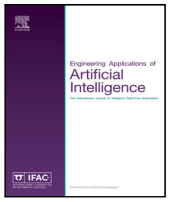
Engineering Applications of Artificial Intelligence 123 (2023) 106474

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



SMINet:Semantics-aware multi-level feature interaction network for surface defect detection

Bin Wan [a], Xiaofei Zhou [a,*], Yaoqi Sun [c], Zunjie Zhu [b], Haibing Yin [b], Ji Hu [d], Jiyong Zhang [a], Chenggang Yan [a,*]

[a] School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China
[b] Lishui Institute of Hangzhou Dianzi University, and School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
[c] Lishui Institute of Hangzhou Dianzi University, and School of Automation Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
[d] Lishui Institute of Hangzhou Dianzi University, Hangzhou Dianzi University, Hangzhou 310018, China

ARTICLE INFO

Keywords:
Surface defect detection
Salient object detection
Cross-layer feature fusion
Semantic-aware feature extraction

ABSTRACT

To boost the product quality, numerous saliency-based surface defect detection methods have been devoted to the areas of industrial production, construction consumable, road construction. However, the existing salient object detection (SOD) methods not only consume a significant amount of computing resources but also fail to meet the detection efficiency requirements of enterprises. Therefore, this paper proposes a lightweight semantics-aware multi-level feature interaction network (SMINet), to address the above issues. In the encoder phase, we integrate multiple adjacent level features in the cross-layer feature fusion (CFF) module to alleviate the discrepancy between multi-scale features. In the decoder phase, we first employ the semantic-aware feature extraction (SFE) module to mine the location cues embedded in the high-level features. Afterwards, we introduce the detail-aware context attention (DCA) module based on the attention mechanism to recover more spatial details. Extensive experiments on four surface defect datasets validate that our SMINet outperforms the existing state-of-the-art methods.

1. Introduction

Defect detection as a branch of saliency detection aims to locate and highlight the defect regions, which is widely applied to various production activities, such as steel manufacturing, road maintenance, equipment overhaul, and so on. With the development of computer vision, there have been various types of tasks based on machine vision in recent years such as object detection (Dong et al., 2021b; Zhang et al., 2022; Zhou et al., 2022), object tracking (Hare et al., 2015; Lu et al., 2019), and image classification (He et al., 2019; Park and Yang, 2019)), and defect detection (Zhou et al., 2021b; Li et al., 2019) with a wide range of application scenarios has attracted the attention of many researchers.

In the last two decades, traditional computer vision-based defect detection methods, which mainly rely on the hand-crafted cues (*i.e.*, color, contrast, and texture), have become the leading way in many industrial scenarios (Sharifzadeh et al., 2008), and it is gradually applied to the civilian realm (Tajeripour et al., 2007). For example, in Ng (2006), Fang et al. proposed a revised Otsu method for defect detection by automatically selecting optimal threshold values for both unimodal and bimodal distributions. Mak et al. (2009) leveraged pre-trained gabor wavelet network to extract texture features of fabric which facilitate the construction of structuring elements. Moreover, Ngan et al. (2005) proposed an automated inspection method that consists of golden image subtraction method and wavelet pre-processed golden image subtraction for defect detection on patterned fabric. Although traditional detection methods achieve a certain degree of improvement in efficiency and accuracy, these methods only extract the shallow features, ignoring the location and detail information embedded in deep features and causing incomplete segmentation of defect regions when dealing with low-quality input images (*i.e.*, low contrast and low resolution)

With the rapid development of deep learning, this technology has become the mainstream way to cope with various vision tasks. Benefiting from the ability of convolutional neural network (CNN) to extract deep features as the convolutional layer deepens, all kinds of information are revealed, among which high-level semantic information is conducive to locating the objects, and low-level discriminative information contributes to the recovery of detail regions. Therefore, the CNN-based defect detection method has become a promising research field, which replaces the traditional methods and further promotes the

* Corresponding authors.
E-mail addresses: wanbinxueshu@icloud.com (B. Wan), zxforchid@outlook.com (X. Zhou), syq@hdu.edu.cn (Y. Sun), zunjiezhu@hdu.edu.cn (Z. Zhu), yhb@hdu.edu.cn (H. Yin), huji@hdu.edu.cn (J. Hu), jzhang@hdu.edu.cn (J. Zhang), cgyan@hdu.edu.cn (C. Yan).

https://doi.org/10.1016/j.engappai.2023.106474
Received 9 November 2022; Received in revised form 29 March 2023; Accepted 13 May 2023
Available online 1 June 2023
0952-1976/

accuracy of detection results. For instance, Li and Xi (2021) proposed a novel defect detection network which consists of a binary classification network to determine whether the image contains defects and a detection network to detect the defects, improving the detection speed. Cheng and Yu (2020) leveraged search-based anchor optimization to improve the accuracy and proposed a novel channel attention mechanism to reduce information loss. Besides, adaptive spatial feature fusion module was introduced to fuse shallow and deep features effectively. Wang and Wu (2021) adopted skip layer connection module and pyramid feature fusion module to improve the performance of defect images with significant intraclass differences and high interclass similarity and solve the problem of low detection speed. In recent works (Konovalenko et al., 2022a,b), Konovalenko investigated the impact of the illumination level on the rolled strip and proposed a U-net architecture to detect the detects.

Despite the outstanding performance of CNN-based defect detection methods, there are still multiple issues that need to be addressed. First, it is vital to take advantage of multi-scale features from the encoder for saliency detection. However, most existing methods based on the fully convolutional network (FCN) directly feed multi-scale features from the backbone to the decoder and simply integrate these features in a bottom-to-top manner via upsampling and the continuous reduction in feature resolution, which is introduced by a series of convolution and pooling operations causes significant gaps between multi-scale features, diluting the defect region information in the process of transferring from low resolution to high resolution and reducing the accuracy of detection results greatly. Although some current methods (Pang et al., 2020) fuse adjacent two- or three-layer features before the next stage and reduce the impact of information dilution to a certain extent, these methods do not take into account the connection of more cross-layer features, resulting in a limitation of information interaction between features with large resolution discrepancy. Second, digging for effective semantic information embedded in the high-level features becomes a key component of determining object location in the SOD. Many existing methods (Zhai et al., 2021) adopt atrous spatial pyramid pooling (ASPP) (Chen et al., 2017) behind the high-level features, which utilize convolutional layers with different receptive fields to further extract rich semantics. However, ASPP gives rise to the grid artifacts and increases the number of parameters and lowers the calculation speed. Third, compared to the high-level features, low-level features with larger size have more spatial structural details but also contains excessive noise. Many previous methods (Yang et al., 2019; Chen et al., 2019) directly integrate low-level features from the encoder via a feature pyramid manner, but the background noise cannot be filtered out when dealing with low-contrast defects.

To address the above challenges, we propose a novel defect detection network named SMINet, which accomplishes remarkable performance improvement in detecting different kinds of surface defects. First, to eliminate the feature information discrepancy induced by scale differences as much as possible and avoid the information dilution introduced by bottom-to-top connection, we propose a cross-layer feature fusion (CFF) module. Specifically, CFF module shown in Fig. 2 adopts cross-layer fusion strategy to integrate features of multiple adjacent layers, where the complementarity between multi-scale features can be reflected. Thus, for the low-level features, it can fuse with other scale features to the greatest extent while retaining more detailed information. Besides, integrating high-level features with their adjacent features can avoid the interference of low level noise and enhance the high level semantic cues. By introducing CFF module, the integrity of multi-scale information can be guaranteed during the transmission between different layers. Second, different from the commonly used ASPP module, we introduce the semantic-aware feature extraction (SFE) module shown in Fig. 3 which consists of two stages (*i.e.*, feature interaction (FI) and feature refinement (FR)) to mine the semantic information. In stage one, GCN-like structure is adopted to achieve the correlation between two features and explore the semantic relation between each pixel and each channel. After that, stage two aggregates the semantic-aware features from FI and leverages attention mechanism to realize the refinement. Third, after exploiting the cross-layer feature fusion module (CFF), although low level features contain more detailed structural cues, much non-related noise information also is introduced. Hence, we deploy the detail-aware context attention (DCA) module, as shown in Fig. 4, after two low level features, which takes advantage of different attention mechanisms for the features of different channels. Finally, to demonstrate the remarkable performance of our proposed SMINet, we conduct extensive experiments on four public datasets, and the visualization and quantitative results prove that our method outperforms other state-of-the-art methods on effectiveness and superiority.

The contributions of this paper can be summarized as follows:

1. A novel saliency detection method named SMINet, is introduced to detect four different types of surface defects, which consists of the cross-layer feature fusion (CFF) module, semantic-aware feature extraction (SFE) module, and detail-aware context attention (DCA) module.
2. The CFF module is proposed to enhance the complementarity of multi-scale features and low the impact of information dilution caused by bottom-to-top connection. Visualized results shown in Fig. 5 also verify the effectiveness of CFF module.
3. To extract semantic cues embedded in the high-level features, we propose an SFE module consisting of feature interaction and feature refinement, which adopt GCN-like structure and channel attention mechanism.
4. We propose the DCA module that leverages two attention strategies to filter out redundant information introduced by the CFF module.

2. Related works

Here, we briefly introduce some related works about salient object detection, surface defect detection, and multi-level fusion-based methods.

2.1. Salient object detection

Salient object detection has been applied to various scenarios (*i.e.*, remote sensing image, medical image, and video) and gains excellent results. In recent years, there are lots of saliency models have been proposed. For example, Wei et al. (2012) proposed a geodesic saliency which exploits boundary prior and connectivity prior to focus more on the background instead of the foreground. Ju et al. (2014) detected the saliency of point depending on how much it outstands from the background and utilized depth and location priors for refinement. Recently, deep learning-based saliency methods have verified their effectiveness. Zhang et al. (2020a) proposed a novel saliency detection network which incorporates attention mechanism and explores context-aware mechanism to capture long-term semantic context relationship. Zhang and Ma (2021) proposed a SOD method that combines the benefits of weakly and fully supervised learning, where pseudo-label generation method is applied to reduce the demands of large scale pixel-wise annotations. Besides, a feedback saliency analysis network was constructed to generate pixel-wise labels. Wang et al. (2021) proposed a joint learning network which consists of cross-task feature subnet, 3d lesion subnet, and classification subnet to cope with the shortage of medical resources. Fan et al. (2020) proposed an automatic detection network, where parallel partial decoder is used to integrate high-level features and two attention mechanisms are utilized to enhance the representations of infected regions. Currently, SOD methods are gradually applied to the video field. Wang et al. (2017) proposed a novel data augmentation technique to make network learn more diverse saliency information and captured temporal saliency from frame pairs. Ji et al.

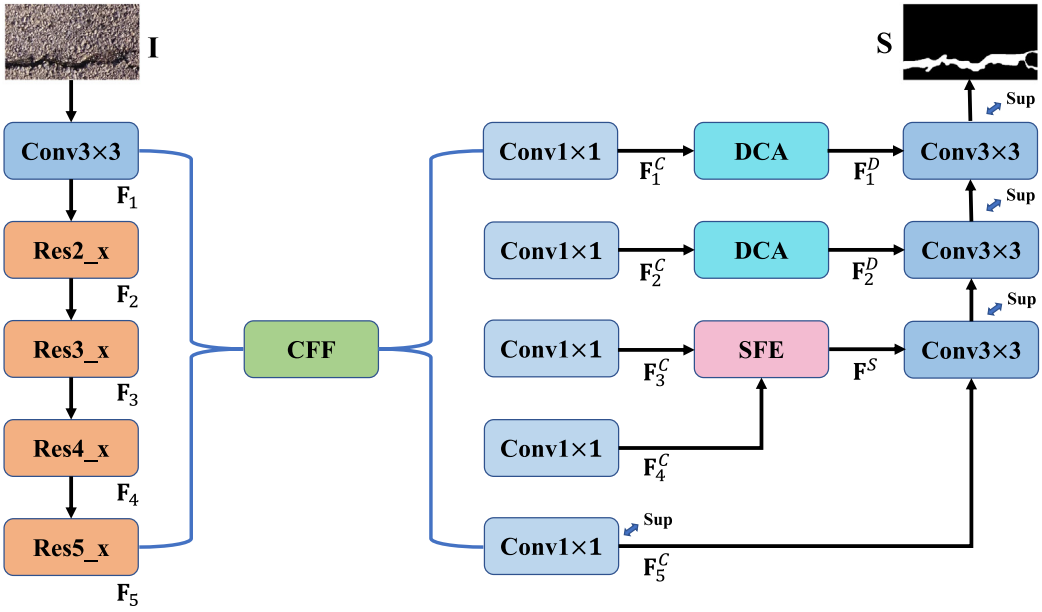


Fig. 1. The overall architecture of proposed SMINet.

(2020) developed a novel encoder–decoder siamese framework for video salient object detection and designed the combination of self- and cross-attention modules to preserve the spatial–temporal saliency correlation.

2.2. Surface defect detection

Benefiting from the development of computer vision, saliency-based defect detection methods have achieved a huge process. Quintana et al. (2015) developed a simpler detection system which consists of hard shoulder detection, cell candidate proposal, and crack classification to split up the image into small grid cells and extract the texture features. Yu et al. (2018) proposed a coarse-to-fine model from three levels (*i.e.*, subimage level, region level, and pixel level) to distinguish the defect regions from noise. In recent years, deep learning is widely applied to the surface defect detection. Pandiyan et al. (2019) proposed an encoder–decoder convolutional neural networks which adopts the modified VGG-16 to achieve semantic segmentation of weld seam removal states. Luo et al. (2021) proposed a decoupled two-stage object detection framework, where multi-hierarchical aggregation blocks and locally non-local blocks are designed for localization and classification. Song et al. (2020) proposed a novel detection method which adopts channel weighted block and residual decoder block to integrate multi-level features. Besides, a residual refinement structure with 1d filters was deployed to refine the coarse features. Niu et al. (2021) employed adaptive pyramid graph and variation residual to capture the correlation description and enhance the detection of abnormal defects, which improves the robustness of railway defect detection. Zheng et al. (2021) designed an end-to-end residual U-structure framework where multi-scale and multi-level features are integrated, and s coordinate attention module is introduced to extract useful features. Besides, the development of imaging equipment also provides multiple possibilities for defect detection. Zhang et al. (2020b) leveraged the color line scan camera, strip light, and displacement platform to capture the rail surface defect images. Zhang et al. (2020d) combined the flexible eddy current array probe and cartesian coordinate robot to image the curved surface defects.

2.3. Multi-level fusion-based detection

Multi-level fusion as a mean of information interaction has been proven effective in the SOD. Chen et al. (2021) integrated the current level feature with each features which is at the lower level to exploit complementary information. Dong et al. (2021b) incorporated the top-level feature into the others to propagate the highest semantic representations. Zhang et al. (2020c) deployed deep high-resolution parallel structure to alleviate the problem of losing valuable information in the encoder phase. Moreover, Pang et al. (2020), Zhou et al. (2021a) integrated features of three adjacent layers to extract multi-scale information and be better able to handle scale variation.

3. Proposed framework

3.1. Overview of proposed SMINet

In this paper, we propose a novel defect detection method shown in Fig. 1 which consists of three key components, including the cross-layer feature fusion (CFF) module, semantic-aware feature extraction (SFE) module, and detail-aware context attention (DCA) module. First, multi-scale features $\{\mathbf{F}_i\}_{i=1}^{5}$ extracted from the encoder network (ResNet-18 He et al., 2016) are fed into the CFF module, where cross-layer fusion strategy is introduced to facilitate information exchange between different scale features, yielding the enhanced features $\{\mathbf{F}_i^C\}_{i=1}^{5}$. After that, the SFE module and the DCA module are deployed in high level features $\{\mathbf{F}_i^C\}_{i=3}^{5}$ and low level features $\{\mathbf{F}_i^C\}_{i=1}^{2}$, respectively. Concretely, to give a sufficient exploration of semantic descriptions, we employ the SFE module which leverages the correlation between $\mathbf{F}_3^C$ and $\mathbf{F}_4^C$ to generate the deep semantic feature $\mathbf{F}^S$. Then, the DAC module adopts a split attention mechanism to give more concerns on spatial details and filter out the background noise, acquiring the detail-aware features $\{\mathbf{F}_i^D\}_{i=1}^{2}$. Finally, by combining the outputs of all modules in a top-down way, we can generate the final saliency map $\mathbf{S}$.

3.2. Cross-layer feature fusion module

By employing the encoder network, we can obtain multi-level features, of which low-level features focus on the spatial details and high-level features contain rich location information. To enhance the representation of multi-level features, the existing methods (Pang et al., 2020; Zhou et al., 2021a) try to integrate features from adjacent encoder layers. Unfortunately, these methods fail to explore the relationship of features with large resolution discrepancies. Therefore, to give a sufficient exploration of the complementarity between multi-scale features is the departure point in designing the cross-layer feature fusion (CFF) module.

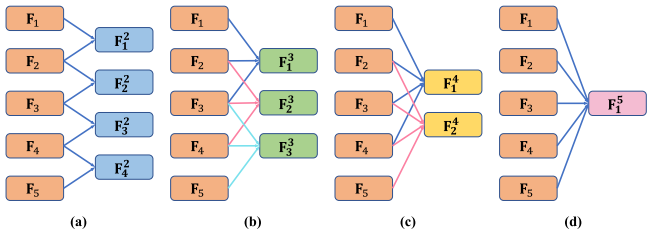


Fig. 2. Architecture of the cross-layer feature fusion module.

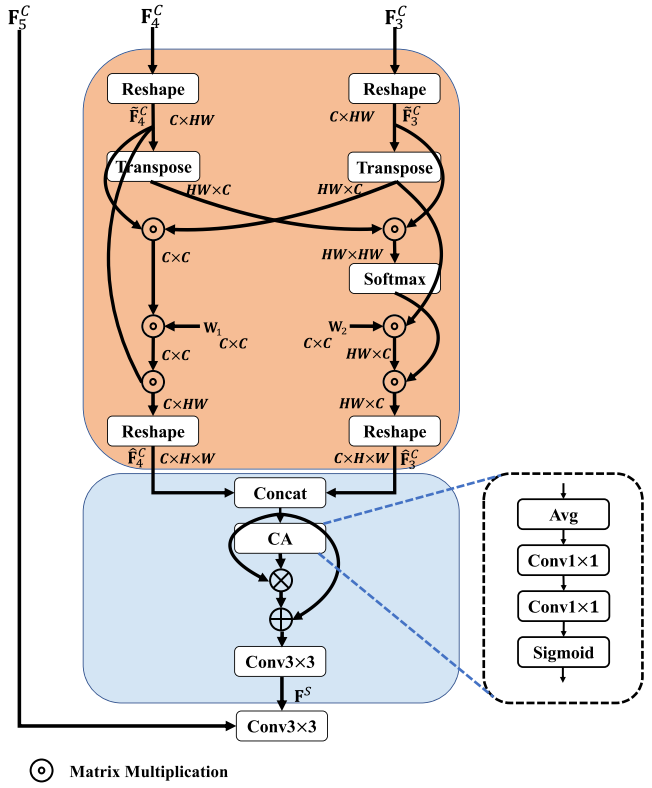


Fig. 3. Architecture of the semantic-aware feature extraction module.

Specifically, first, to capture the correlation of adjacent layer features, we propose adjacent interaction modules shown in Fig. 2 (a,b) to aggregate adjacent two- or three-layer features. In this stage, adjacent layer features $(\mathbf{F}_i, \mathbf{F}_{i+1})$ or $(\mathbf{F}_i, \mathbf{F}_{i+1}, \mathbf{F}_{i+2})$ are concatenated along channel dimension, where we further deploy a 3 × 3 convolutional layer to reduce the channel dimension. The entire process is formulated as follows

$$\begin{cases} \mathbf{F}_i^2 = f_{3\times3}(Cat(\mathbf{F}_i, U_{\times2}(\mathbf{F}_{i+1}))) \quad i = 1,2,3,4 \\ \mathbf{F}_i^3 = f_{3\times3}(Cat(\mathbf{F}_i, U_{\times2}(\mathbf{F}_{i+1}), U_{\times4}(\mathbf{F}_{i+2}))) \quad i = 1,2,3, \end{cases} \tag{1}$$

where Cat means the concatenation operation, $f_{3\times3}$ denotes the 3 × 3 convolutional layer, U denotes upsampling operation.

Meanwhile, we propose multi-level feature fusion modules, as shown in Fig. 2 (c,d), which fuse consecutive four- or five layers of encoder features. This alleviates the information dilution problem in the subsequent decoding phase. The entire process can be depicted as follows

$$\begin{cases} \mathbf{F}_i^4 = f_{3\times3}(Cat(\mathbf{F}_i, U_{\times2}(\mathbf{F}_{i+1}), U_{\times4}(\mathbf{F}_{i+2}), \\ \qquad U_{\times8}(\mathbf{F}_{i+3}))) \quad i = 1,2 \\ \mathbf{F}_i^5 = f_{3\times3}(Cat(\mathbf{F}_i, U_{\times2}(\mathbf{F}_{i+1}), U_{\times4}(\mathbf{F}_{i+2}), \\ \qquad U_{\times8}(\mathbf{F}_{i+3}), U_{\times16}(\mathbf{F}_{i+4}))) \quad i = 1 \end{cases}. \tag{2}$$

Finally, to enhance the representation of network, we integrate features with same resolution into the deep fused features $\{\mathbf{F}_i^C\}_{i=1}^5$. The corresponding process can be represented as follows

$$\begin{cases} \mathbf{F}_1^C = f_{3\times3}(Cat(\mathbf{F}_1^2, \mathbf{F}_1^3, \mathbf{F}_1^4, \mathbf{F}_1^5)) \\ \mathbf{F}_2^C = f_{3\times3}(Cat(\mathbf{F}_2^2, \mathbf{F}_2^3, \mathbf{F}_2^4)) \\ \mathbf{F}_3^C = f_{3\times3}(Cat(\mathbf{F}_3^2, \mathbf{F}_4^3)) \\ \mathbf{F}_4^C = f_{3\times3}(\mathbf{F}_4^2) \\ \mathbf{F}_5^C = \mathbf{F}_5. \end{cases} \tag{3}$$

Following this way, our network can endow the low-level features with abundant spatial structural information, and equip high-level features with sufficient semantic or location cues.

3.3. Semantic-aware feature extraction module

Most of existing methods attempt to dig the semantic information to indicate the location of defects, however, some of them (Zhang et al., 2020b; Dong et al., 2021a) overlook the correlation between features. Inspired by Zhang et al. (2019), Li et al. (2022), we attempt to extract the global cues via exploring the pertinence between $\mathbf{F}_3^C$ and $\mathbf{F}_4^C$. Meanwhile, we find that the effort (Li et al., 2022) only generates a single correlation matrix to guide two input features at the spatial dimension, and ignores the feature interaction at the channel dimension which is more worthy of attention in high-level features. Therefore, we propose a semantic-aware feature extraction (SFE) module, where two interaction matrices are generate to learn the complementary relationship at the channel and spatial dimensions simultaneously. As shown in Fig. 3, semantic-aware feature extraction (SFE) module consists of two stages (*i.e.*, feature interaction (orange part) and feature refinement (blue part)).

3.3.1. Feature interaction

As shown in Fig. 3 (orange part), we first reshape the input features $\mathbf{F}_3^C$ and $\mathbf{F}_4^C \in R^{C\times H\times W}$ to features $\tilde{\mathbf{F}}_3^C$ and $\tilde{\mathbf{F}}_4^C \in R^{C\times HW}$. Then, we perform corresponding operations on two branches, respectively. To be specific, for the branch one (left), we construct an interaction matrix $\mathbf{M}_1 \in R^{C\times C}$ by deploying matrix multiplication and softmax function on $\tilde{\mathbf{F}}_4^C$ and $(\tilde{\mathbf{F}}_3^C)^T$, where a learnable weight matrix $\mathbf{W}_1 \in R^{C\times C}$ is imposed for $\mathbf{M}_1$. The above process is formulated as follows:

$$\begin{cases} \tilde{\mathbf{F}}_3^C = RE(\mathbf{F}_3^C) \\ \tilde{\mathbf{F}}_4^C = RE(\mathbf{F}_4^C) \\ \mathbf{M}_1 = \sigma(\tilde{\mathbf{F}}_4^C \odot (\tilde{\mathbf{F}}_3^C)^T \odot \mathbf{W}_1), \end{cases} \tag{4}$$

where $\odot$ means the matrix multiplication, RE is reshape operation, σ is the softmax function, and $()^T$ is transposition operation. After that, we employ matrix multiplication followed by a reshape operation on $\tilde{\mathbf{F}}_4^C$ and $\mathbf{M}_1$, yielding the semantic feature $\hat{\mathbf{F}}_4^C \in R^{C\times H\times W}$.

$$\hat{\mathbf{F}}_4^C = RE(\tilde{\mathbf{F}}_4^C \odot \mathbf{M}_1). \tag{5}$$

Following this way, we can give more concerns on the channel dimension of high-level deep features.

For the right branch, we first compute the similar matrix $\mathbf{M}_2 \in R^{HW\times HW}$ via matrix multiplication and softmax function to find the correlation between $\mathbf{F}_3^C$ and $\mathbf{F}_4^C$. Then, we define a learnable weight matrix $\mathbf{W}_2 \in R^{C\times C}$ which provides a weight for the feature $\tilde{\mathbf{F}}_3^C$. After multiplying weighted $\tilde{\mathbf{F}}_3^C$ and matrix $\mathbf{M}_2$, we can generate another semantic feature $\hat{\mathbf{F}}_3^C \in R^{C\times H\times W}$, which is represented as follows:

$$\begin{cases} \mathbf{M}_2 = \sigma((\tilde{\mathbf{F}}_4^C)^T \odot \tilde{\mathbf{F}}_3^C) \\ \hat{\mathbf{F}}_3^C = RE((\tilde{\mathbf{F}}_3^C)^T \odot \mathbf{W}_2 \odot \mathbf{M}_2). \end{cases} \tag{6}$$

In this way, we explore the correlation between adjacent layer features from the perspective of spatial dimension and channel dimension.

3.3.2. Feature refinement

After the first stage, the location of defect regions can be coarsely inferred, but there are still some disturbances of the irrelevant information. To further filter out the noise and optimize semantic features, we conduct the feature refinement stage based on the channel attention.

As shown in Fig. 3 (blue part), the concatenation result of $\hat{\mathbf{F}}_3^C$ and $\hat{\mathbf{F}}_4^C$ is fed into the channel attention (CA) block (Hu et al., 2018) shown in the dashed box. In detail, global average pooling compresses the input feature into a channel-wise vector, and two 1×1 convolutional layers followed by a sigmoid activation function map the vector to $[0, 1]$. After that, the semantic-aware features can be acquired via element-wise multiplication and summation. Besides, we add a 3×3 convolutional layer to the output of the SFE module to further improve the semantic cues. The whole process is formulated as follows:

$$\begin{cases} \mathbf{C} = Cat(\hat{\mathbf{F}}_3^C, \hat{\mathbf{F}}_4^C) \\ \mathbf{V} = \sigma(f_{1\times1,2}(g(\mathbf{C}))) \\ \mathbf{F}^S = f_{3\times3}(\mathbf{C} \oplus (\mathbf{C} \otimes \mathbf{V})), \end{cases} \tag{7}$$

where g denotes the global average pooling operation, σ means the sigmoid activation function, $f_{1\times1,2}$ means two 1×1 convolutional layers, $\oplus$ and $\otimes$ are element-wise summing and multiplication, respectively.

By constructing the SFE module, the proposed network has strong capabilities to locate defect regions. Meanwhile, we should note that feature $\mathbf{F}_5^C$ contains precise semantic information, and thus we do not deploy the SFE module on $\mathbf{F}_5^C$ and $\mathbf{F}_4^C$. Instead, we just concatenate $\mathbf{F}_5^C$ and $\mathbf{F}^S$, and add a 3×3 convolutional layer to fuse the two features.

3.4. Detail-aware context attention module

After the cross-layer feature fusion (CFF) module, a great deal of feature information is aggregated in $\mathbf{F}_1^C$ and $\mathbf{F}_2^C$, which not only include rich spatial structural details but also contain a certain amount of background noises introduced by multi-scale feature fusion. Under this condition, we propose the detail-aware context attention (DCA) module, as shown in Fig. 4, to better distinguish the defect regions from background. Concretely, we first evenly split the input feature $\mathbf{F}^C$ into four equal parts $\mathbf{F}^{C_j}(j = 1, 2, 3, 4)$ with 32 channels along the channel dimension and construct four branches. Then, we deploy the channel attention to the 1st and 3rd branches, and employ spatial attention to 2nd and 4th branches. Next, four representative features are generated by multiplying the attention weights, which are further concatenated to generate the final output $\{\mathbf{F}_i^D\}_{i=1}^2$. The entire process is depicted as follows

$$\begin{cases} \mathbf{F}_i^{C_j} = Sp(\mathbf{F}_i^C) \quad j = 1, 2, 3, 4 \\ \mathbf{V}_i^{C_j} = CA(\mathbf{F}_i^{C_j}) \quad j = 1, 3 \\ \mathbf{W}_i^{C_j} = SA(\mathbf{F}_i^{C_j}) \quad j = 2, 4 \\ \mathbf{F}_i^D = Cat(\mathbf{F}_i^{C_1} \otimes \mathbf{V}_i^{C_1}, \mathbf{F}_i^{C_3} \otimes \mathbf{V}_i^{C_3}, \\ \qquad \mathbf{F}_i^{C_2} \otimes \mathbf{W}_i^{C_2}, \mathbf{F}_i^{C_4} \otimes \mathbf{W}_i^{C_4}) \end{cases}, \tag{8}$$

where $Sp(\cdot)$ means split operation, and CA and SA denote channel attention and spatial attention,respectively.

3.5. Loss function

To further promote the detection performance of our network, we adopt deep supervision to train our model, as shown in Fig. 1. Besides, we introduce F-measure loss (Zhao et al., 2019) to compute the loss value of our network. The loss function L_f is formulated as follows:

$$L_f(\mathbf{G}, \mathbf{S}) = \frac{1}{N} \sum_{i=1}^{N} (1 - \frac{(1+\beta^2) TP_i}{H_i}), \tag{9}$$

where $\mathbf{G}$ and $\mathbf{S}$ denote the saliency groundtruth and predicted saliency map, and N is batch size. $H = \beta^2(TP + FN) + (TP + FP)$, where TP, FP, and FN denote true positive, false positive, and false negative, respectively.

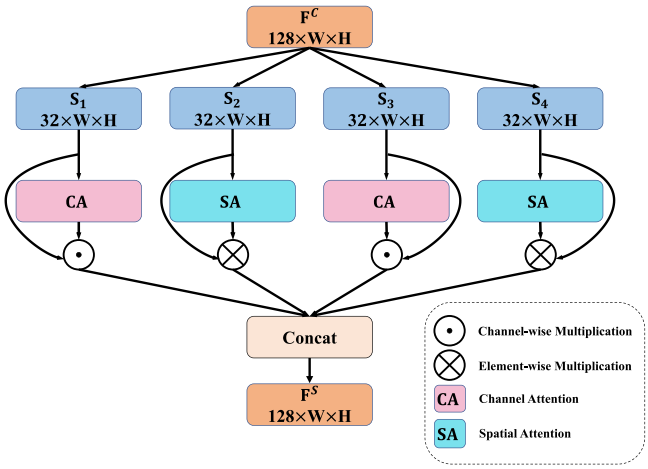


Fig. 4. Architecture of the detail-aware context attention module.

4. Experiments and analyses

4.1. Implementation details and evaluation metrics

4.1.1. Implementation details

We train and test our network on four public defect datasets, including road defect dataset, steel surface defect dataset, magnetic tile defect dataset, and DAGM2007 dataset. The road defect dataset includes 2512 images for training and 856 images for testing. The steel surface defect dataset contains 2160 images for training and 360 images for testing. The magnetic tile defect dataset contains 392 images, of which 349 images are selected as training set and 43 images as test set. The DAGM2007 dataset includes 2100 images with ten types of defects, 1026 images are training set, other 1054 images are test set. Before training our proposed SMINet, all images are resized to 224×224 to reduce the occupation of computing sources. Moreover, we implement SMINet on the Ubuntu system with Pytorch toolbox and adopt a workstation equipped with RTX2080 Ti GPU (with 12G memory) to accelerate the training process. Besides, we select the RMSprop (Hinton et al., 2012) optimizer to train model and minimize the loss function. The learning rate and momentum are set $1e-4$ and 0.9, respectively.

4.1.2. Evaluation metrics

To evaluate the performance of our proposed SMINet, we employ six popular evaluation methods, including mean absolute error (MAE) (Perazzi et al., 2012), F-measure (FM) (Achanta et al., 2009), weighted F-measure (WF) (Margolin et al., 2014), E-measure (EM) (Fan et al., 2018), structure-measure (SM) (Fan et al., 2017), and Precision-Recall (PR) curve. The MAE is the smaller the better, the other five metrics are bigger the better. In this article, we adopt the widely used evaluation code based on python and matlab.

4.2. Comparison with state-of-the-art methods

In experiments, we compare the proposed SMINet with other 14 state-of-the-art salient object detection methods, including three SOD methods for surface defect (*i.e.*, MC (Zhang et al., 2020b), PGANet (Dong et al., 2019), and AEP (Dong et al., 2021a)), and eleven SOD methods for other scenarios, of which four methods (*i.e.*, SF (Perazzi et al., 2012), DM (Yang et al., 2013), GB (Wei et al., 2012), and BC (Zhu et al., 2014)) are traditional methods and other seven methods (*i.e.*, F3Net (Wei et al., 2020), ITSD (Zhou et al., 2020), BASNet (Qin et al., 2019), GATE (Zhao et al., 2020), CPD (Wu et al., 2019), VST (Liu et al., 2021), and CorrNet (Li et al., 2022)) are based on deep learning. For fair comparison, we retrain the compared models, where the source codes are released by the authors, on four surface defect datasets.

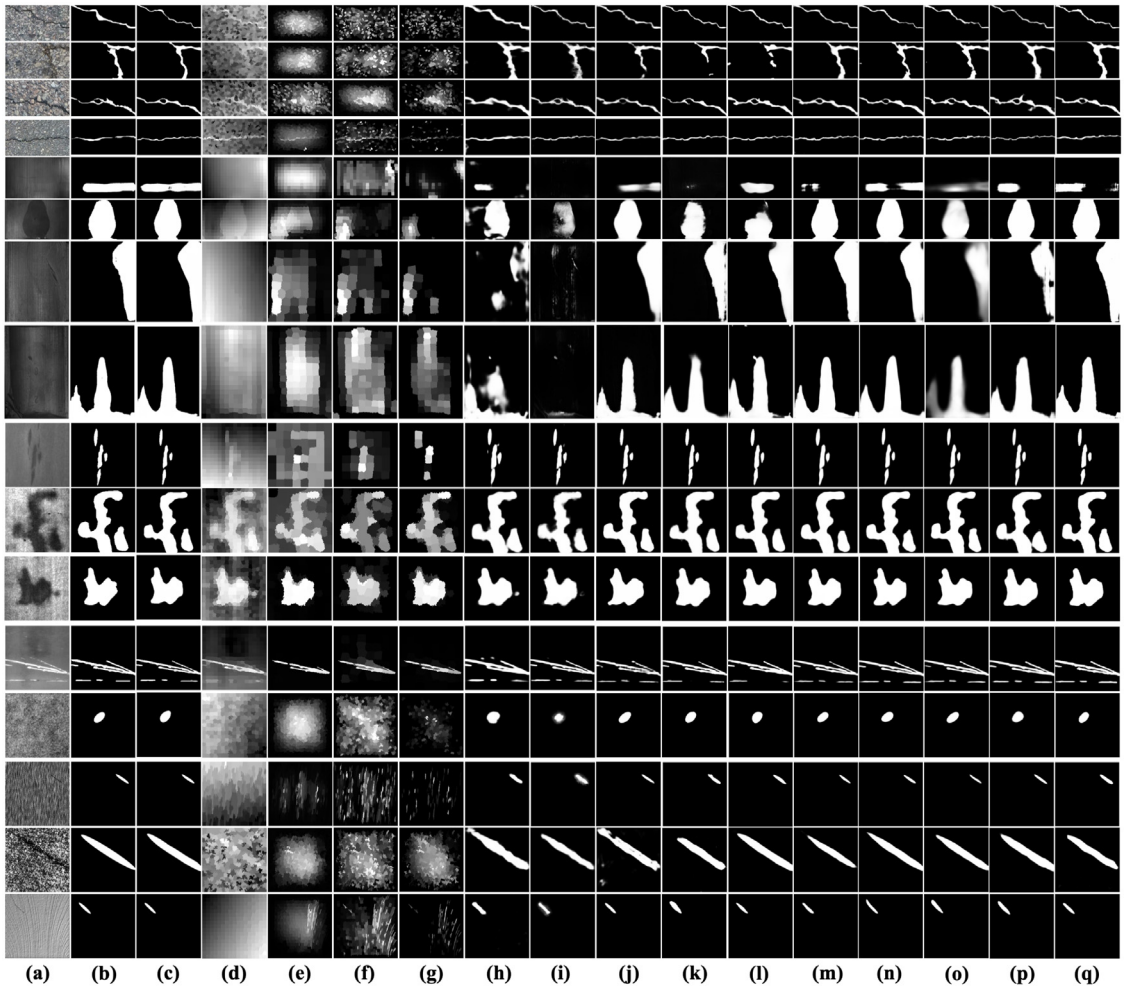


Fig. 5. Visual comparison of saliency maps. (a) Source image. (b) Groundtruth. (c) Ours. (d) SF (Perazzi et al., 2012). (e) DM (Yang et al., 2013). (f) GB (Wei et al., 2012). (g) BC (Zhu et al., 2014). (h) F3 (Wei et al., 2020). (i) ITSD (Zhou et al., 2020). (j) MC (Zhang et al., 2020b). (k) CorrNet (Li et al., 2022). (l) BAS (Qin et al., 2019). (m) GATE (Zhao et al., 2020). (n) PGA (Dong et al., 2019). (o) CPD (Wu et al., 2019). (p) VST (Liu et al., 2021). (q) AEP (Dong et al., 2021a).

Table 1
Ablation study on different architectures and backbones.

	Settings								Steel Defect Dataset				
#	Base	SFE	DCA	CFF	ResNet-18	ResNet-34	Vgg-16	Decoder	MAE	FM	WF	EM	SM
1	✓				✓				0.0297	0.7997	0.7962	0.9448	0.8400
2	✓	✓			✓				0.0281	0.8203	0.8255	0.9424	0.8482
3	✓	✓	✓		✓				0.0277	0.8505	0.8305	0.9572	0.8509
4	✓	✓	✓	✓			✓		0.0252	0.8462	0.8419	0.9574	0.8688
5	✓	✓	✓	✓		✓			0.0253	0.8485	0.8462	0.9581	0.8699
6	✓	✓	✓	✓	✓			(a)	0.0267	0.8223	0.8320	0.9458	0.8691
7	✓	✓	✓	✓	✓			(b)	0.0258	0.8429	0.8326	0.9560	0.8641
8	✓	✓	✓	✓	✓			(ours)	**0.0247**	**0.8623**	**0.8534**	**0.9592**	**0.8721**

4.2.1. Qualitative comparison

To demonstrate the effectiveness and accuracy of our proposed SMINet, we list some test results of typical defects as shown in Fig. 5, where we can find that SMINet with three designed modules gives full play to its advantages.

Advantages in low constrast. For the low contrast images, as shown in Fig. 5 (5th, 6th, and 7th rows), where the defect regions have similar characteristics to the background, most methods fail to distinguish the defects from background noise. For example, in the 5th row image containing the magnetic tile defects, which are difficult to identify by manual means, the deep learning-based SOD methods (*i.e.*, h-q) are incapable of segmenting the defect region integrally, and other traditional SOD methods (*i.e.*, d-g) even cannot recognize it. On the contrary, from our result (c), we can find that SMINet achieves better performance in highlighting the defects.

Advantages in defect details. For the defect regions with complex appearance, some details such as boundaries and small areas are ignored. For example, in the 2nd row image, which contains three interacting lathy pavement cracks, the saliency maps generated from other methods have unclear boundary details. In contrast, as shown in Fig. 5(c), our method not only restores clear frontiers between defect regions and background but also achieves the digging of small area information.

Table 2
Quantitative comparisons with 15 methods on four datasets. The top two results are marked in red and green.

Dataset	Road Defect Dataset					Steel Defect Dataset					Magnetic Tile Defect Dataset					DAGM 2007				
Metric	MAE↓	FM↑	WF↑	EM↑	SM ↑	MAE↓	FM↑	WF↑	EM↑	SM↑	MAE↓	FM↑	WF↑	EM↑	SM↑	MAE↓	FM↑	WF↑	EM↑	SM↑
SF (Perazzi et al., 2012)	0.4711	0.1489	0.0668	0.5018	0.7563	0.4417	0.2018	0.1538	0.4889	0.4166	0.4899	0.0248	0.0893	0.5479	0.3564	0.4888	0.0331	0.0258	0.6382	0.3500
DM (Yang et al., 2013)	0.2605	0.0775	0.0628	0.4578	0.4236	0.4140	0.3040	0.2234	0.6793	0.4534	0.4570	0.0561	0.0645	0.4323	0.3405	0.2409	0.0363	0.0281	0.3357	0.4280
GB (Wei et al., 2012)	0.2508	0.1162	0.0895	0.5370	0.4411	0.1664	0.3615	0.2932	0.6483	0.5685	0.2358	0.0772	0.0713	0.3851	0.4210	0.2483	0.0365	0.0288	0.3827	0.4225
BC (Zhu et al., 2014)	0.1306	0.1733	0.1211	0.5944	0.4846	0.1560	0.4699	0.3785	0.7356	0.5925	0.1618	0.0880	0.0631	0.4124	0.4341	0.0714	0.0342	0.0299	0.4851	0.4766
F3 (Wei et al., 2020)	0.0391	0.5796	0.5851	0.8274	0.6573	0.0381	0.7033	0.7442	0.8853	0.8274	0.0589	0.4274	0.5023	0.6046	0.6935	0.0128	0.5109	0.5645	0.7702	0.7668
ITSD (Zhou et al., 2020)	0.0370	0.5940	0.6055	0.8195	0.7599	0.0289	0.7875	0.8144	0.9223	0.8753	0.1056	0.1371	0.0944	0.4184	0.4903	0.0095	0.5239	0.6811	0.7393	0.8581
MC (Zhang et al., 2020b)	0.0360	0.6120	0.5867	0.8611	0.7358	0.0271	0.8028	0.8199	0.9346	0.8690	0.0271	0.4574	0.5543	0.6394	0.7564	0.0085	0.7828	0.8101	0.9303	0.8870
CorrNet (Li et al., 2022)	0.0351	0.6736	0.5921	0.8713	0.7058	0.0303	0.8326	0.8271	0.9462	0.8550	0.0863	0.4055	0.2796	0.5956	0.6075	0.0069	0.7905	0.7959	0.9357	0.8711
BASNet (Qin et al., 2019)	0.0342	0.6638	0.6277	0.8940	0.7368	0.0251	0.8222	0.8391	0.9501	0.8694	0.0346	0.6594	0.7102	0.8284	0.8186	0.0061	0.8248	0.8301	0.9523	0.8949
GATE (Zhao et al., 2020)	0.0335	0.6350	0.6203	0.8700	0.7481	0.0266	0.7933	0.8197	0.9287	0.8683	0.0183	0.6611	0.8086	0.8160	0.8796	0.0059	0.7906	0.8265	0.9282	0.8999
PGANet (Dong et al., 2019)	0.0331	0.6560	0.6265	0.8935	0.7378	0.0288	0.7761	0.7941	0.9222	0.8537	0.0187	0.6392	0.7870	0.7989	0.8692	0.0083	0.7391	0.7608	0.9048	0.8628
CPD (Wu et al., 2019)	0.0330	0.6597	0.6349	0.8998	0.7419	0.0244	0.8186	0.8315	0.9453	0.8714	0.025	0.5123	0.6220	0.6761	0.7963	0.0051	0.7851	0.8359	0.9291	0.9041
VST (Liu et al., 2021)	0.0328	0.6413	0.6341	0.8766	0.7556	0.0267	0.7934	0.8172	0.9336	0.8664	0.0210	0.6251	0.7359	0.7917	0.8494	0.0068	0.6943	0.7437	0.8877	0.8600
AEP (Dong et al., 2021a)	0.0311	0.6697	0.6332	0.9006	0.7500	0.0258	0.7979	0.8274	0.9315	0.8684	0.0285	0.7705	0.7851	0.9068	0.8473	0.0062	0.8082	0.8126	0.9384	0.8784
Ours	0.0304	0.7120	0.6585	0.9157	0.7408	0.0247	0.8623	0.8534	0.9592	0.8721	0.0158	0.8393	0.8640	0.9543	0.8956	0.0053	0.8561	0.8516	0.9580	0.9004

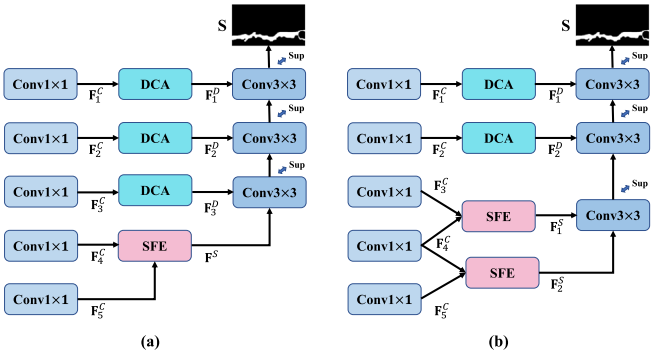


Fig. 6. Architecture of different decoder structure.

Advantages in multiple defects. Multi-object detection has always been a major sticking point in saliency detection. For example, in the images (*i.e.*, 9th and 12th rows), where multiple defect regions are adjacent to each other, most methods have difficulty in distinguishing the frontier between these defects, but our method can yield the saliency map that is closet to the ground truth and contains a clear position boundary between different defects.

Advantages in complex background. For the images with complex background, as shown in Fig. 5 (14th, 15th, and 16th rows), traditional SOD methods (*i.e.*, d-g) falsely detect the background textures as defects and most existing deep learning-based SOD methods fail to uniformly segment defect regions, yet our SMINet can highlight defect regions accurately.

4.2.2. Quantitative comparison

To illustrate the superiority of our method, we plot the PR and F-measure curves of all methods on four defect datasets, as shown in Fig. 7. As visible, it is clear that our SMINet has a position closest to the upper right corner of the diagrams. Besides, the proposed SMINet is in a higher position for most thresholds on the F-measure curve compared with other methods. Moreover, we provide the numerical results in terms of five evaluation metrics (*i.e.*, MAE, FM, WF, EM, and SM) as shown in Table 2, where we can find the proposed SMINet achieves competitive performance against other methods in five metrics on the four datasets. Particularly, compared with the novel method VST, our method gains improvements of 34.3% in terms of FM, 17.4% in terms of WF, 20.5% in terms of EM, and 5.4% in terms of SM, and gains a decrease 24.8% in terms of MAE. In addition, compared with state-of-the-art method AEP, SMINet also achieves significant performance improvements, where FM, EM, and SM increase by 8.9%, 5.2%, and 5.7%, MAE decreases by 44.6%. The above numerical comparison results are calculated on the magnetic tile defect dataset. Furthermore, we list the average running time and model size of fifteen detection methods. As shown in Table 3, compared with deep learning-based methods, SMINet has a relatively small model size, meeting the demand for lightweight. For the average running time, the FPS of our proposed SMINet outperforms most comparison methods.

Table 3
Comparison of the model size and the average running time.

Method	SF	DM	GB	BC	F3	ITSD	MC	CorrNet
Time(s)	0.32	0.26	0.005	0.054	0.010	0.02	0.028	0.041
Size(MB)	–	–	–	–	98	65	163	16
Method	BASNet	GATE	PGANet	CPD	VST	AEP	Ours	
Time(s)	0.046	0.26	0.025	0.055	0.046	0.055	0.06	
Size(MB)	332	491	210	183	170	114	57	

4.3. Ablation study

To demonstrate the effectiveness of each component (*i.e.*, cross-layer feature fusion (CFF), semantic-aware feature extraction (SFE), and detail-aware context attention (DCA) modules), we conduct several ablation experiments by gradually adding modules to the baseline. Furthermore, we also conduct two other experiments which replace ResNet-18 of our SMINet with ResNet-34 or Vgg-16 to verify the feature extraction capability of different backbones. As shown in Fig. 8, we list visualization results of ablation experiments. Compared with the baseline (c), it can be found that the SFE module can capture the defect location, but the detailed information (*i.e.*, boundary and structure) is scarce. After adding DCA module, as shown in Fig. 8(e), it is clear that the context information is dug and background noise is suppressed. Finally, on the basis of the above, we add CFF module to form the final network, where more defect information is recovered by fusing multi-scale features.

To intuitively illustrate the performance of each module, we provide the quantitative results in Table 1. From these numerical results, we can find that with the increase of modules, the performance of the proposed SMINet is promoted progressively. In detail, compared with the baseline, SMINet greatly improves the defect detection performance, where FM and SM are improved by 7.2% and 3.8%, and MAE is decreased by 16.8%. In addition, compared with different backbones, as shown in Table 1 (#4 and #5), the performance of SMINet is improved to a certain extent. Furthermore, we compare the decoder structure our method with two other structure as shown in Fig. 6 to verify the validity, the

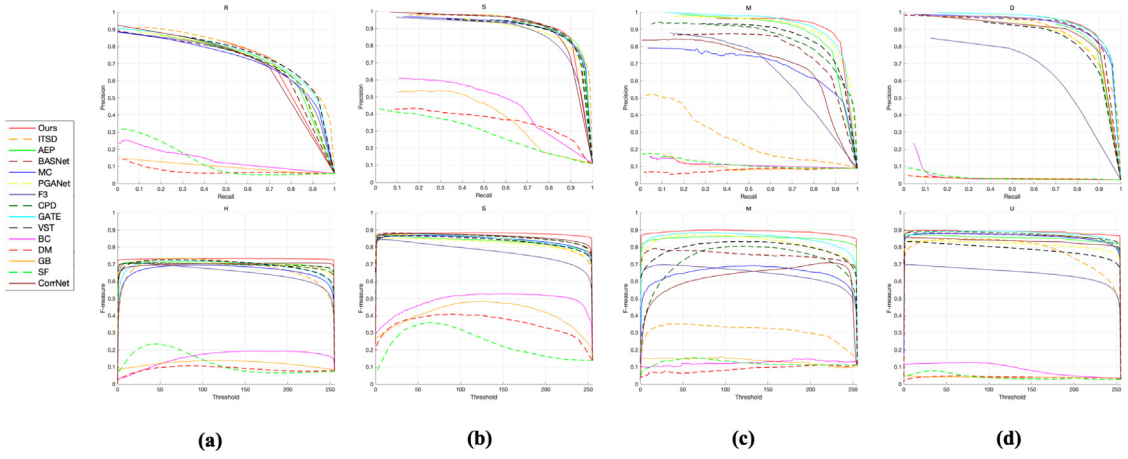


Fig. 7. PR and F-measure curves of different methods. (a) Results on the Road Defect Dataset. (b) Results on the Steel Defect Dataset. (c) Results on the Magnetic Tile Defect Dataset. (d) Results on the DAGM2007 Defect Dataset.

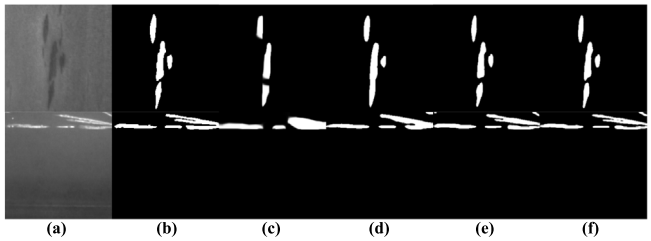


Fig. 8. Visualization results of ablation studies. (a) Soure image, (b) Ground truth, (c) #1, (d) #2, (e) #3, (f) #8.

comparison results are list in Table 1 (#6, #7 and #8). From the above results, we can see that our decoder structure has obvious advantages, where FM and SM are improved by average of 3.3% and 0.6%.

5. Conclusion

In this paper, a novel salient object detection for surface defects is proposed, named SMINet, which is capable of integrating multi-scale features and capturing detailed and semantic information. Firstly, to take full advantage of multi-level features, we design a cross-layer feature fusion (CFF) module that integrates multiple adjacent levels of the encoder. Particularly, the detailed and semantic cues are aggregated to the low- and high-level features. Secondly, we propose semantic-aware feature extraction (SFE) module that exploits GCN-like structure to explore the semantic relationship of defects. In addition, we utilize the split attention mechanism to construct detail-aware context attention (DCA) module which filters out background noise and enhances the detection accuracy. Comprehensive experiments over four defect datasets validate that our proposed SMINet achieves competitive or even better performance than state-of-the-art methods.

CRediT authorship contribution statement

Bin Wan: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Xiaofei Zhou:** Data curation, Writing – review & editing. **Yaoqi Sun:** Data curation. **Zunjie Zhu:** Visualization, Investigation. **Haibing Yin:** Software, Validation. **Ji Hu:** Writing – review & editing. **Jiyong Zhang:** Visualization, Writing – review & editing. **Chenggang Yan:** Conceptualization, Funding acquisition, Resources, 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.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grants 2020YFB1406604; the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grants GK229909299001-009; the National Natural Science Foundation of China under Grants 62271180, 62171002, 61901145, U21B2024, 61931008, 62071415, 61972123, 62001146; the Zhejiang Province Nature Science Foundation of China under Grants LR17F030006, LY19F030022, LZ22F020003; the Hangzhou Dianzi University (HDU) and the China Electronics Corporation DATA (CEC-DATA) Joint Research Center of Big Data Technologies under Grants KYH063120009; the 111 Project under Grants D17019; and the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grants GK219909299001-407.

References

Achanta, Radhakrishna, Hemami, Sheila, Estrada, Francisco, Susstrunk, Sabine, 2009. Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1597–1604.

Chen, Tianyou, Hu, Xiaoguang, Xiao, Jin, Zhang, Guofeng, Wang, Shaojie, 2021. BINet: Bidirectional interactive network for salient object detection. Neurocomputing 465, 490–502.

Chen, Hanshen, Lin, Huiping, Yao, Minghai, 2019. Improving the efficiency of encoder-decoder architecture for pixel-level crack detection. IEEE Access 7, 186657–186670.

Chen, Liang-Chieh, Papandreou, George, Kokkinos, Iasonas, Murphy, Kevin, Yuille, Alan L, 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Cheng, Xun, Yu, Jianbo, 2020. RetinaNet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection. IEEE Trans. Instrum. Meas. 70, 1–11.

Dong, Hongwen, Song, Kechen, He, Yu, Xu, Jing, Yan, Yunhui, Meng, Qinggang, 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, Hongwen, Song, Kechen, Wang, Yanyan, Yan, Yunhui, Jiang, Peng, 2021a. Automatic inspection and evaluation system for pavement distress. IEEE Trans. Intell. Transp. Syst..

Dong, Bo, Zhou, Yan, Hu, Chuanfei, Fu, Keren, Chen, Geng, 2021b. BCNet: bidirectional collaboration network for edge-guided salient object detection. Neurocomputing 437, 58–71.

Fan, Deng-Ping, Cheng, Ming-Ming, Liu, Yun, Li, Tao, Borji, Ali, 2017. Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4548–4557.

Fan, Deng-Ping, Gong, Cheng, Cao, Yang, Ren, Bo, Cheng, Ming-Ming, Borji, Ali, 2018. Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421.

Fan, Deng-Ping, Zhou, Tao, Ji, Ge-Peng, Zhou, Yi, Chen, Geng, Fu, Huazhu, Shen, Jianbing, Shao, Ling, 2020. Inf-Net: Automatic COVID-19 lung infection segmentation from CT images. IEEE Trans. Med. Imaging 39 (8), 2626–2637.

Hare, Sam, Golodetz, Stuart, Saffari, Amir, Vineet, Vibhav, Cheng, Ming-Ming, Hicks, Stephen L, Torr, Philip HS, 2015. Struck: Structured output tracking with kernels. IEEE Trans. Pattern Anal. Mach. Intell. 38 (10), 2096–2109.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

He, Tong, Zhang, Zhi, Zhang, Hang, Zhang, Zhongyue, Xie, Junyuan, Li, Mu, 2019. Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 558–567.

Hinton, Geoffrey, Srivastava, Nitish, Swersky, Kevin, 2012. Rmsprop: Divide the gradient by a running average of its recent magnitude. In: Neural Networks for Machine Learning, Coursera Lecture 6e. p. 13.

Hu, Jie, Shen, Li, Sun, Gang, 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Ji, Yuzhu, Zhang, Haijun, Jie, Zequn, Ma, Lin, Wu, QM Jonathan, 2020. CASNet: A cross-attention siamese network for video salient object detection. IEEE Trans. Neural Netw. Learn. Syst. 32 (6), 2676–2690.

Ju, Ran, Ge, Ling, Geng, Wenjing, Ren, Tongwei, Wu, Gangshan, 2014. Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 1115–1119.

Konovalenko, Ihor, Maruschak, Pavlo, Brezinová, Janette, Prentkovskis, Olegas, Brezina, Jakub, 2022a. Research of U-Net-based CNN architectures for metal surface defect detection. Machines 10 (5), 327.

Konovalenko, Ihor, Maruschak, Pavlo, Kozbur, Halyna, Brezinová, Janette, Brezina, Jakub, Nazarevich, Bohdan, Shkira, Yaroslav, 2022b. Influence of uneven lighting on quantitative indicators of surface defects. Machines 10 (3), 194.

Li, Gongyang, Liu, Zhi, Bai, Zhen, Lin, Weisi, Ling, Haibin, 2022. Lightweight salient object detection in optical remote sensing images via feature correlation. IEEE Trans. Geosci. Remote Sens..

Li, Feng, Xi, QingGang, 2021. DefectNet: toward fast and effective defect detection. IEEE Trans. Instrum. Meas. 70, 1–9.

Li, Yuyuan, Zhang, Dong, Lee, Dah-Jye, 2019. Automatic fabric defect detection with a wide-and-compact network. Neurocomputing 329, 329–338.

Liu, Nian, Zhang, Ni, Wan, Kaiyuan, Shao, Ling, Han, Junwei, 2021. Visual saliency transformer. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4722–4732.

Lu, Xiankai, Ni, Bingbing, Ma, Chao, Yang, Xiaokang, 2019. Learning transform-aware attentive network for object tracking. Neurocomputing 349, 133–144.

Luo, Jiaxiang, Yang, Zhiyu, Li, Shipeng, Wu, Yilin, 2021. FPCB surface defect detection: A decoupled two-stage object detection framework. IEEE Trans. Instrum. Meas. 70, 1–11.

Mak, Kai-Ling, Peng, P., Yiu, Ka Fai Cedric, 2009. Fabric defect detection using morphological filters. Image Vis. Comput. 27 (10), 1585–1592.

Margolin, Ran, Zelnik-Manor, Lihi, Tal, Ayellet, 2014. How to evaluate foreground maps? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255.

Ng, Hui-Fuang, 2006. Automatic thresholding for defect detection. Pattern Recognit. Lett. 27 (14), 1644–1649.

Ngan, Henry YT, Pang, Grantham KH, Yung, Siu-Pang, Ng, Michael K, 2005. Wavelet based methods on patterned fabric defect detection. Pattern Recognit. 38 (4), 559–576.

Niu, Menghui, Wang, Yanyan, Song, Kechen, Wang, Qi, Zhao, Yongjie, Yan, Yunhui, 2021. An adaptive pyramid graph and variation residual-based anomaly detection network for rail surface defects. IEEE Trans. Instrum. Meas. 70, 1–13.

Pandiyan, Vigneashwara, Murugan, Pushparaja, Tjahjowidodo, Tegoeh, Caesarendra, Wahyu, Manyar, Omey Mohan, Then, David Jin Hong, 2019. In-process virtual verification of weld seam removal in robotic abrasive belt grinding process using deep learning. Robot. Comput.-Integr. Manuf. 57, 477–487.

Pang, Youwei, Zhao, Xiaoqi, Zhang, Lihe, Lu, Huchuan, 2020. Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9413–9422.

Park, Youngmin, Yang, Hyun S., 2019. Convolutional neural network based on an extreme learning machine for image classification. Neurocomputing 339, 66–76.

Perazzi, Federico, Krähenbühl, Philipp, Pritch, Yael, Hornung, Alexander, 2012. Saliency filters: Contrast based filtering for salient region detection. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 733–740.

Qin, Xuebin, Zhang, Zichen, Huang, Chenyang, Gao, Chao, Dehghan, Masood, Jagersand, Martin, 2019. Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7479–7489.

Quintana, Marcos, Torres, Juan, Menéndez, José Manuel, 2015. A simplified computer vision system for road surface inspection and maintenance. IEEE Trans. Intell. Transp. Syst. 17 (3), 608–619.

Sharifzadeh, M, Amirfattahi, R, Sadri, S, Alirezaee, S, Ahmadi, M, 2008. Detection of steel defect using the image processing algorithms. In: The International Conference on Electrical Engineering. Vol. 6. Military Technical College, pp. 1–7, 6th International Conference on Electrical Engineering ICEENG 2008.

Song, Guorong, Song, Kechen, Yan, Yunhui, 2020. EDRNet: Encoder–decoder residual network for salient object detection of strip steel surface defects. IEEE Trans. Instrum. Meas. 69 (12), 9709–9719.

Tajeripour, Farshad, Kabir, Ehsanollah, Sheikhi, Abbas, 2007. Fabric defect detection using modified local binary patterns. EURASIP J. Adv. Signal Process. 2008, 1–12.

Wang, Xiaofei, Jiang, Lai, Li, Liu, Xu, Mai, Deng, Xin, Dai, Lisong, Xu, Xiangyang, Li, Tianyi, Guo, Yichen, Wang, Zulin, et al., 2021. Joint learning of 3D lesion segmentation and classification for explainable COVID-19 diagnosis. IEEE Trans. Med. Imaging 40 (9), 2463–2476.

Wang, Wenguan, Shen, Jianbing, Shao, Ling, 2017. Video salient object detection via fully convolutional networks. IEEE Trans. Image Process. 27 (1), 38–49.

Wang, Mi, Wu, 2021. A real-time steel surface defect detection approach with high accuracy. IEEE Trans. Instrum. Meas. 71.

Wei, Jun, Wang, Shuhui, Huang, Qingming, 2020. F^3Net: Fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. pp. 12321–12328.

Wei, Yichen, Wen, Fang, Zhu, Wangjiang, Sun, Jian, 2012. Geodesic saliency using background priors. In: European Conference on Computer Vision. Springer, pp. 29–42.

Wu, Zhe, Su, Li, Huang, Qingming, 2019. Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3907–3916.

Yang, Chuan, Zhang, Lihe, Lu, Huchuan, Ruan, Xiang, Yang, Ming-Hsuan, 2013. Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3166–3173.

Yang, Fan, Zhang, Lei, Yu, Sijia, Prokhorov, Danil, Mei, Xue, Ling, Haibin, 2019. Feature pyramid and hierarchical boosting network for pavement crack detection. IEEE Trans. Intell. Transp. Syst. 21 (4), 1525–1535.

Yu, Haomin, Li, Qingyong, Tan, Yunqiang, Gan, Jinrui, Wang, Jianzhu, Geng, Yangli-ao, Jia, Lei, 2018. A coarse-to-fine model for rail surface defect detection. IEEE Trans. Instrum. Meas. 68 (3), 656–666.

Zhai, Yingjie, Fan, Deng-Ping, Yang, Jufeng, Borji, Ali, Shao, Ling, Han, Junwei, Wang, Liang, 2021. Bifurcated backbone strategy for rgb-d salient object detection. IEEE Trans. Image Process. 30, 8727–8742.

Zhang, Qijian, Cong, Runmin, Li, Chongyi, Cheng, Ming-Ming, Fang, Yuming, Cao, Xiaochun, Zhao, Yao, Kwong, Sam, 2020a. Dense attention fluid network for salient object detection in optical remote sensing images. IEEE Trans. Image Process. 30, 1305–1317.

Zhang, Xin, Liu, Yicheng, Huo, Chunlei, Xu, Nuo, Wang, Lingfeng, Pan, Chunhong, 2022. PSNet: Perspective-sensitive convolutional network for object detection. Neurocomputing 468, 384–395.

Zhang, Libao, Ma, Jie, 2021. Salient object detection based on progressively supervised learning for remote sensing images. IEEE Trans. Geosci. Remote Sens. 59 (11), 9682–9696.

Zhang, Defu, Song, Kechen, Xu, Jing, He, Yu, Niu, Menghui, Yan, Yunhui, 2020b. MCnet: Multiple context information segmentation network of no-service rail surface defects. IEEE Trans. Instrum. Meas. 70, 1–9.

Zhang, Si, Tong, Hanghang, Xu, Jiejun, Maciejewski, Ross, 2019. Graph convolutional networks: a comprehensive review. Comput. Soc. Netw. 6 (1), 1–23.

Zhang, Xue, Wang, Zheng, Hu, Qinghua, Ren, Jinchang, Sun, Meijun, 2020c. Boundary-aware High-resolution Network with region enhancement for salient object detection. Neurocomputing 418, 91–101.

Zhang, Weipeng, Wang, Chenglong, Xie, Fengqin, Zhang, Huayu, 2020d. Defect imaging curved surface based on flexible eddy current array sensor. Measurement 151, 107280.

Zhao, Kai, Gao, Shanghua, Wang, Wenguan, Cheng, Ming-Ming, 2019. Optimizing the f-measure for threshold-free salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8849–8857.

Zhao, Xiaoqi, Pang, Youwei, Zhang, Lihe, Lu, Huchuan, Zhang, Lei, 2020. Suppress and balance: A simple gated network for salient object detection. In: European Conference on Computer Vision. Springer, pp. 35–51.

Zheng, Zhouzhou, Yang, Huanbo, Zhou, Liang, Yu, Bin, Zhang, Yan, 2021. HLU 2-Net: A residual U-structure embedded U-Net with hybrid loss for tire defect inspection. IEEE Trans. Instrum. Meas. 70, 1–11.

Zhou, Xiaofei, Fang, Hao, Fei, Xiaobo, Shi, Ran, Zhang, Jiyong, 2021a. Edge-aware multi-level interactive network for salient object detection of strip steel surface defects. IEEE Access 9, 149465–149476.

Zhou, Xiaofei, Fang, Hao, Liu, Zhi, Zheng, Bolun, Sun, Yaoqi, Zhang, Jiyong, Yan, Chenggang, 2021b. Dense attention-guided cascaded network for salient object detection of strip steel surface defects. IEEE Trans. Instrum. Meas. 71, 1–14.

Zhou, Xiaofei, Shen, Kunye, Weng, Li, Cong, Runmin, Zheng, Bolun, Zhang, Jiyong, Yan, Chenggang, 2022. Edge-guided recurrent positioning network for salient object detection in optical remote sensing images. IEEE Trans. Cybern. http://dx.doi.org/10.1109/TCYB.2022.3163152.

Zhou, Huajun, Xie, Xiaohua, Lai, Jian-Huang, Chen, Zixuan, Yang, Lingxiao, 2020. Interactive two-stream decoder for accurate and fast saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9141–9150.

Zhu, Wangjiang, Liang, Shuang, Wei, Yichen, Sun, Jian, 2014. Saliency optimization from robust background detection. In: Proceedings of the IEEE Conference on Computer VVsion and Pattern Recognition. pp. 2814–2821.