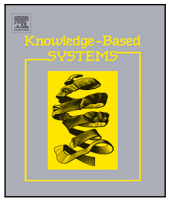
Knowledge-Based Systems 287 (2024) 111451

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Feature matching driven background generalization neural networks for surface defect segmentation

Biao Chen [a], Tongzhi Niu [a,*], Ruoqi Zhang [c], Hang Zhang [a], Yuchen Lin [a], Bin Li [a,b]

[a] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuahn, 430074, Hubei, China
[b] Wuhan Intelligent Equipment Industrial Institute Co., Ltd, 8 Ligou South Road, Wuahn, 430074, Hubei, China
[c] China-EU Institute for Clean and Renewable Energy, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuahn, 430074, Hubei, China

ARTICLE INFO

Keywords:
Surface defect detection
Neural networks
Feature matching
Background generalization

ABSTRACT

In this paper, we address the challenge of background generalization in surface defect segmentation for surface-mounted device chips, particularly focusing on template-sample comparison algorithms. These algorithms often struggle with background features in templates and samples that exhibit spatial variations, including translation and rotation. The inherent spatial equivariance in CNN-based algorithms complicates the elimination of noise attributed to these spatial variations. To address this issue, we developed the Background Generalization Networks (BGNet). BGNet effectively reduces spatial variation noise by subtracting background features of samples and templates based on their matching relationships. It starts by extracting dense features rich in global and interactive information via a Siamese network and then applies self- and cross-attention mechanisms from Transformers. The matching score is calculated based on feature similarity, with matching relations established using the Mutual Nearest Neighbour (MNN) algorithm. These relations enable us to mitigate the noise caused by spatial variations and implement a multiscale fusion of detailed and semantic information, leading to more accurate segmentation results. Our experiments on OCDs and PCBs have shown that BGNet surpasses existing state-of-the-art methods in terms of performance. Furthermore, the code for this work is available on GitHub: https://github.com/Max-Chenb/BG-Net.

1. Introduction

In recent years, surface defect detection technology based on deep learning has been widely researched and applied in industries such as semiconductors [1], aerospace [2], transportation [3,4], and textiles [5]. Extensive and comprehensive studies have addressed challenges such as data imbalance [6], inconsistent label [7], multiple scales and shapes [8], and significant intraclass variations versus minor interclass differences [9]. This paper focuses on an emerging issue: achieving batch-to-batch background generalization through template-sample comparison [10–13].

The success of traditional deep learning relies on the assumption that training and testing datasets share the same distribution. However, in surface defect detection for chips of surface-mounted devices (including printed circuit boards (PCBs) and optical communication devices (OCDs)), variations in device types and distribution across different batches can lead to distinct data distributions. If defects are defined as foreground and nondefects as background, then the challenge of distribution inconsistency can be described as background generalization. Existing methods [10–13] aim to learn how to compare the differences between the template and the samples. For new batches, generalization can be achieved by collecting templates. The primary challenge with this approach is the noise that arises from inconsistencies in device and fabrication processes, which contributes to the background variations between the template and samples beyond just the defect foreground features.

Typically, the background features of templates and samples exhibit spatial variations such as translation, rotation, and scale, in addition to texture variations such as colour changes, due to variations in device types and fabrication processes, as illustrated in Fig. 1. For texture variations, existing CNNs have the powerful ability to extract features and demonstrate robust performance. However, in regard to spatial variations, the spatial invariance of methods based on CNNs is limited [14].

Existing methods such as Siamese UNet [15], DSSSNet [11], and GWNet [11] have shown some effectiveness in handling spatial variations in surface defect segmentation. However, their operational mechanisms and design principles present certain limitations. Siamese UNet, which relies on direct feature subtraction across layers, may not effectively address complex spatial transformations. DSSSNet achieves spatial invariance predominantly through global pooling, a technique that

* Corresponding author.
E-mail address: tzniu@hust.edu.cn (T. Niu).

https://doi.org/10.1016/j.knosys.2024.111451
Received 1 August 2023; Received in revised form 12 January 2024; Accepted 26 January 2024

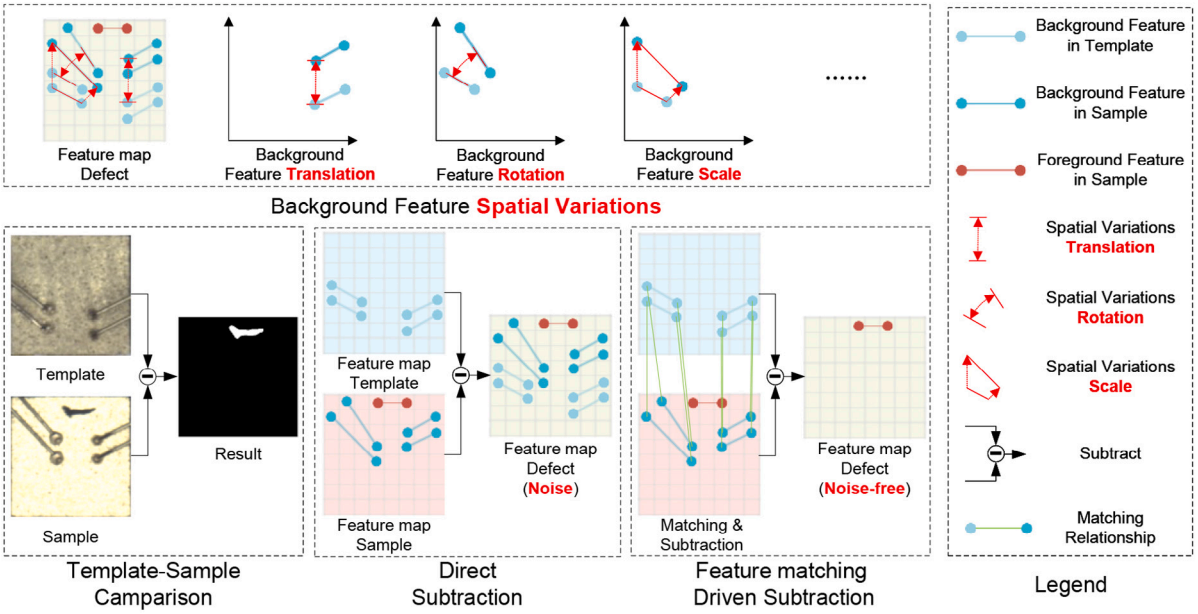


Fig. 1. Spatial variation noise and different subtraction comparisons.

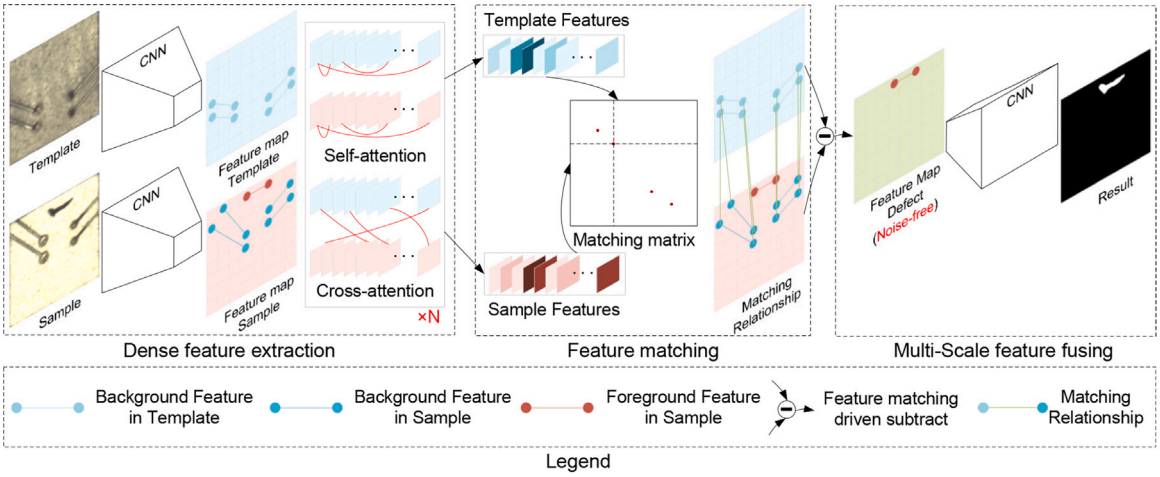


Fig. 2. Feature matching driven background generalization neural networks for surface defect segmentation.

can oversimplify spatial arrangements and miss finer details. GWNet, which incorporates self- and cross- attention mechanisms from Transformers, is an innovative approach but may not comprehensively address the complex spatial relationships crucial for background generalization. These observations highlight the need for a more robust solution that can thoroughly manage spatial variations, a gap our proposed BGNet aims to fill.

In this paper, a transparent and accountable Background Generalization Networks (BGNet) based on feature matching is introduced, as illustrated in Fig. 2. Despite spatial variations in displacement, rotation, and mapping among the background features of the template and sample, a one-to-one correlation persists. Therefore, a naive idea is to compute the matching relationship and adjust feature positions accordingly, allowing for one-to-one subtraction and achieving background generalization.

Given the unavailability of interest point labels, obtaining dense features equipped with matching information through a Convolutional Neural Network (CNN) based Siamese network is an initial step. However, these dense features have a limited receptive field and lack interrelated information between the template and sample. To address this, existing detect-free local feature matching methods [16–19] use a Transformer [20] to obtain the global context through self-attention and interrelated information through cross-attention. Multiple iterations of self- and cross-attention are layered for a more robust representation.

After obtaining dense features of both the template and sample, similarity measures, as recommended in [21], are employed to calculate matching scores among feature points and derive a matching matrix. This matrix is preprocessed using a dual-softmax function [16] with a threshold for isolating and matching significant features. Matching relationships are ultimately determined using a mutual nearest neighbour (MNN) strategy.

Based on the matching relationships derived, matching features in the sample are subtracted from those in the template to acquire noise-free defect foreground features. A multiscale fusion technique is employed to achieve more accurate segmentation. However, considering computational demands, the feature matching network is applied separately to detailed (1/32 scale) and semantic information (1/8 scale) within BGNet, as suggested by BiSeNet [22].

The performance of BGNet was evaluated using surface-mounted device chip datasets for both the OCDs [13] and PCBs [11] datasets.

The experimental results shown that BGNet outperforms current state-of-the-art techniques. Visualization experiments demonstrated BGNet's ability to accurately identify features undergoing spatial transformations within the background. After successful matching and subtraction, only foreground defect features remain.

To summarize, the main contributions of this paper are as follows:

- A novel feature matching driven Background Generalization Network (BGNet) for surface defect segmentation is proposed that achieves spatial invariance in template-sample comparisons.
- A Transformer-based dense feature extraction method coupled with a MNN algorithm for feature matching was proposed to calculate the correspondence relationships between features.
- The accuracy and robustness of BGNet are confirmed through comparisons with 16 state-of-the-art methods across two datasets.

2. Related work

2.1. Surface defect detection

In recent years, the issue of data imbalance in surface defect detection has received widespread attention. Anomaly detection [23] algorithms based on positive samples, data generation algorithms [24], domain adaption [25], and generalizable algorithms for new foreground (defect) types and background (defect-free) types have been extensively researched. This paper focuses on the study of generalizable algorithms and provides a detailed introduction to both foreground and background generalization.

2.1.1. Foreground generalization

Several novel approaches have been proposed for surface defect classification. The graph embedding and distribution transformation (GEDT) model [26], in combination with the optimal transport (OPT) module, can identify new defect classes even with a limited number of labelled samples. The FSDR approach [27] advances a coarse-to-fine few-shot defect classification strategy that employs dynamic weighting and joint metrics, easing the data collection process and enabling the classification of novel defect categories. FaNet [28] introduces a feature-attention convolution module that excels at extracting comprehensive feature details from base classes while enhancing semantic integration by capitalizing on long-range feature interconnections.

In the context of surface defect segmentation, several notable methodologies have emerged. TGRNet [29] applies few-shot learning theory to generic metal surface defect segmentation and devises a C-way N-shot W-normal learning method that includes a surface defect triplet to independently segment the background and defect areas. It also incorporates a multigraph reasoning module to explore similarity relationships among different images. Simultaneously, OBFTNet [30] introduces background images as supplementary learning information and treats few-shot segmentation as an optimal bilateral transport problem, adaptively generating task-specific semantic correspondences to ensure the model's ability to generalize to unseen materials. Recently, a comparative dataset known as Industrial-5^i [31] has been constructed using public datasets.

2.1.2. Background generalization

In some flexible production lines, particularly with chips of surface-mounted devices, the types of defect foregrounds rarely increase, while the backgrounds vary with batch changes. As a result, background generalization is a valuable research topic.

DSSSNet [11] establishes a deep Siamese semantic segmentation network by combining the similarity measurement capabilities of the Siamese network with an encoder–decoder semantic segmentation network, resulting in an effective tool for PCB welding defect detection. Concurrently, SC-OSDA [12] presents a shape-consistent style transfer module to address the issue of insufficient target domain samples by performing pixel-level distribution alignment between training and test images. This approach, which requires only a single target domain sample, significantly enhances the model's robustness to domain shifts. GWNet [13] introduces a dual-attention mechanism (DAM) for feature extraction and a recurrent residual attention mechanism (RRAM) for feature fusion, enabling the model to effectively generalize to new batches of unseen data during training by utilizing collected templates.

In summary, adapting models to new defects or data is a significant challenge, with current methods still being explored and not yet ready for practical implementation. Given the consistent nature of defect features, background generalization is a more feasible and practical approach at this stage, particularly in the context of flexible production lines. This paper proposes an explicit and explainable method for this task, building upon prior research.

2.2. Local feature matching

In general, local feature matching between images is the foundation of many 3D computer vision tasks, including structure from motion, simultaneous localization and mapping, and visual localization. Image matching methods typically use a three-stage process: feature detection, description, and matching. In the detection stage, significant points are identified in each image. Local descriptors are then extracted from the areas around these points. The result is two sets of descriptors whose correspondences are established using nearest neighbour searches or advanced matching algorithms. Based on these stages, existing techniques can be divided into two categories: detector-based and detector-free local feature matching methods.

2.2.1. Detector-based local feature matching

Before the advent of deep learning, handcrafted methods were often based on SIFT [32] and ORB [33]. SIFT characterizes distinctive keypoints by constructing a high-dimensional vector that represents the image gradients within a localized region of the image. ORB proposes an extremely fast binary descriptor based on BRIEF [34], which is two orders of magnitude faster than SIFT. Notably, both ORB and SIFT demonstrate rotation invariance and robustness to noise.

Due to their powerful feature extraction capabilities, deep learning-based methods significantly improve performance under substantial viewpoint and illumination changes. LIFT [35] was the first to introduce an end-to-end differentiable complete feature point handling pipeline, which includes detection, orientation estimation, and feature description. Most recent research [36–39] on deep learning for matching has typically focused on learning superior sparse detectors and local descriptors from data using CNNs.

However, methods based on CNNs typically use the nearest neighbour search to find matches among the extracted points of interest. SuperGlue [21] learns matches with a graph neural network (GNN), which is a generalized form of Transformer [20]. Although SuperGlue demonstrates impressive performance, it fails to detect repeatable points of interest in indistinct regions.

2.2.2. Detector-free local feature matching

Detector-free methods bypass the feature detection phase and directly generate dense descriptors or dense feature matches. SIFT Flow [40] was the first to propose pixelwise SIFT features between two images while preserving spatial discontinuities.

In NCNet [41], exhaustive pairwise cosine similarities between two dense feature descriptors are computed and stored in a 4D tensor known as a correlation map. This map is subsequently input into a neighbourhood consensus CNN (4D-CNN), which learns dense correspondences by regularizing the cost volume and enforcing neighbourhood consensus among all matches. Following this line of work, Sparse-NCNet [42] employs sparse convolutions to improve efficiency. Moreover, DRC-Net [43] combines multiscale information in a coarse-to-fine approach.

Like detector-based methods, the aforementioned detector-free methods rely solely on local features to obtain descriptors. By utilizing both self- and cross-attention layers within the Transformer and repeatedly interleaving these layers, LoFTR [16] generates feature descriptors that are conditioned on both images, learning densely arranged globally consented matching priors inherent in ground-truth matches. In addition, transfusion [17] and GMFlow [18] designs matching algorithms based on the Transformer. However, these works have rarely focused on the scale difference between image pairs. PATS [19] proposes patch area transportation with subdivision to obtain a significantly larger and more accurate number of matches. Additionally, soft matching [44] introduces a learning-based soft template matching network tailored for defect detection that incorporates an innovative attention mechanism.

This paper focuses on matching background features between templates and samples, which exhibit spatial variations. Despite the lack of interest point annotations, we built upon previous research on detector-free local feature matching and proposed a background feature matching algorithm.

3. Methodology

3.1. Problem definition

This paper focuses on the challenge of background generalization, particularly in template-sample matching algorithms that address spatial variations in template and sample background features, including aspects such as translation, rotation and mapping. Given an image pair consisting of a template I^T and a sample I^S, they are input into a Siamese network, resulting in corresponding features at five different scales, denoted as $\{F_i^T\}_{i=1}^5$ and $\{F_i^S\}_{i=1}^5$. The feature map is represented as $F=(f_{x,y})\in\mathbb{F}^{C\times H\times W}$, where C, H, and W represent the channel, height, and width of the feature map, respectively. The terms x and y correspond to the coordinates of a specific feature f within the feature map F. The feature maps of template F^T, sample F^S, sample with translation $\overline{F}^S$, and sample with rotation $\widetilde{F}^S$ are represented as follows:

$$F^T=\begin{bmatrix} f_{1,1}^T & f_{1,2}^T & f_{1,3}^T & \cdots & f_{1,W}^T \\ f_{2,1}^T & f_{2,2}^T & f_{2,3}^T & \cdots & f_{2,W}^T \\ f_{3,1}^T & f_{3,2}^T & f_{3,3}^T & \cdots & f_{3,W}^T \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{H,1}^T & f_{H,2}^T & f_{H,2}^T & \cdots & f_{H,W}^T \end{bmatrix} \tag{1}$$

$$F^S=\begin{bmatrix} f_{1,1}^S & f_{1,2}^S & f_{1,3}^S & \cdots & f_{1,W}^S \\ f_{2,1}^S & f_{2,2}^S & f_{2,3}^S & \cdots & f_{2,W}^S \\ f_{3,1}^S & \mathbf{d_{3,2}} & \mathbf{d_{3,3}} & \cdots & f_{3,W}^S \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{H,1}^S & f_{H,2}^S & f_{H,2}^S & \cdots & f_{H,W}^S \end{bmatrix} \tag{2}$$

$$\overline{F}^S=\begin{bmatrix} f_{1,1}^S & \mathbf{f_{2,2}^S} & f_{1,3}^S & \cdots & f_{1,W}^S \\ \mathbf{f_{3,1}^S} & \mathbf{f_{1,2}^S} & f_{2,3}^S & \cdots & f_{2,W}^S \\ \mathbf{f_{2,1}^S} & d_{3,2} & d_{3,3} & \cdots & f_{3,W}^S \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{H,1}^S & f_{H,2}^S & f_{H,2}^S & \cdots & f_{H,W}^S \end{bmatrix} \tag{3}$$

$$\widetilde{F}^S=\begin{bmatrix} \mathbf{f_{1,3}^S} & f_{1,2}^S & \mathbf{f_{1,1}^S} & \cdots & f_{1,W}^S \\ \mathbf{f_{2,2}^S} & \mathbf{f_{2,1}^S} & f_{2,3}^S & \cdots & f_{2,W}^S \\ \mathbf{f_{3,1}^S} & d_{3,2} & d_{3,3} & \cdots & f_{3,W}^S \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{H,1}^S & f_{H,2}^S & f_{H,2}^S & \cdots & f_{H,W}^S \end{bmatrix} \tag{4}$$

where d represents the defective feature. The utilization of an underline (as seen in $\underline{f}$ and $\underline{d}$) indicates a change in feature location or type. Values in boldface (in **f** and **d**) denote the results of defects, translation, and rotation processes.

This study acknowledges that spatial variation, resulting in shifts in the positions of background features in both the template and sample, renders direct subtraction ineffective. The primary focus of this paper is the development of a technique that matches these dynamic background features, enables corresponding subtractions, and thus produces more accurate segmentation results.

3.2. Framework overview

As shown in Fig. 3, BGNet consists of three components: dense feature extraction, feature matching, and multiscale feature fusion.

In the dense feature extraction section, the template image I^T and sample image I^S are input into the Siamese network to obtain feature maps of different scales $\{F_i^T\}_{i=1}^5$ and $\{F_i^S\}_{i=1}^5$, where $F_i=(f_{x,y})\in\mathbb{F}^{C_i\times H_i\times W_i}$. Then, dense features $\hat{F}_i^T$ and $\hat{F}_i^S$ are extracted specifically from the 1/8 ($i=3$) and 1/32 ($i=5$) scaled features. After applying position encoding to the feature map, it is fed into multiple layers of self-attention $Atten_{self}(\cdot,\cdot)$ and cross-attention $Atten_{cross}(\cdot,\cdot)$ mechanisms.

$$\hat{F}_i^T=Atten_{cross}\left(Atten_{self}\left(F_i^T,F_i^T\right),F_i^S\right) \tag{5}$$

$$\hat{F}_i^S=Atten_{cross}\left(Atten_{self}\left(F_i^S,F_i^S\right),F_i^T\right) \tag{6}$$

In the feature matching section, a dual-softmax operation is employed to derive the matching similarity matrix. The MNN algorithm is then utilized to establish the matching relationships between feature points. Direct subtraction of the corresponding features at the 1/8 and 1/32 scales is performed to eliminate deformation noise from the feature maps. The matching algorithm $Match(\cdot,\cdot)$ is represented as $\mathcal{F}_i^M: M_i^S\rightarrow M_i^T$, where $M_i^S=\left\{\left(x_j^S,y_j^S\right)\right\}_{j=1}^J$ and $M_i^T=\left\{\left(x_j^T,y_j^T\right)\right\}_{j=1}^J$. This implies that the features $f_{x_j^S,y_j^S}^S$ and $f_{x_j^T,y_j^T}^T$ correspond to each other in a one-to-one matching relationship. Here, J denotes the number of matching features.

$$\mathcal{F}_i^M=Match\left(\hat{F}_i^T,\hat{F}_i^S\right) \tag{7}$$

Then, the noise-free features $F_i^D=\left(f_{x,y}^D\right)\in\mathbb{F}^{C_i\times H_i\times W_i}$ are eliminated by utilizing the matching relationship for the corresponding subtractions.

$$f_{x,y}^D=\begin{cases} f_{x^S,y^S}^S & x^S,y^S\notin M_i^S \\ f_{x^S,y^S}^S-f_{x^T,y^T}^T & x^S,y^S\in M_i^S \end{cases} \tag{8}$$

In the multiscale feature fusion process, to balance computational complexity, the feature matching network is independently applied to both detailed information (1/8 scale) and semantic information (1/32 scale). For the other scales (1/2, 1/4, and 1/16), direct subtraction is employed. These multiscale features are fused through a skip-connection approach.

3.3. Dense feature extraction

3.3.1. Siamese network

This study employs a Siamese network to extract features from both the template and sample, which consists of two subnetworks with shared weights. Resnet-18 is used as the subnetwork, and pretraining weights based on ImageNet are utilized during the training process. The template I^T and sample I^S are input into the Siamese network, resulting in corresponding features at five different scales $\{F_i^T\}_{i=1}^5$ and $\{F_i^S\}_{i=1}^5$.

3.3.2. Positional encoding

In contrast to CNNs, Transformers input the entire feature map simultaneously, leading to the loss of inherent positional information in

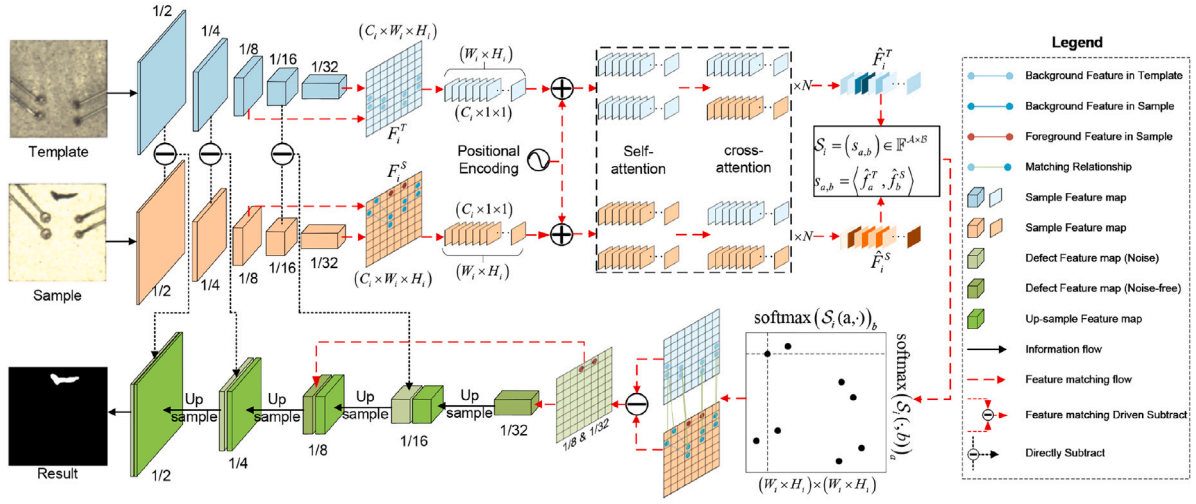


Fig. 3. The BGNet architecture.

the image. To ensure appropriate matching of background features, this study augments the feature map with positional encoding information. Unlike previous methodologies, this work focus on adding positional encoding to two-dimensional feature maps.

In our research, we utilize the 2D sinusoidal position encoding technique. This method involves mapping each position within a two-dimensional space onto a uniquely defined vector characterized by a specific mathematical pattern. The core of this encoding process hinges on the strategic use of sine and cosine functions. These functions exhibit variations in frequency that are systematically distributed across the different elements of the encoding vector. This approach allows for the effective capture and representation of positional information within a two-dimensional context, which is essential for the accurate processing and interpretation of spatial data. Consequently, the positional encoding $p_{x,y} \in \mathbb{F}^{C_i \times H_i \times W_i}$ is defined as follows:

$$p_{x,y}^{(c)} = \begin{cases} sin\left(x \times \frac{1}{10000^{2k/C_i}}\right) & c = 4k \\ cos\left(x \times \frac{1}{10000^{2k/C_i}}\right) & c = 4k+1 \\ sin\left(y \times \frac{1}{10000^{2k/C_i}}\right) & c = 4k+2 \\ cos\left(y \times \frac{1}{10000^{2k/C_i}}\right) & c = 4k+3 \end{cases} \tag{9}$$

where k ranges from 0 to $C_i/4 - 1$, x ranges from 1 to H_i (height indices), and y ranges from 1 to W_i (width indices). Here, c represents the dimension of the channel, and C_i represents the dimension of the position encoding.

Subsequently, the feature map, after adding positional encoding, is expressed as follows:

$$F_i = \left(f_{x,y} + p_{x,y}\right) \in \mathbb{F}^{C_i \times H_i \times W_i} \tag{10}$$

3.3.3. Self- and cross-attention mechanism

After adding the positional encoding, local feature maps of the template and sample F_i^T and F_i^S are input into self- and cross-attention mechanisms, respectively, to extract global and interactive information. Consequently, dense feature maps $\hat{F}_i^T$ and $\hat{F}_i^S$ are derived.

First, F_i^T, F_i^S are resized to $F_i^T, F_i^S \in \mathbb{F}^{L_i \times C_i}$, where $L_i = H_i \times W_i$.

Next, as shown in Fig. 4, the mechanism of self- and cross-attention is depicted in the diagram. Due to the difference in inputs to self- and cross-attention, $F', F'' \in \mathbb{F}^{L_i \times C_i}$ are used for representation. $Q, K, V \in \mathbb{F}^{L \times \hat{C}}$ are computed by fully connected networks W^Q, W^K, W^V, and $C \neq \hat{C}$.

$$\begin{aligned} Q &= W^Q F' = \left(W^Q f'_{x,y}\right) \in \mathbb{F}^{L \times \hat{C}} \\ K &= W^K F'' = \left(W^K f''_{x,y}\right) \in \mathbb{F}^{L \times \hat{C}} \\ V &= W^V F'' = \left(W^V f''_{x,y}\right) \in \mathbb{F}^{L \times \hat{C}} \end{aligned} \tag{11}$$

Referencing the Linear Transformer [45], $Atten(Q, K, V)$ is defined as follows:

$$Atten(Q, K, V) = \phi(Q)\left(\phi(K)^\top V\right) \tag{12}$$

where $\phi(\cdot) = elu(\cdot) + 1$.

Then, the attention map $\hat{F}'$ is obtained by concatenating and applying a residual operation to Q. Before the concatenation and residual operations, there are fully connected networks W_{L1}, W_{L2} and normalization.

To compute the self-attention features of F_i^T, let $F' = F_i^T, F'' = F_i^T$. To compute the cross-attention features of F_i^T, let $F' = F_i^T, F'' = F_i^S$. Conversely, the computations for the self- and cross-attention features of F_i^S are performed similarly.

Finally, the dense feature maps of the template $\hat{F}_i^T$ and sample $\hat{F}_i^S$ are calculated as per Eqs. (5) and (6). To endow the model with stronger representational capacity, this study adopts a strategy similar to SuperGlue [21] and LoFTR [16], stacking multiple instances of self- and cross-attention. For dense feature extraction at different scales, different numbers of computations are stacked. More stacks (4 stacks on the 1/32 layer) are used for semantic features (deep features), while fewer stacks (2 stacks on the 1/8 layer) are used for detailed features (shallow features).

3.4. Feature matching

In the preceding section, the dense feature maps of the template and sample, denoted as $\hat{F}_i^T = \left(\hat{f}_a^T\right) \in \mathbb{F}^{L_i \times C_i}$ and $\hat{F}_i^S = \left(\hat{f}_b^S\right) \in \mathbb{F}^{L_i \times C_i}$, respectively, are obtained, where $\hat{f}_a^T, \hat{f}_b^S \in \mathbb{F}^C$ and $a, b \in \mathcal{A}, \mathcal{B}, \mathcal{A} = \mathcal{B} = [1, 2, \ldots, L]$.

First, the similarity of matching descriptors is expressed as a score matrix $\mathcal{S}_i = \left(s_{a,b}\right) \in \mathbb{F}^{\mathcal{A} \times \mathcal{B}}$:

$$s_{a,b} = \left\langle \hat{f}_a^T, \hat{f}_b^S \right\rangle, \forall (a, b) \in \mathcal{A} \times \mathcal{B} \tag{13}$$

where $\langle \cdot, \cdot \rangle$ is the inner product.

Then, a dual-softmax operator [41] is applied on both dimensions of $\mathcal{S}_i$ to obtain the probability of soft MNN matching. The matching probability $\mathcal{P}_i = \left(p_{a,b}\right) \in \mathbb{F}^{\mathcal{A} \times \mathcal{B}}$ is obtained by:

$$\mathcal{P}_i(a, b) = softmax\left(\mathcal{S}_i(a, \cdot)\right)_b \cdot softmax\left(\mathcal{S}_i(\cdot, b)\right)_a \tag{14}$$

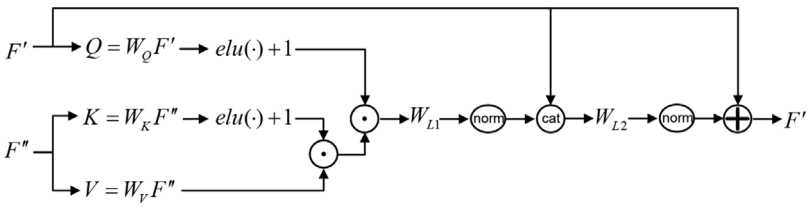


Fig. 4. The self- and cross-attention mechanism.

Based on the matching probability $\mathcal{P}_i$, potential background matching features are selected by enforcing the MNN criterion:

$$\mathcal{M}_i = \left\{ \left(\widetilde{a}, \widetilde{b}\right) | \forall \left(\widetilde{a}, \widetilde{b}\right) \in MNN\left(\mathcal{P}_i\right), \mathcal{P}_i\left(\widetilde{a}, \widetilde{b}\right) \geqslant \theta \right\} \tag{15}$$

where $\mathcal{M}_i = \left\{ \left(\widetilde{a}, \widetilde{b}\right) \right\}_{j=1}^{N}$ represents the matching pairs. The pseudocode for MNN can be found in Algorithm 1. Additionally, a threshold of θ is applied to filter out noise and maintain high confidence matches. In this paper, θ is an empirical parameter, which we set to 0.2.

Algorithm 1 Mutual Nearest Neighbour (MNN) Algorithm

Require: Distance matrix $\mathcal{P}$, Index sets $\mathcal{A}$, $\mathcal{B}$
Ensure: Set of matching pairs $\mathcal{M}$
Initialize $\mathcal{M} \leftarrow \emptyset$
for $a \in \mathcal{A}$ **do**
  $b \leftarrow \arg\min_{k \in \mathcal{B}}(\mathcal{P}_{a,k})$
  if $a == \arg\min_{l \in \mathcal{A}}(P_{l,b})$ **then**
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{(a, b)\}$
  end if
end for
return $\mathcal{M}$

Finally, a one-to-one correspondence between the template and sample features $\mathcal{F}_i^M : M_i^S \rightarrow M_i^T$ is obtained based on $\mathcal{M}_i$, where $M_i^S = \left\{ \left(x^S, y^S\right)_j \right\}_{j=1}^{N}$ and $M_i^T = \left\{ \left(x^T, y^T\right)_j \right\}_{j=1}^{N}$. The noise-free features F_i^D are eliminated by Eq. (8).

3.5. Multifeature fusion

Multiscale fusion is an effective method for improving segmentation accuracy according to existing methodologies [46,46–49]. Five scales of features are obtained through the Siamese network. However, matching each scale would lead to significant computational overhead. In reference to BiSeNet [22], fusing detailed and semantic features not only reduces computational costs but also enhances accuracy. Therefore, this study solely achieves noise-free feature maps through matching at the 1/8 and 1/32 scales, while direct subtraction is applied at other scales. Ultimately, multiscale fusion is conducted in a manner similar to that of UNet, as shown in Fig. 3.

3.6. Loss function

In surface defect detection images, the foreground is sparse compared to the background. Therefore, this study employs focal loss [50] as the loss function.

The focal loss is a loss function designed specifically to address the class imbalance problem in one-stage object detection. It has proven to be effective at assigning more importance to hard-to-classify instances. The focal loss is designed to add a modulating factor to the standard cross entropy criterion to downweight easy examples and thus focus training on hard negatives. The focal loss is defined as follows:

$$FL(p, y) = \begin{cases} -(1-p)^{\gamma} log(p) & \text{if } y = 1 \\ -p^{\gamma} log(1-p) & \text{otherwise} \end{cases} \tag{16}$$

where p is the model's estimated probability for the class with label y and γ is the focusing parameter that should be greater than 0. In this paper, $\gamma = 2$.

4. Experiments and results

4.1. Experimental setup

4.1.1. Implementation details

Employing an NVIDIA GeForce RTX 4090 GPU facilitated efficient data processing, which is ideal for complex machine learning tasks. A PyTorch-based model was utilized and optimized via the Adam optimizer set at a learning rate of 10e-5. This arrangement ensured a balance between convergence speed and training stability. Further optimization occurred through data processing in minibatches of eight, enabling superior GPU utilization and accelerated model updates.

4.1.2. Evaluation metrics

In this paper, we use six key metrics: precision (Pre), recall (Rec), F-measure (F2), mean intersection over union (mIoU), mean accuracy (mACC), and parameter size (MB).

$$Pre = \frac{TP}{TP + FP} \tag{17}$$

$$Rec = \frac{TP}{TP + FN} \tag{18}$$

$$F2 = (1 + 2^2) \cdot \frac{Pre \cdot Rec}{(2^2 \cdot Pre) + Rec} \tag{19}$$

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{20}$$

$$mACC = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i} \tag{21}$$

where TP represents the number of true positives, FP represents the number of false positives, FN represents the number of false negatives, and N represents the total number of classes.

4.1.3. Dataset description

In this paper, we focus on the challenge of background generalization for chips of surface-mounted devices, such as OCDs and PCBs. In these cases, background features in templates and samples exhibit spatial variations, such as shifts and rotations.

OCDs are devices that convert optical and electrical signals in gigabit passive optical networks and optical network terminals. These devices are composed of a base, pins, and various surface-mounted device (SMD) components interconnected by jump wires. The OCDs dataset [13] contains a total of 918 datasets, including 60 instances of base crushing, 27 instances of base scratches, 375 instances of component contamination, 240 instances of component breakage, and 216 instances of varying numbers of jump wires.

PCBs serve as foundational building blocks in electronics, providing a platform that connects and supports various electronic components through conductive pathways etched from copper sheets laminated onto a nonconductive substrate. The PCB dataset [11] consists of 340 pairs of images from a PCB manufacturer. Each pair of images includes a defective image (also referred to as an NG image) and a nondefective image (alternatively known as a template image or an OK image).

Data split strategy: We adopted a data split ratio of 6:2:2 for dividing the dataset into training, validation, and test sets, ensuring this distribution across each category. Specifically, the training set is used for model training, the validation set is used for model selection, and the test set is used for evaluating the model's performance.

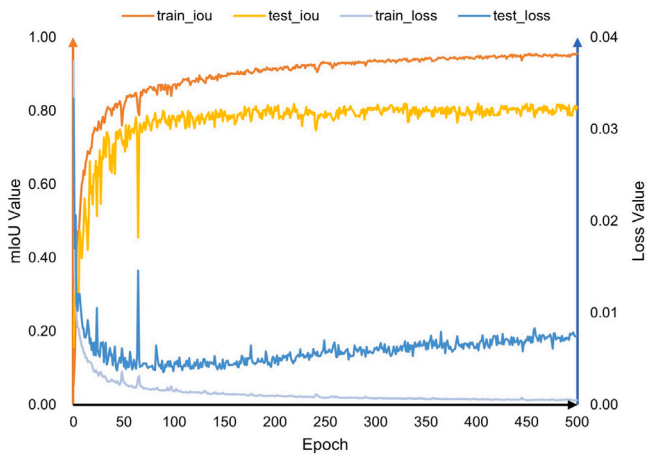


Fig. 5. Convergence performance of BGNet on the OCDs Dataset.

4.2. Convergence performance

In this section, we evaluate the convergence performance of BGNet by utilizing the OCDs dataset to verify that BGNet is not overfitting. We plot the loss and mIoU values of the training and test trajectories over the training epochs to visualize the model convergence pattern, as shown in Fig. 5.

During the progression of the model's training across 500 epochs, a distinct pattern was observed in the performance metrics. The intersection over union (IoU) for the training set demonstrated a gradual increase, indicative of the model's ability to improve the generalizability of the training data. However, the rate of this increase decelerated over time, eventually reaching a plateau. This behaviour suggested that the model's learning capacity was saturated on the training dataset. Conversely, the IoU for the test set exhibited a rapid increase, peaking at the 200th epoch. Initially, the test set's IoU showed considerable volatility, which stabilized over time, indicating that the model's robustness increased. Concerning loss metrics, a consistent decline was noted for the training set, albeit with decreasing speed, aligning with the observed trends in IoU. Intriguingly, there was a marked decrease in the loss of the test set up until the 200th epoch, followed by a gradual increase which then stabilized, maintaining a nearly constant rate thereafter. This divergence in loss trends, particularly the gradual increase in the test set loss beyond the 200th epoch, potentially signals the onset of overfitting. However, the degree of overfitting might be relatively mild, as indicated by the stabilization of the test set's IoU and the absence of a drastic divergence between the training and test loss metrics. This suggests that while the model may exhibit initial signs of overfitting after the 200th epoch, its overall performance remains robust up to this point.

4.3. Visualization of feature matching

Fig. 6 shows the class activation maps (CAMs) for the template and sample features. As depicted, BGNet accurately matches the background features between the template and the sample at the 1/32 scale, focusing exclusively on the foreground defect features after the corresponding subtraction. At the 1/8 scale, the detailed features are highly dense, and the matching relationship is generally accurate. The corresponding subtraction retains the defect features. It should also be noted that the corresponding subtraction operation eliminates only significant features, not all background features, yet it remains quite effective.

4.4. Ablation studies and discussion

4.4.1. Ablation experiment setting

Ablation experiments to validate the effectiveness of the proposed BGNet were performed as follows:

S1: Concatenate the template and sample directly and input them into a UNet-like base network, testing whether a CNN-based network inherently possesses template-sample contrast capabilities.

S2: Utilize a Siamese network in the encoder, input the template and sample into a weight-shared backbone, directly subtract the features obtained from the five scales, and achieve segmentation results after feature fusion.

S3: In the ablation study of positional encoding, features at the 1/8 and 1/32 scales are directly input into self- and cross-attention mechanisms without undergoing positional encoding, resulting in the extraction of dense features. These features are then subjected to subtraction following feature matching. Finally, the segmentation results are obtained through the fusion of these processed features.

S4: In the ablation study focusing on self- and cross-attention, features extracted from the Siamese network at the 1/8 and 1/32 scales are directly matched. Subsequent to this matching, appropriate subtractions are performed. Segmentation results are then acquired following the fusion of these processed features.

S5: In the ablation study of feature matching, dense features at the 1/8 and 1/32 scales are further extracted through the employment of self- and cross-attention. These features are then directly subtracted, and segmentation results are subsequently obtained following the process of feature fusion.

S6: In the ablation study of MNN in feature matching, dense features are extracted at the 1/8 and 1/32 scales through positional encoding, along with the implementation of self- and cross-attention. Subsequently, these features undergo subtraction after feature matching, which is performed without the MNN. The segmentation results are subsequently obtained following the fusion of these features.

S7: Extract dense features through positional encoding and self- and cross-attention at the 1/8 and 1/32 scales, subtract after feature matching, and obtain segmentation results following feature fusion. This represents the complete method proposed in this study.

4.4.2. Discussion of the results of the ablation experiment

The quantitative results of the ablation experiments are shown in Table 1, and the typified results are depicted in Fig. 7.

When the template and sample are concatenated and input into the network (S1), a limited implicit contrasting capability of the network is observed. Utilizing the Siamese network, where corresponding features at different scales are subtracted (S2), the mIoU improved by 2.50%. In S3, the elimination of positional encoding leads to a 5.24% increase in mIoU. Despite this significant improvement, it falls short by 0.49% compared to S7. This indicates that while positional encoding contributes to the construction of BGNet, its impact is somewhat constrained. In S4, there is a 2.48% increase in mIoU, almost mirroring the result achieved with the Siamese network in S2. This suggests that in the absence of global and mutual information, the matching algorithm contributes minimally to the mIoU. However, a notable improvement in the F2 score highlights a substantial increase in recall, demonstrating that direct matching significantly aids in background noise reduction. In S5, the mIoU increases by 4.34%. Employing positional encoding and self- and cross-attention to extract dense features containing global and mutual information aids the model in focusing on defect features while somewhat disregarding background features. In S6, the absence of MNN in feature matching leads to a 3.61% increase in mIoU, confirming that matching enhances mIoU; however, the accuracy of the matching relationship critically influences the final outcome. Finally, in S7, there is a 5.73% increase in mIoU, and the F2 score is the highest, indicating that the dense features extracted through self- and cross-attention significantly improve the matching algorithm. This validates

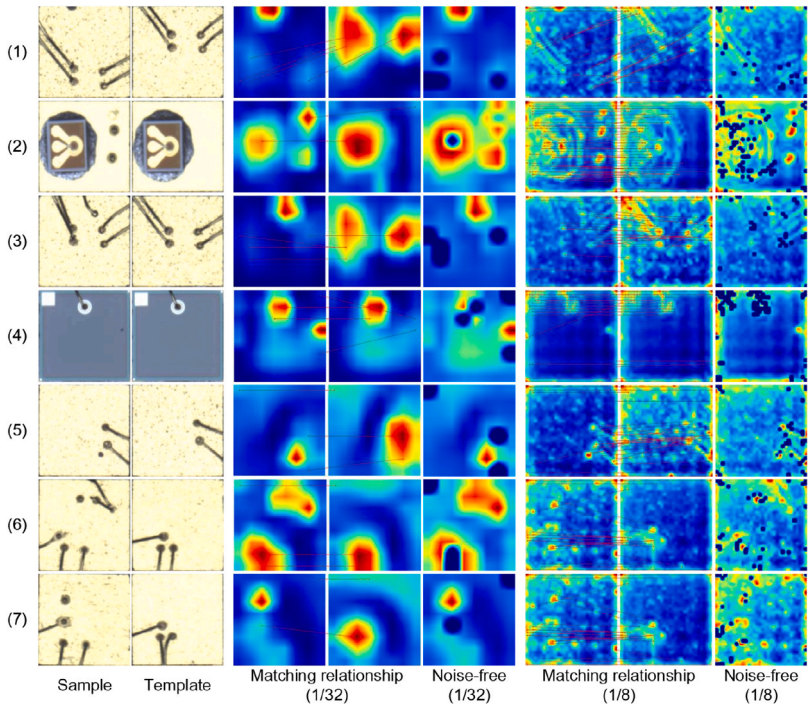


Fig. 6. Visualization of feature matching results on the OCDs dataset. The dots represent the positions of matching points, while the lines illustrate the matching relationships between these points.

Table 1
Results of ablation on the OCDs dataset.

Modules	Baseline	Siamese	Positional encoding	Self- and Cross-attention	Feature matching[a]	mIoU	F2
S1	✓					0.7637	0.8683
S2	✓	✓				0.7887	0.8850
S3	✓	✓		✓	✓	0.8161	0.9145
S4	✓	✓	✓		✓	0.7885	0.9048
S5	✓	✓	✓	✓		0.8071	0.8948
S6	✓	✓	✓	✓	◯[a]	0.7998	0.8923
S7	✓	✓	✓	✓	✓	0.8210	0.9101

[a] The symbol ◯ represents feature matching without MNN.

the importance of each component of BGNet, with the most effective results achieved through their combined application.

Qualitatively, S7 demonstrates a precise focus on the defect foreground features at the 1/8 and 1/16 scales. The deep blue areas, which are indicative of regions subtracted through matching, reveal spatial changes in the sample background. In contrast to S7, S3 exhibits a notable reduction in the number of matching points, slightly diminishing its ability to eliminate spatial variation noise. Furthermore, S4 effectively removes noise features at the 1/32 scale but shows less efficacy at the 1/8 scale. In comparison with S2, S5 more attentively highlights defect foregrounds at the 1/8 scale; however, it still retains a significant amount of noise due to spatial variations at the 1/32 scale. Finally, S6, which underwent ablation of the MNN in feature matching, displayed inaccurate matching relationships, impacting the overall result.

4.5. Comparison with the state-of-the-art models

BGNet effectiveness is demonstrated by comparing it with fifteen existing methods, including the following: Five classical methods concatenate the sample and template and input them into the network. This approach exploits the network's potential to adapt to background spatial changes. Four classic semantic segmentation networks (UNet [46], FCN [47], SegNet [48], and DeepLabV3+ [51]) and a classic surface defect detection network (PGANet [49]) were selected. Four methods based on attention mechanisms, including three classical attention mechanism methods (CCNet [52], DUNet [52], and DANet [53]) and a recent method based on the Transformer Swin UNet [54], were used. Two methods for foreground generalization, TGRNet [29] and PFENet [55], were selected to validate that foreground generalization has limited applicability to background generalization. The three methods for background generalization include all the contrast-based background generalization methods, such as Siamese UNet [15], DSSS-Net [11], and GWNet [13]. In addition, we conducted a comparison between two feature matching methods. These include soft matching [44], which operates on feature maps, and SIFT+UNet, which involves original sample and template matching. Specifically, for the SIFT+UNet approach, the SIFT [32] algorithm is employed to match layers. Subsequently, pixel subtraction is carried out in accordance with the established matching relationships. The resulting image is subsequently fed into UNet for segmentation.

4.5.1. Comparison on the OCDs dataset

In this section, we compare our approach with state-of-the-art methods from both quantitative (as depicted in Table 2) and qualitative perspectives (as illustrated in Fig. 8).

Quantitatively, classic CNN-based networks exhibit limited implicit contrasting capabilities between templates and samples. Similarly, the performance of attention mechanism-based methods is also restricted. Foreground generalization methods do not provide positive

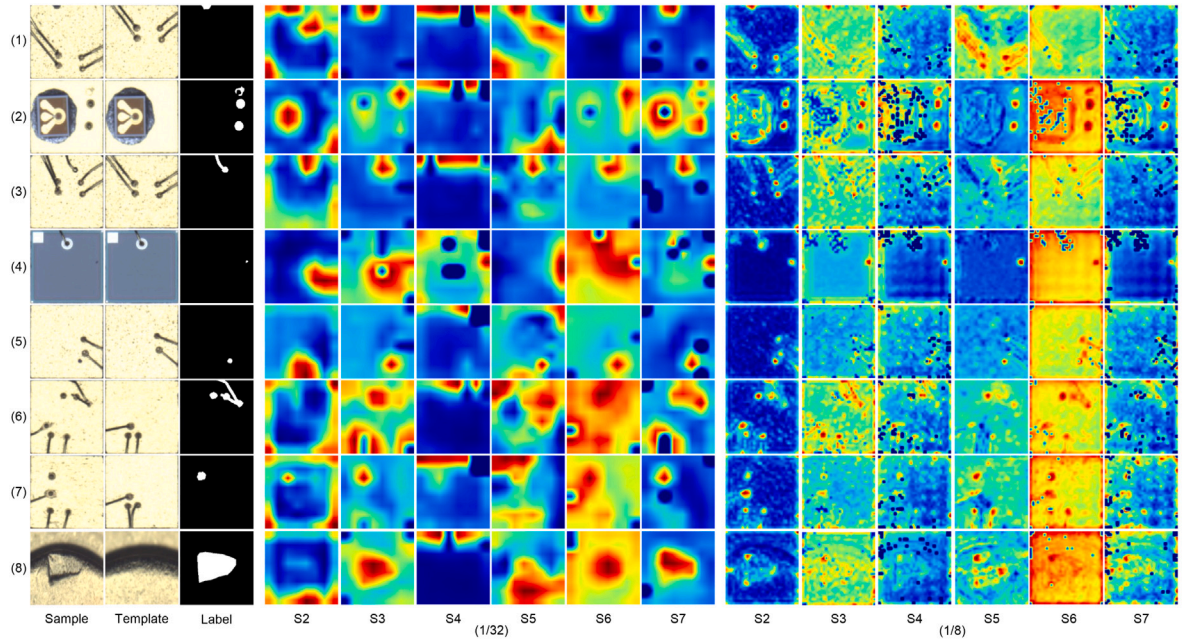


Fig. 7. Visualization of ablation studies on the OCDs dataset.

Table 2
Quantitative comparison with state-of-the-arts methods on the OCDs datasets.

	Method	Pre	Recall	F2	mIoU
Classical methods	UNet	0.8597	0.6926	0.7177	0.6325
	FCN	0.8831	0.7376	0.7627	0.6580
	SegNet	0.8949	0.3907	0.4403	0.3662
	DeepLabV3+	0.8295	0.7967	0.8031	0.6702
	PGANet	0.9186	0.4793	0.5300	0.4483
Attention-based methods	CCNet	0.8224	0.3875	0.4333	0.3614
	DUNet	0.8716	0.3100	0.3559	0.2942
	DANet	0.8220	0.5748	0.6116	0.5130
	Swin UNet	0.6612	0.2569	0.2927	0.2076
Foreground generalization methods	TGRNet	0.2446	0.4770	0.4008	0.1638
	PFENet	0.1929	0.2713	0.2509	0.1233
Background generalization methods	Siamese UNet	0.8913	0.6946	0.7267	0.6243
	DSSSNet	0.8931	0.8148	0.8293	0.7405
	GWNet	**0.9070**	0.8891	0.8926	0.8074
Feature matching based methods	SIFT+UNet	0.8416	0.7275	0.7478	0.6661
	Soft Matching	0.5276	0.3157	0.3433	0.3000
Ours	BGNet	0.8961	**0.9137**	**0.9101**	**0.8210**

contributions to the problem of background generalization that we study; their performance is even worse.

Among the background generalization methods, Siamese UNet has a similar structure to UNet. Despite the explicit feature subtraction, the performance hardly improves (the mIoU is 62.43%). This demonstrates that concatenating and inputting into the network can exploit the potential of CNN-based networks to contrast templates and samples. However, direct subtraction cannot resolve the noise caused by spatial changes in background features. DSSSNet adopts a metric similar to max pooling, achieving some degree of noise elimination due to spatial variations, leading to a significant improvement in the mIoU (74.05%). This finding substantiates that eliminating noise caused by background changes is a key measure for achieving background generalization. GWNet employs an attention mechanism, and based on the location-independent characteristics of the attention mechanism, it further enhances the model's ability to eliminate background noise, increasing the mIoU to 80.74%.

In terms of feature matching-based methods, SIFT+UNet operates at the image-level, matches and then subtracts pixels based on their matching relationships. This method achieves a modest mIoU of 66.61%. Compared to UNet, which attains an mIoU of 63.25% with a similar network structure, and Siamese UNet, achieving an mIoU of 62.43%, the improvement provided by SIFT+UNet is evidently limited. This finding suggests that matching subtraction is somewhat effective for images with spatially varying noise. However, its performance is substantially inferior to that of BGNet, achieving a significantly higher mIoU of 82.10%. This disparity underscores the limited accuracy of direct image-level feature matching. The soft matching method focuses on the entire image's characteristics and matches the position of the fabric pattern. However, this approach fails to adequately address local deformations, leading to a relatively low mIoU of only 30.00% on OCD datasets.

The feature matching-based method proposed in this paper explicitly eliminates background noise and further increases the mIoU to 82.10%, surpassing state-of-the-art methods.

Qualitatively, classical methods exhibit significant false negatives and false positives. False positives primarily occur in areas with spatial changes in background features, such as the gold wire in row (6). False negatives mainly occur in areas where foreground defect features overlap with spatially varying background features, such as rows (3) and (4). Attention-based methods, such as CCNet and Swin UNet, perform poorly, with numerous false negatives and severe false positives, respectively. Although these methods have achieved some differences between the template and the sample, their ability to distinguish between the background and the foreground is limited. Foreground generalization methods, such as PEFNet and TGRNet, perform the worst and are almost incapable of correctly detecting defects.

Among background generalization methods, DSSSNet reduces the incidence of false positives compared to Siamese UNet and greatly improves the noise removal ability, as shown in row (6). However, it also exhibits some false negatives, such as in row (4). GWNet shows promising detection results but lacks accuracy in detecting details compared to BGNet, as shown in rows (4) and (10). It also exhibits some minor noise, such as in rows (1), (3), (8), and (10).

Among the feature matching-based methods, SIFT+UNet clearly focuses on the foreground area, effectively minimizing spatial variation noise. However, for the image in row (10), characterized by a complex background, the presence of noticeable spatial variation noise can be observed. This issue may stem from inaccurate matching relationships.

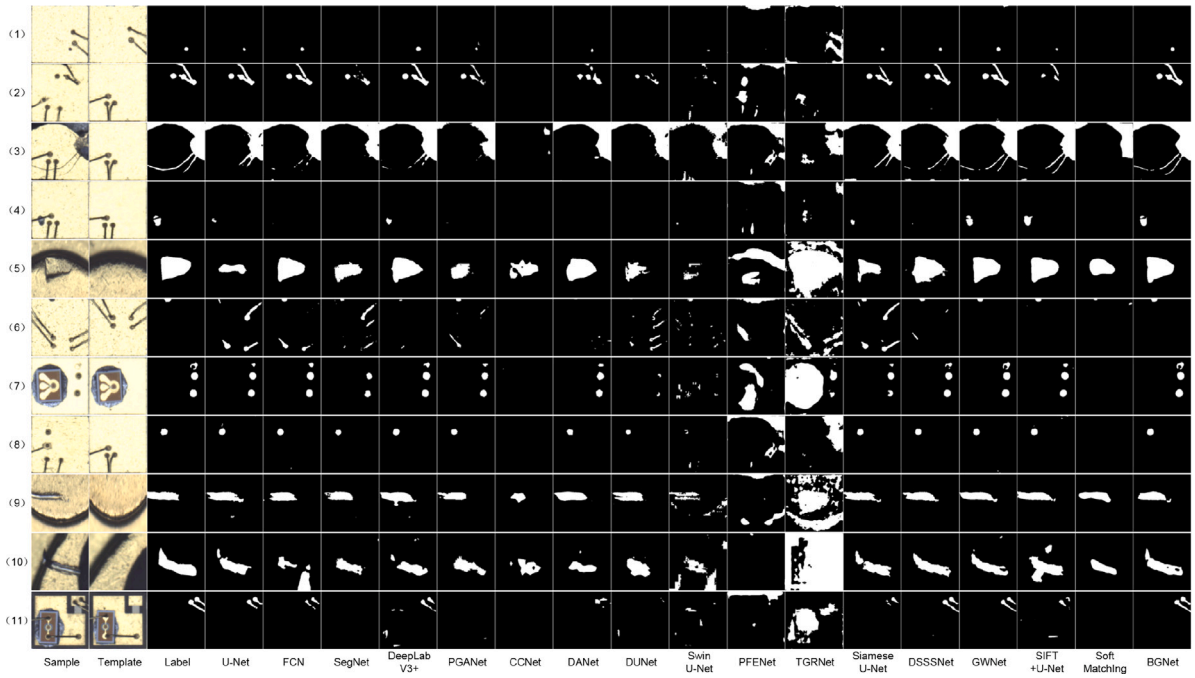


Fig. 8. Visualization of comparative analysis with state-of-the-art methods on the OCDs dataset.

Table 3
Quantitative comparison with state-of-the-arts methods on the PCBs datasets.

	Method	mIoU	mACC	Params (MB)
Classical methods	UNet	0.5981	0.9046	7.86
	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-UNet	0.7719	0.9972	27.16
Foreground generalization methods	TGRNet	0.7068	0.8988	32.12
	PFENet	0.7214	0.9105	30.25
Background generalization methods	Siamese UNet	0.7837	0.9954	**7.85**
	DSSSNet	0.7634	0.9678	33.60
	GWNet	0.8243	0.9978	26.54
Feature matching based methods	SIFT+UNet	0.0510	0.2988	7.86
	Soft Matching	0.7890	0.8719	16.43
Ours	BGNet	**0.8393**	**0.9996**	47.70

Additionally, the soft matching algorithm has certain limitations. In many cases, such as in rows (1), (2), and (3), the algorithm fails to detect images entirely, likely due to its inherent methodological characteristics.

4.5.2. Comparison on the PCB dataset

To further validate the effectiveness of BGNet, we added a multi-class dataset, the PCB dataset. The quantitative results are presented in Table 3, and the qualitative results are shown in Fig. 9.

Quantitatively, Like on the OCDs dataset, conventional methods based on CNNs and attention mechanisms demonstrated limited effectiveness, while foreground generalization methods performed the poorest. Among the background generalization methods, Siamese UNet and GWNet outperform DSSSNet. This may be attributed to the fact that methods similar to max pooling have predefined pooling ranges, which necessitate adaptation to the scale of defects and background features. Such predefined spatial variation adaptation methods have inherent limitations. In feature matching-based methods, a stark contrast is observed when comparing results from the OCDs and PCBs datasets. Soft matching achieved significantly better outcomes than SIFT+UNet on the PCB dataset. This difference may stem from the high density and similarity of surface components in the PCB dataset, which poses challenges for accurately computing SIFT matching relationships. Furthermore, variations in the component space in PCBs tend to be more prominent in overall shifts than in minute feature changes. BGNet achieves the best results among all the methods. Although the number of parameters in BGNet increased, it remains within an acceptable range compared to that of existing methods.

Qualitatively, the background components are more numerous and densely distributed in the PCB dataset, and the spatial variation in background features between samples and templates is relatively small. This finding is consistent with the small difference in the mIoU values of the various methods shown in Table 3. In Fig. 6, the classical methods based on CNNs and those based on attention mechanisms have many false positives. Many foreground generalization methods have many false negatives. The background generalization methods generally performed well, and BGNet achieved very high accuracy.

5. Discussion and conclusion

5.1. Discussion

A comparison of templates and samples is an effective and widely used method in defect detection models. This paper tackles the challenge of background generalization, with a particular emphasis on the spatial variation of background features, which introduces noise into the comparative analysis. Based on our ideas and experimental results, we discuss the following points:

(1) Our experimental results show that CNN networks do not exhibit spatial invariance properties. Existing research also indicates that CNN networks are spatially equivariant. The pooling operation can provide CNNs with limited spatial invariance properties, as demonstrated by the improved performance of DSSSNet on OCDs. However, due to the definition of the pooling operation, the receptive field is fixed, resulting in poor performance of DSSSNet on PCBs.

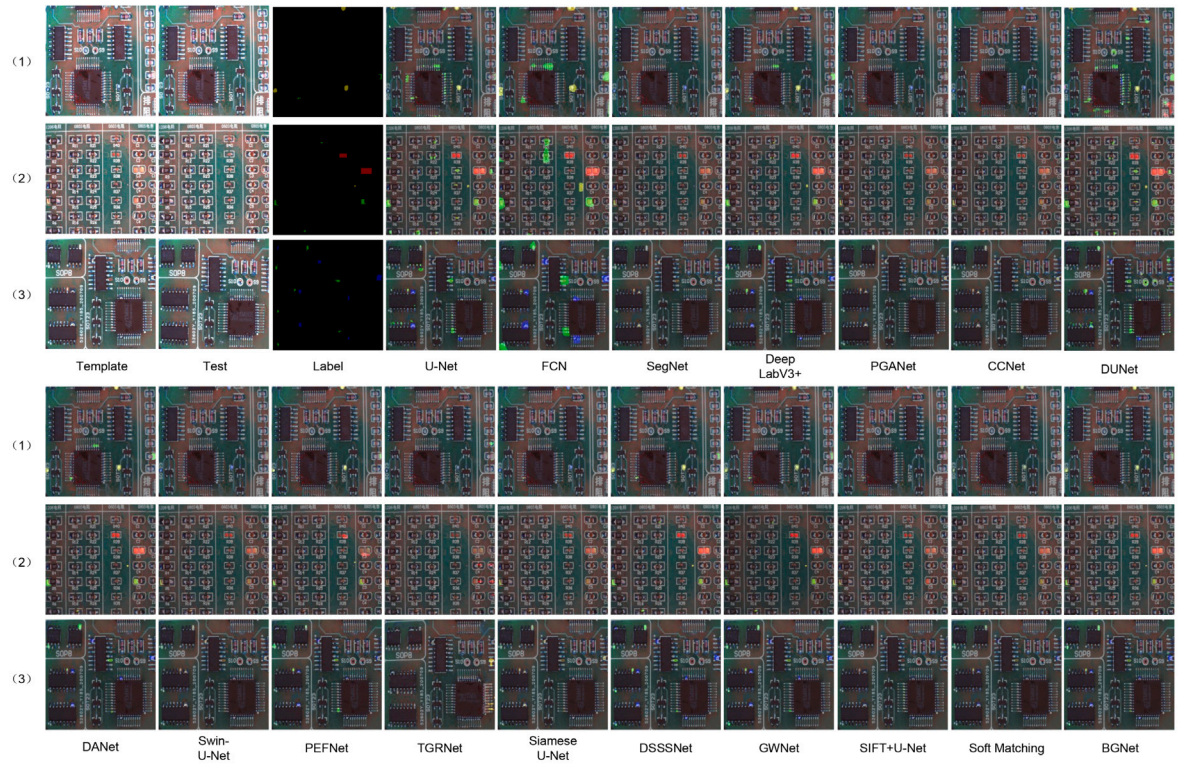


Fig. 9. Visualization of comparative analysis with state-of-the-art methods on the PCBs dataset.

(2) The self- and cross-attention mechanisms in Transformers have greater equivariance properties for spatial changes in background features due to their ability to capture global and interactive information. However, the underlying principle of this equivariance remains unclear. In GWNet, position encoding was not performed prior to calculating self-attention. This approach leverages the location-independent nature of self-attention to reduce the impact of spatial variation noise on the results. In contrast, BGNet incorporates position encoding, which also helps mitigate the effects of spatial variation noise.

(3) When subtracting feature matches with background variation, it is not necessary to subtract all backgrounds individually. Instead, subtracting only significant features allows the network to eliminate noise caused by spatial variations.

5.2. Conclusion

In this study, we introduce a novel background generalization network (BGNet) that leverages feature matching to achieve state-of-the-art results. Our network employs self- and cross-attention mechanisms to extract dense features containing global and interactive information. Feature matching is accomplished by using the MNN algorithm, and subtraction is performed based on the matching relationship to explicitly eliminate spatially variant background features. Our proposed method demonstrates exceptional performance on both the OCD and PCB datasets. Future work will focus on exploring the mathematical principles underlying spatial variations and designing networks based on matrix translation, rotation, and affine transformation to further elucidate the mechanisms governing spatial variations in background features.

CRediT authorship contribution statement

Biao Chen: Writing – original draft, Visualization, Validation, Resources, Methodology, Formal analysis, Data curation. **Tongzhi Niu:** Writing – review & editing, Validation, Supervision, Methodology, Data curation, Conceptualization. **Ruoqi Zhang:** Visualization, Methodology, Data curation. **Hang Zhang:** Methodology, Data curation. **Yuchen Lin:** Methodology, Data curation. **Bin Li:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

[1] M.M. Ferdaus, B. Zhou, J.W. Yoon, K.L. Low, J. Pan, J. Ghosh, M. Wu, X. Li, A.V.-Y. Thean, J. Senthilnath, Significance of activation functions in developing an online classifier for semiconductor defect detection, Knowl.-Based Syst. 248 (2022) 108818, http://dx.doi.org/10.1016/j.knosys.2022.108818.

[2] X. Dong, C.J. Taylor, T.F. Cootes, Automatic aerospace weld inspection using unsupervised local deep feature learning, Knowl.-Based Syst. 221 (2021) 106892, http://dx.doi.org/10.1016/j.knosys.2021.106892.

[3] L. Yang, J. Fan, B. Huo, E. Li, Y. Liu, A nondestructive automatic defect detection method with pixelwise segmentation, Knowl.-Based Syst. 242 (2022) 108338, http://dx.doi.org/10.1016/j.knosys.2022.108338.

[4] G. Wang, M. Chen, Y. Lin, X. Tan, C. Zhang, W. Yao, B. Gao, K. Li, Z. Li, W. Zeng, Efficient multi-branch dynamic fusion network for super-resolution of industrial component image, Displays (2023) 102633, http://dx.doi.org/10.1016/j.displa.2023.102633.

[5] X. Yu, W. Lyu, C. Wang, Q. Guo, D. Zhou, W. Xu, Progressive refined redistribution pyramid network for defect detection in complex scenarios, Knowl.-Based Syst. 260 (2023) 110176, http://dx.doi.org/10.1016/j.knosys.2022.110176.

[6] G. Tong, Q. Li, Y. Song, Two-stage reverse knowledge distillation incorporated and Self-Supervised Masking strategy for industrial anomaly detection, Knowl.-Based Syst. 273 (2023) 110611, http://dx.doi.org/10.1016/j.knosys.2023.110611.

[7] T. Niu, B. Chen, Q. Lyu, B. Li, W. Luo, Z. Wang, B. Li, Scoring Bayesian Neural Networks for learning from inconsistent labels in surface defect segmentation, Measurement 225 (2024) 113998, http://dx.doi.org/10.1016/j.measurement.2023.113998.

[8] X. Yu, W. Lyu, C. Wang, Q. Guo, D. Zhou, W. Xu, Progressive refined redistribution pyramid network for defect detection in complex scenarios, Knowl.-Based Syst. 260 (2023) 110176, http://dx.doi.org/10.1016/j.knosys.2022.110176.

[9] T. Liu, Z. He, Z. Lin, G.-Z. Cao, W. Su, S. Xie, An adaptive image segmentation network for surface defect detection, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–14, http://dx.doi.org/10.1109/TNNLS.2022.3230426.

[10] H. Zhang, Y. Chen, B. Liu, X. Guan, X. Le, Soft matching network with application to defect inspection, Knowl.-Based Syst. 225 (2021) 107045, http://dx.doi.org/10.1016/j.knosys.2021.107045.

[11] Z. Ling, A. Zhang, D. Ma, Y. Shi, H. Wen, Deep siamese semantic segmentation network for PCB welding defect detection, IEEE Trans. Instrum. Meas. 71 (2022) 1–11, http://dx.doi.org/10.1109/TIM.2022.3154814.

[12] S. Ma, K. Song, M. Niu, H. Tian, Y. Wang, Y. Yan, Shape consistent one-shot unsupervised domain adaptation for rail surface defect segmentation, IEEE Trans. Ind. Inform. (2023) http://dx.doi.org/10.1109/TII.2022.3233654.

[13] T. Niu, Z. Xie, J. Zhang, L. Tang, B. Li, H. Wang, A generalized well neural network for surface defect segmentation in Optical Communication Devices via Template-Testing comparison, Comput. Ind. 151 (2023) 103978, http://dx.doi.org/10.1016/j.compind.2023.103978.

[14] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Advances in Neural Information Processing Systems, Vol. 28, 2015.

[15] D. Kwon, J. Ahn, J. Kim, I. Choi, S. Jeong, Y.-S. Lee, J. Park, M. Lee, Siamese U-net with healthy template for accurate segmentation of intracranial hemorrhage, in: D. Shen, T. Liu, T.M. Peters, L.H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan (Eds.), Medical Image Computing and Computer Assisted Intervention, MICCAI 2019, Springer International Publishing, Cham, 2019, pp. 848–855.

[16] J. Sun, Z. Shen, Y. Wang, H. Bao, X. Zhou, LoFTR: Detector-free local feature matching with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8922–8931.

[17] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, C.-L. Tai, Transfusion: Robust lidar-camera fusion for 3d object detection with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1090–1099.

[18] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, D. Tao, Gmflow: Learning optical flow via global matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8121–8130.

[19] S. Zhu, X. Liu, PMatch: Paired masked image modeling for dense geometric matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21909–21918.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, 2017.

[21] P.-E. Sarlin, D. DeTone, T. Malisiewicz, A. Rabinovich, Superglue: Learning feature matching with graph neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4938–4947.

[22] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, N. Sang, Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation, Int. J. Comput. Vis. 129 (2021) 3051–3068, http://dx.doi.org/10.1007/s11263-021-01515-2.

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

[24] J.P. Yun, W.C. Shin, G. Koo, M.S. Kim, C. Lee, S.J. Lee, Automated defect inspection system for metal surfaces based on deep learning and data augmentation, J. Manuf. Syst. 55 (2020) 317–324, http://dx.doi.org/10.1016/j.jmsy.2020.03.009.

[25] K. Li, Z. Li, X. Jia, L. Liu, M. Chen, A domain adversarial graph convolutional network for intelligent monitoring of tool wear in machine tools, Comput. Ind. Eng. 187 (2024) 109795, http://dx.doi.org/10.1016/j.cie.2023.109795.

[26] W. Xiao, K. Song, J. Liu, Y. Yan, Graph embedding and optimal transport for few-shot classification of metal surface defect, IEEE Trans. Instrum. Meas. 71 (2022) 1–10, http://dx.doi.org/10.1109/TIM.2022.3169547.

[27] Y. Song, Z. Liu, S. Ling, R. Tang, G. Duan, J. Tan, Coarse-to-fine few-shot defect recognition with dynamic weighting and joint metric, IEEE Trans. Instrum. Meas. 71 (2022) 1–10, http://dx.doi.org/10.1109/TIM.2022.3193204.

[28] W. Zhao, K. Song, Y. Wang, S. Liang, Y. Yan, FaNet: Feature-aware network for few shot classification of strip steel surface defects, Measurement 208 (2023) 112446, http://dx.doi.org/10.1016/j.cosrev.2020.100359.

[29] Y. Bao, K. Song, J. Liu, Y. Wang, Y. Yan, H. Yu, X. Li, Triplet-graph reasoning network for few-shot metal generic surface defect segmentation, IEEE Trans. Instrum. Meas. 70 (2021) 1–11, http://dx.doi.org/10.1109/TIM.2021.3083561.

[30] D. Shan, Y. Zhang, S. Coleman, D. Kerr, S. Liu, Z. Hu, Unseen-material few-shot defect segmentation with optimal bilateral feature transport network, IEEE Trans. Ind. Inform. 19 (7) (2023) 8072–8082, http://dx.doi.org/10.1109/TII.2022.3216900.

[31] X. Shi, S. Zhang, M. Cheng, L. He, X. Tang, Z. Cui, Few-shot semantic segmentation for industrial defect recognition, Comput. Ind. 148 (2023) 103901, http://dx.doi.org/10.1016/j.compind.2023.103901.

[32] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110, http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

[33] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision, 2011, pp. 2564–2571, http://dx.doi.org/10.1109/ICCV.2011.6126544.

[34] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features, in: Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV, Vol. 11, Springer, 2010, pp. 778–792.

[35] K.M. Yi, E. Trulls, V. Lepetit, P. Fua, Lift: Learned invariant feature transform, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VI, vol. 14, Springer, 2016, pp. 467–483.

[36] D. DeTone, T. Malisiewicz, A. Rabinovich, Superpoint: Self-supervised interest point detection and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 224–236.

[37] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, T. Sattler, D2-net: A trainable cnn for joint detection and description of local features, 2019.

[38] Y. Ono, E. Trulls, P. Fua, K.M. Yi, LF-Net: Learning local features from images, in: Advances in Neural Information Processing Systems, vol. 31, 2018.

[39] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon, M. Humenberger, R2D2: repeatable and reliable detector and descriptor, 2019.

[40] C. Liu, J. Yuen, A. Torralba, Sift flow: Dense correspondence across scenes and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2010) 978–994, http://dx.doi.org/10.1109/TPAMI.2010.147.

[41] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, J. Sivic, Neighbourhood consensus networks, in: Advances in Neural Information Processing Systems, vol. 31, 2018.

[42] I. Rocco, R. Arandjelović, J. Sivic, Efficient neighbourhood consensus networks via submanifold sparse convolutions, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX, vol. 16, Springer, 2020, pp. 605–621.

[43] X. Li, K. Han, S. Li, V. Prisacariu, Dual-resolution correspondence networks, Adv. Neural Inf. Process. Syst. 33 (2020) 17346–17357.

[44] H. Zhang, Y. Chen, B. Liu, X. Guan, X. Le, Soft matching network with application to defect inspection, Knowl.-Based Syst. 225 (2021) 107045, http://dx.doi.org/10.1016/j.knosys.2021.107045.

[45] A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret, Transformers are rnns: Fast autoregressive transformers with linear attention, in: International Conference on Machine Learning, PMLR, 2020, pp. 5156–5165.

[46] O. Ronneberger, P. Fischer, T. Brox, 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, vol. 18, Springer, 2015, pp. 234–241.

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

[48] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495, http://dx.doi.org/10.1109/TPAMI.2016.2644615.

[49] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, Q. Meng, PGA-Net: Pyramid feature fusion and global context attention network for automated surface defect detection, IEEE Trans. Ind. Inform. 16 (12) (2019) 7448–7458, http://dx.doi.org/10.1109/TII.2019.2958826.

[50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

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

[52] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 603–612.

[53] Q. Jin, Z. Meng, T.D. Pham, Q. Chen, L. Wei, R. Su, DUNet: A deformable network for retinal vessel segmentation, Knowl.-Based Syst. 178 (2019) 149–162, http://dx.doi.org/10.1016/j.knosys.2019.04.025.

[54] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.

[55] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, J. Jia, Prior guided feature enrichment network for few-shot segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2) (2020) 1050–1065, http://dx.doi.org/10.1109/TPAMI.2020.3013717.