q., 2022. A survey on multi-task learning. IEEE Trans. Knowl. Data Eng. 34 (12), 5586–5609. http://dx.doi.org/10.1109/TKDE.2021.3070203.

Zhuxi, M., Li, Y., Huang, M., Huang, Q., Cheng, J., Tang, S., 2022. A lightweight detector based on attention mechanism for aluminum strip surface defect detection. Comput. Ind. 136, 103585.