Optics and Lasers in Engineering 163 (2023) 107488

Contents lists available at ScienceDirect

Optics and Lasers in Engineering

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

Defect classification for specular surfaces based on deflectometry and multi-modal fusion network

Jingtian Guan[a,b], Jingjing Fei[b], Wei Li[b], Xiaoke Jiang[b], Liwei Wu[b], Yakun Liu[b], Juntong Xi[a,c,d,*]

[a] School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] Sensetime Research, Shanghai 201908, China
[c] Shanghai Key Laboratory of Advanced Manufacturing Environment, Shanghai 200030, China
[d] State Key Laboratory of Mechanical System and Vibration, Shanghai 200240, China

ARTICLE INFO

Keywords:
Specular surface
Defect classification
Deflectometry
Dataset
Deep learning

ABSTRACT

Automated defect inspection for specular surfaces is still a challenge in the manufacturing industry because of their specular reflection property. Deflectometry provides surface information based on the captured fringe patterns through the reflection of the specular surfaces and has been widely applied in defect detection for specular surfaces. Conventional methods combined deflectometry with machine learning approaches, but the hand-crafted features need to be defined for each specific task. Combined with the deep neural network, the input images are obtained from deflectometry, and the network completes the identification of the defects. Nevertheless, conventional deep-learning-based defect inspection methods approached the problem as a binary classification, or only certain obvious defects can be correctly classified. In this study, we generated and released, for the first time, to the best of our knowledge, the benchmark dataset named SpecularDefect9 with various defects for specular surfaces, and the classification accuracy of some kinds of defects may be low with only one kind of input image. To classify all kinds of defects accurately, the proposed method applied the light intensity contrast map combined with the original captured fringe pattern as the input of the network, and a fusion network was introduced to extract features from multi-modal inputs. Experimental results based on the released benchmark dataset verified the effectiveness and robustness of the proposed multi-modal defect classification method.

1. Introduction

Specular reflection materials, such as mirrors with high manufactured accuracy, wafers, and specular vehicle surfaces, are widely applied in industries. Traditional defect inspection for specular surfaces was usually performed by human-vision [1,2], which was subjective and time-consuming.

Deflectometry [3–5] is a non-contact and full-field measurement technology for the shape reconstruction and defect inspection of specular surfaces. Combined with machine learning and deflectometry, defect inspection methods have been introduced for many kinds of specular surfaces, such as molded plastic parts [6], car bumpers [7], metal sheets [8], and silicon wafers [9]. Nevertheless, different hand-crafted features need to be defined for each specific task in all the above methods, and the generalization ability of these machine-learning-based methods is limited.

Defect detection methods with the deep neural network have been widely applied in diffuse surfaces, such as concrete bridge surfaces [10], fabric [11], and power line insulators [12]. Diffuse surfaces can be directly captured by cameras and utilized as the inputs of the network. However, the directly captured images of specular surfaces are unsuitable for the network because of the specular reflection property. To solve this challenge, the deflectometry system uses a liquid crystal display (LCD) as the active target to display sinusoidal fringe patterns. The cameras capture the deformed patterns via the reflection of the inspected specular surfaces, which contain abundant geometrical and textural information. The local curvature map and light intensity contrast map [13] can be obtained based on the captured images, which are meaningful in defect inspection for specular surfaces. Combining the light intensity contrast map with the local curvature map, Maestro-Watson et al. [14] proposed a simple network for defect detection in specular surfaces and then introduced a data segmentation method [15] for specular surface inspection with a fully convolutional network. However, all the two methods simply predicted the parts as “ok” or “nok”, which only achieved a binary classification. Zhou et al. [16] proposed an end-to-end attention-based fully convolutional neural network named “DeepInspection” for automotive parts. Nevertheless, the network can only classify the obvious geometrical defects like pits and scratches. Based on one

* Corresponding author at: School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.
E-mail address: jtxi@sjtu.edu.cn (J. Xi).

https://doi.org/10.1016/j.optlaseng.2023.107488
Received 12 August 2022; Received in revised form 20 December 2022; Accepted 14 January 2023

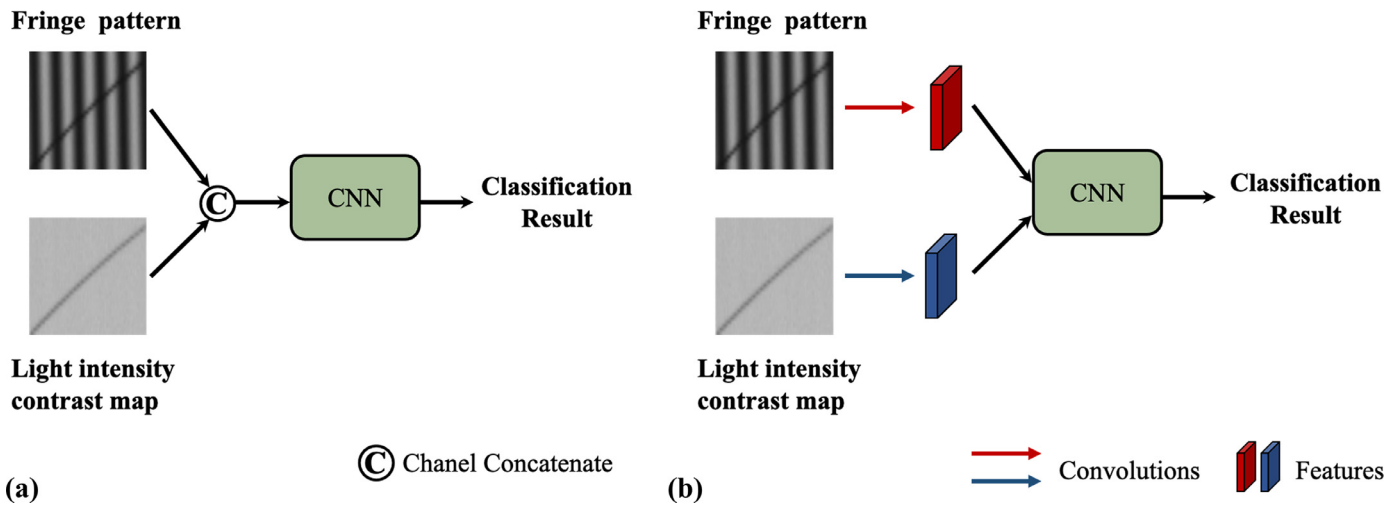

Fig. 1. Two main existing fusion strategies:(a) image-level fusion;(b) feature-level fusion.

light intensity contrast map, Guan et al. [17] introduced a network that can detect both geometrical defects like scratches, scuffings, and textural defects like stains and fingerprints but the defects can still be distinguished with only one input image. Qi et al. [18] proposed a phase-modulation combined network for accurate defect detection of specular surfaces. The network has a dual branch, and the inputs are the phase-shifting pattern sequence and the light intensity contrast map. However, the phase-shifting sequence is processed by three-dimensional convolution kernels, which are of high computational complexity, and the final prediction of the network can only classify defect and non-defect areas.

Inspired by the above methods, the light intensity contrast map can effectively describe the defect features on specular surfaces, but using only single modal information, some confusing defects (e.g., hair and scratch) cannot be clearly identified. The captured fringe patterns are necessary for obtaining auxiliary modal information. The ability to classify multiple defects can be further improved by coupling the features from the light intensity contrast map and the original fringe pattern. Intuitively, the proposed method tries to make the best of two different modal inputs, enforcing the complementary information learning between light intensity contrast maps and captured fringe patterns. As shown in Fig. 1, the main existing fusion strategies include image-level fusion and feature-level fusion. Nevertheless, the image-level fusion strategy is limited in learning the complementary information between sharply different modalities, which may even lead to a decrease in accuracy [19]. The feature-level fusion approach adaptively selects the useful features extracted from images of different modalities and further improves the accuracy due to the adequation of sufficient modality-aware features [20].

In this paper, the proposed method combines deflectometry and the deep neural network to classify multiple kinds of defects on specular surfaces. The main contribution of the study is as follows:

(1) To the best of our knowledge, the benchmark defect dataset for specular surfaces named SpecularDefect9[1] is generated and released based on the deflectometry for the first time.
(2) The multi-modal feature fusion network structure was proposed to further extract the image features from different modalities. The light intensity contrast map and the original fringe pattern are combined as the input of the network.

The benchmark dataset contains nine kinds of defects, including both geometrical defects and textural defects. Moreover, some defects in the dataset, like hairs, fibers, and scratches, may be difficult to classify with only one light intensity contrast map. Then, the multi-modal inputs are processed by the dual branch backbone. The features from the original fringe pattern are compressed and fused with the features from the light intensity contrast map by the fusion module. Furthermore, the convolutional block attention module [21] is used for adaptive feature refinement. Experimental results on the benchmark dataset verify that compared with conventional methods, the classification accuracy based on the proposed defect classification method is improved, and the multi-modal network is effective for the defect classification of specular surfaces.

2. Principle

The proposed defect classification method is shown in Fig. 2. The procedure of the proposed classification method can be divided into two parts. The deflectometry system is the first part, which consists of a camera and an LCD screen. The LCD screen is applied as the active target to display sinusoidal fringe patterns while the camera captures the virtual image of the fringe patterns via the reflection of the specular surface. The light intensity contrast map is calculated based on the captured fringe images. Then, the light intensity contrast map and the originally captured fringe image are combined as the input of the deep neural network. The second part is the proposed multi-modal fusion network, which accomplishes defect identification and classification work on specular surfaces. The proposed network has a dual-branch backbone to extract the features from the two modal input images, respectively. Finally, the features are fused by the fusion module and then processed by the prediction head to obtain the final defect classification results.

2.1. Principle of deflectometry system

The basic principle of deflectometry is the law of reflection. The phase-shifting patterns displayed on the LCD screen $I_n(x, y)$ can be expressed as:

$$I_n(x, y) = a(x, y) + b(x, y) \cdot \cos\left[\phi(x, y) + \frac{2\pi n}{N}\right], \; n = 1, 2, ..., N, \tag{1}$$

where $a(x, y)$,$b(x, y)$ and $\phi(x, y)$ are the background intensity, modulation intensity, and phase of fringes, respectively. N denotes the number of phase-shifting steps. Increasing the number of phase-shifting steps will effectively suppress the non-linear and random errors in the further calculation based on the fringe patterns. However, the measurement efficiency will decrease with the increasing phase-shifting steps. To balance quality and efficiency, we select $N = 8$ in the proposed method.

[1] The dataset is publicly available for download at https://tanasguan.github.io/SpecularDefect9/

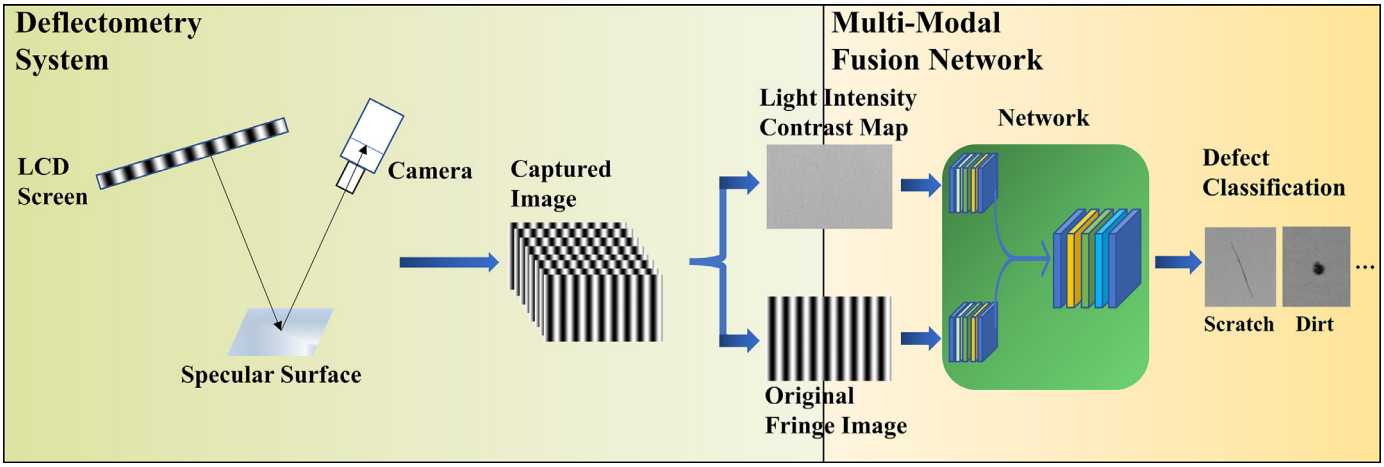

Fig. 2. Illustration of the defect classification method.

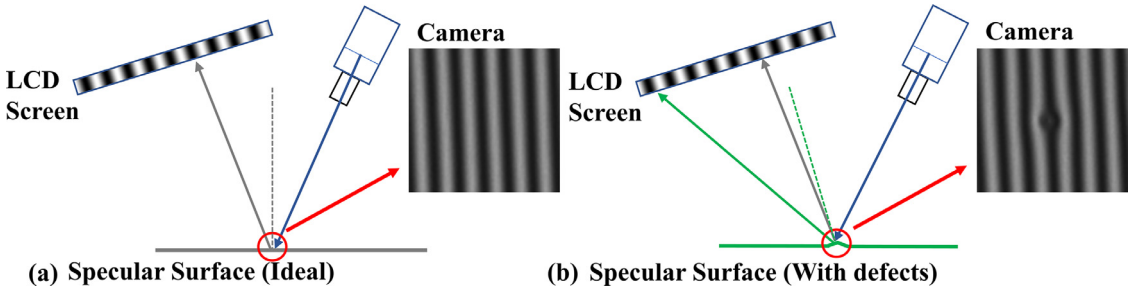

Fig. 3. Fringe images capture on (a) ideal specular surface and (b) specular surface with defects.

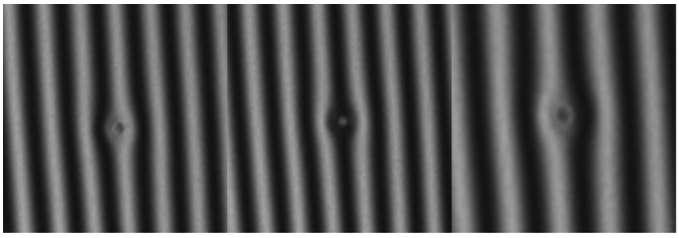

Fig. 4. Convex defects on different parts of fringe patterns.

The captured fringe patterns $I'_n(x,y)$are shown as follows:

$$I'_n(x,y) = I'_{bias}(x,y) + I'_{mod}(x,y) \cdot \cos\left[\phi(x,y) + \frac{2\pi n}{N}\right],\ n = 1,2,...,N, \quad (2)$$

where $I'_{bias}(x,y)$,$I'_{mod}(x,y)$ are the bias intensity and modulation intensity terms, respectively.

As shown in Fig. 3, the camera captures the fringe patterns via the reflection of the specular surface. When the surface has some defects like convex defects, as demonstrated in Fig. 3(b), the reflected fringe patterns will have a phase change, and the captured fringe patterns will have different features compared with the ideal specular surface, which can be applied to detect defects.

Nevertheless, as shown in Fig. 4, the different part of fringe patterns has different intensity, which will cause different image features even on the same defect.

The principle of the light intensity contrast map is the modulation intensity changing caused by defects. Both geometrical and textural defects will alter the amount of reflected light. The modulation intensity $I'_{mod}(x,y)$ is usually low when the defects appear while the $I'_{mod}(x,y)$ is stable on the ideal specular surface. Therefore, the light intensity contrast map can be used to detect both geometrical and textural defects

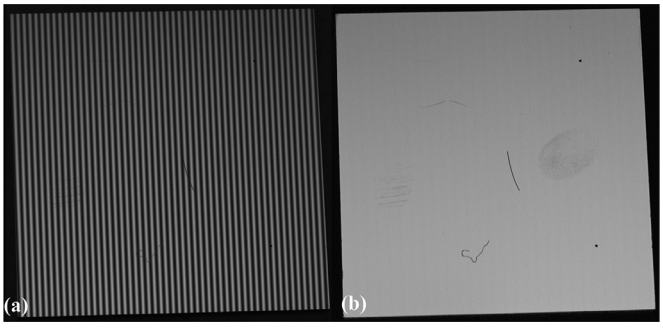

Fig. 5. Illustration of (a) original fringe pattern and (b) light intensity contrast map.

and can be calculated [17] as:

$$\gamma(x,y) = \frac{I'_{mod}(x,y)}{I'_{bias}(x,y)} = \frac{32\sqrt{\left(I'_7 - I'_3 + I'_8 - I'_4 + I'_6 - I'_2\right)^2 + \left(I'_1 - I'_5 + I'_2 - I'_4 + I'_8 - I'_6\right)^2}}{(1+\sqrt{2})\sum_{i=1}^{8} I'_i}, \quad (3)$$

where $\gamma(x,y)$ denotes the light intensity contrast map and I'_i is the i-th captured fringe image.

Fig. 5 illustrates the comparison between the original fringe pattern and the calculated light intensity contrast map. As shown in Fig. 5(b), the light intensity contrast map clearly shows the different kinds of defects (dirt, fiber, dirt, scratch, etc.) and removes the effect of the sinusoidal pattern background.

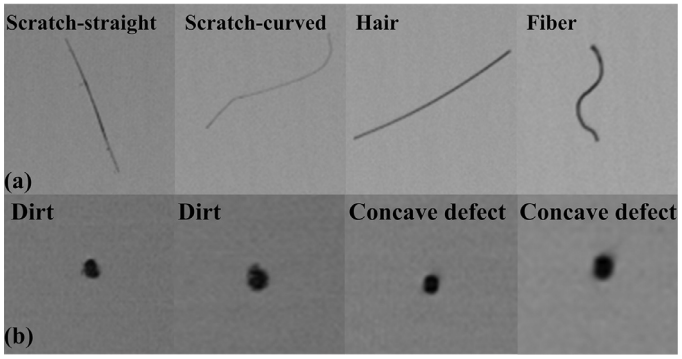

Fig. 6. Hard case among (a) scratch, hair, and fiber; (b) dirt and concave defect.

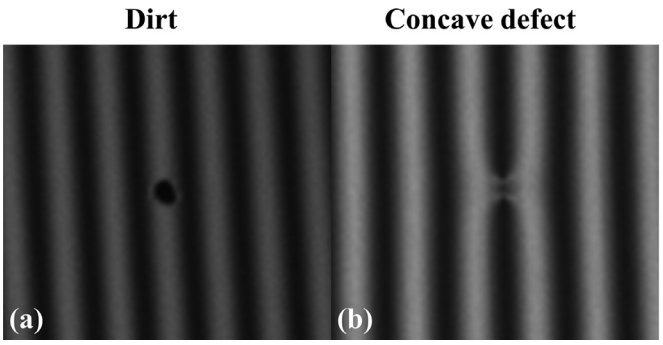

Fig. 8. Original fringe patterns of (a) dirt, and (b) concave defect.

Though both geometrical and textural defects can be shown on the light intensity contrast map, there are still some hard case defects, which may have low classification accuracy based on only one light intensity contrast map. As shown in Fig. 6(a), scratch, hair, and fiber defects are linear-like defects. The straight scratch and hair are hard to distinguish, while the curved scratch and fiber are difficult to classify accurately only based on the light intensity contrast map. The dirt and concave defect on the light contrast map are both like the black spot defect, as shown in Fig. 6(b), which also challenges the classification method.

As demonstrated in Eq. (3), the light intensity contrast map only preserve the light intensity information, and the phase information is lost. As shown in Fig. 7, the hair and fiber are diffuse reflection defects, in which area the camera cannot capture the reflected fringe patterns. In the area of scratch, the defect area can partly reflect the pattern because the specular reflection is not entirely destroyed by the scratch.

Similarly, the dirt is a textural defect, which will have no influence on phase change, while the concave defect is geometrical, and the camera will capture the deformed fringe images. Fig. 8 demonstrates that the dirt area has no reflection of the fringe patterns, while the captured fringe image around the concave defect has a phase change.

Based on the above analysis, we claim that the combination of the light intensity contrast map and the captured fringe pattern provides more information for the network. However, the additional information may be processed as noise and decrease the classification accuracy instead. Hence, the designed multi-modal fusion network is of vital importance for better feature extraction and fusion.

2.2. Principle of proposed fusion network

A novel multi-modal fusion deep architecture is derived, which is consisted of the encoder module, multi-modal fusion module, feature pyramid network (FPN) [22], and decision module, as shown in Fig. 9.

The proposed network employs dual-branch ResNet-18 [23] as the encoder and obtains features of low level, medium level, and high level, which have the feature size of $128 \times 16 \times 16$,$256 \times 8 \times 8$, and $512 \times 4 \times 4$, respectively. The features from different modalities are then processed by the fusion module to combine the multi-modal information.

As shown in Fig. 10, referring to the dynamic fusion module in Ref. [24], the proposed feature fusion module mainly includes three sub-modules: channel reduction, channel fusion, and adaptive feature selection. The channels of the features from the fringe pattern are firstly reduced by the 1×1 convolution kernel and concatenated with the features from the light intensity contrast map. The concatenated features are then filtered by the 3×3 convolution kernel and passed to the adaptive feature selection module. The features are processed by the global adaptive pooling (GAP), 1×1 convolution kernel, and Sigmoid activation function, respectively. Finally, the original fused features are then multiplied with the processed features.

As shown in Fig. 11, the fused features from the fusion module are fed into FPN for multi-scale feature fusion. The channels of the multi-level features are firstly reduced by 1×1 convolution kernel. The higher-level features are up-sampled and added with the lower-level features. The max pooling layer is applied to down-sampling the fused features, and then the multi-level features are concatenated to obtain the decision features which have the size of $96 \times 4 \times 4$. The decision features are then processed by the GAP and the fully connected (FC) layer to obtain the final 10×1 decision result, representing the predicted probability of the nine defects and the non-defect area, respectively.

2.3. Principle of benchmark dataset generation

To evaluate the proposed defect classification method, this study generated a benchmark defect dataset of the specular surfaces for the first time. The deflectometry system is established to acquire the dataset and demonstrated in Fig. 12. An LCD screen (AOC 24P1U, resolution of 1920×1080 pixels, pixel pitch of 0.273 mm) is employed as the active target. The two cameras are CCD sensors (BASLER acA2440-20gm with a resolution of 2448×2048 pixels and pixel pitch of 3.45 μm), and the camera lenses (MORITEX ML-MC16HR) have a focal length of 16 mm. To be mentioned, for defect classification in two dimensions (2D), only the captured images from camera1 are used in the dataset, but this study

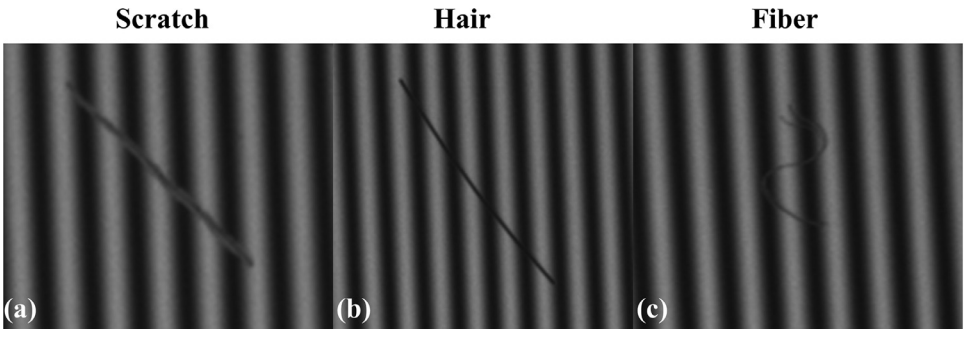

Fig. 7. Original fringe patterns of (a) scratch, (b) hair, and (c) fiber.

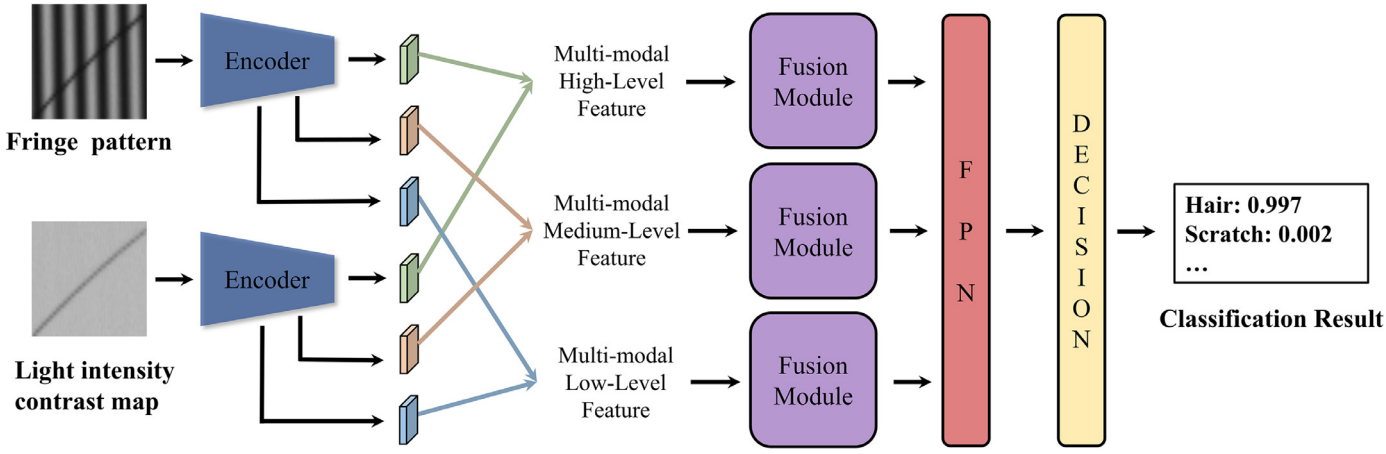

Fig. 9. Architecture of the proposed multi-modal fusion network.

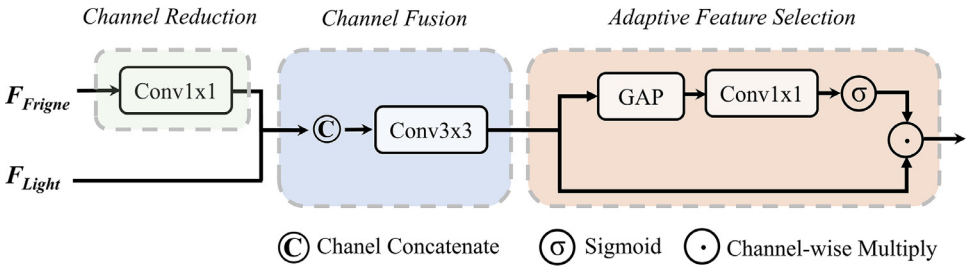

Fig. 10. Details of the fusion module.

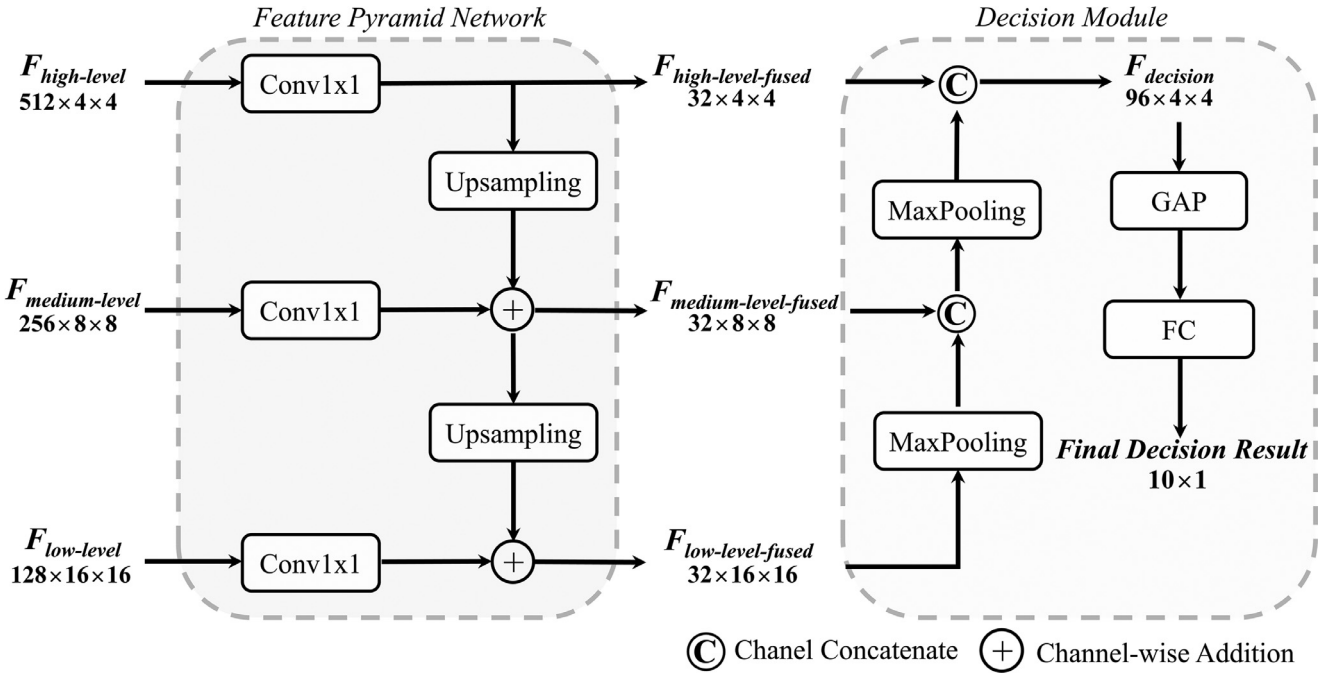

Fig. 11. Details of the FPN and decision module.

collected the fringe images from both two cameras for further study of the three-dimensional (3D) defect classification for specular surfaces in future work.

After the stereo camera calibration [25] and system geometrical parameters calibration [26], the details of the system parameters are shown in Table 1.

For collecting the surface information, the LCD screen displays the fringe patterns with the fringe numbers 225, 224, and 210 in vertical and horizontal directions, respectively. The number of phase-shifting steps for each fringe frequency is set to 8. Therefore, for each camera, 48 fringe images are captured for the tested specular surface in total. Each 8-step phase-shifting fringe sequence in the same frequency can

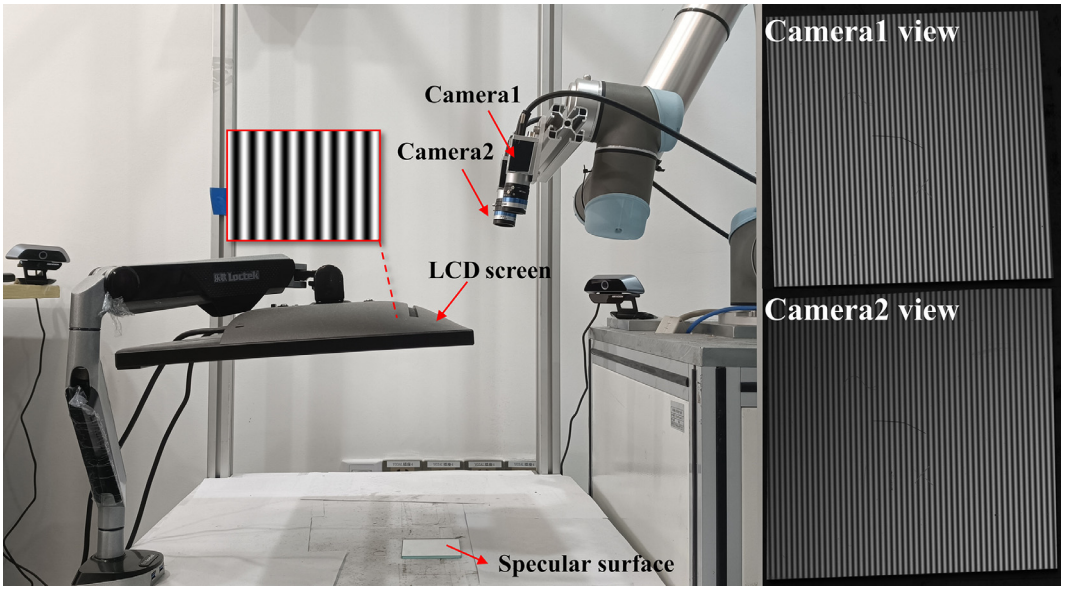

Fig. 12. Illustration of the deflectometry system.

Table 1
Details of the deflectometry system parameters.

System parameters	Value
Mean reprojection error	camera1:0.061*pixel*; camera2: 0.059 *pixel*
Baseline distance	116*mm*
Geometrical transformation (LCD to camera1)	$\begin{bmatrix} 0.9950 & -0.0019 & -0.1000 & -290.0302 \\ 0.0039 & 0.9999 & 0.0116 & -34.6549 \\ 0.1000 & -0.0119 & 0.9949 & 190.3450 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
Work distance (Camera1 to work plane)	348.6*mm*
Lateral resolution	75.2*um*
Gradient resolution	$7.1 \times 10^{-5} rad$
Height resolution	2.5*nm*

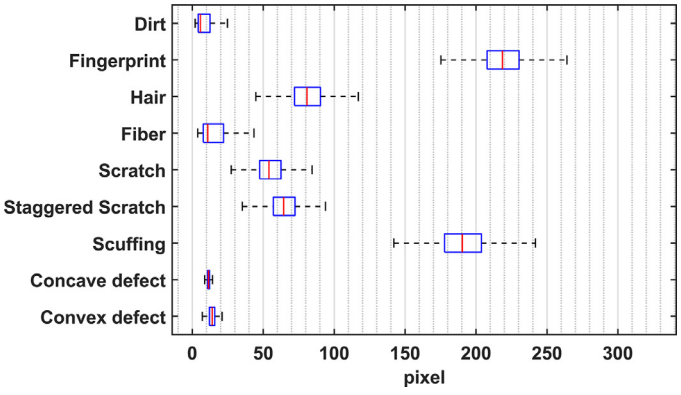

Fig. 14. Statistics of the diagonal length of the bounding box in the dataset.

obtain one light contrast intensity map, and the light contrast intensity map in the dataset is calculated based on the eight fringe images with fringe numbers of 225 in the vertical direction.

For acquiring the defect dataset, 554 float glass with an aluminized front surface and 553 polycarbonate mirrors, which have plastic deformation properties and are suitable for making concave and convex defects. Nine kinds of defects are hand-crafted on specular surfaces. The captured fringe image and the calculated light intensity contrast map are shown in Fig. 13.

The combination of the light intensity contrast map and the fringe image was used to improve the accuracy of defects annotation. The quantity statistics of the dataset are shown in Table 2.

Fig. 14 shows the statistics of the diagonal length of the annotated defect bounding box.

To evaluate the annotation quality of the dataset, the dataset is randomly sampled, and then two annotation experts annotate the sampled dataset independently. The union of the two expert annotation results

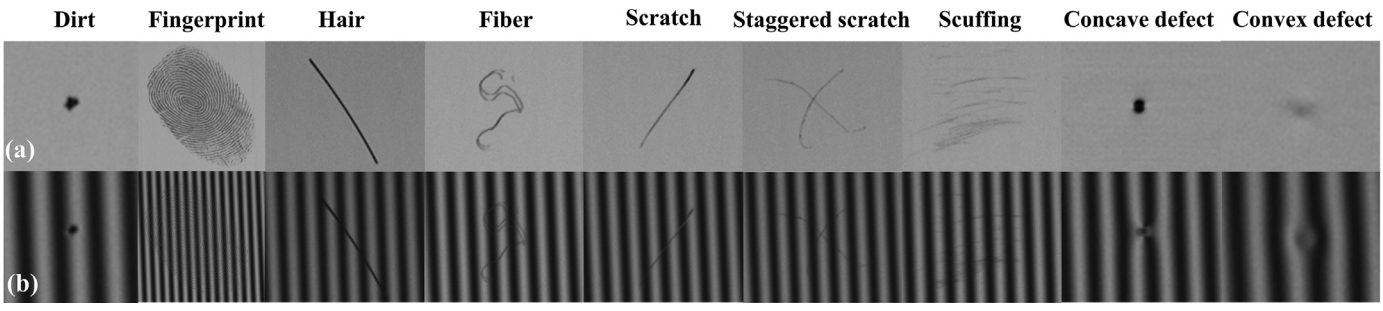

Fig. 13. (a) Light intensity contrast map and (b) fringe image of defects in the dataset.

Table 2
Quantity statistics of the defect dataset.

	Dirt	Fingerprint	Hair	Fiber	Scratch	Staggered scratch	Scuffing	Concave defect	Convex defect
Training	1544	457	460	2266	590	277	448	923	914
Testing	629	202	204	891	250	125	197	401	392

Table 3
Annotation quality statistics of the defect dataset.

	Dirt	Fingerprint	Hair	Fiber	Scratch	Staggered scratch	Scuffing	Concave defect	Convex defect
Recall	97.7%	100%	100%	99.3%	99.3%	100%	100%	99.0%	98.9%
FPR	4.6%	0%	0%	0.7%	0.7%	0%	0%	0.9%	0.9%

Table 4
Hyper-parameter sensitivity study for learning rate, batch size, and weight decay.

Learning rate	Accuracy	Batch size	Accuracy	Weight decay	Accuracy
0.01	96.9 ± 0.1	16	96.5 ± 0.1	0.0001	96.9 ± 0.1
0.02	96.9 ± 0.2	32	96.9 ± 0.3	0.0002	96.9 ± 0.2
0.05	**97.0 ± 0.1**	**64**	**97.0 ± 0.1**	**0.0005**	**97.0 ± 0.1**
0.10	96.8 ± 0.1	128	96.9 ± 0.1	0.0010	96.9 ± 0.3
0.20	96.6 ± 0.1	256	96.6 ± 0.2	0.0020	96.8 ± 0.1

is used as the ground truth data to calculate the recall, while the intersection is utilized to obtain the false positive rate (FPR). The annotation quality statistics are summarized in Table 3.

3. Experiments

For evaluating the performance of the proposed method, the proposed multi-modal feature fusion network is compared with the input fusion network and the single-modal network, which are based on the captured fringe pattern and the light intensity contrast map, respectively. The network is trained and tested based on the benchmark dataset. For the network training, we use stochastic gradient descent (SGD) optimizer with the input size of 128 × 128. The sensitivity study of hyper-parameters has been done to train the network effectively and improve the performance of the network. The detailed studies are summarized in Table 4, and finally, the network is trained with an initial learning rate of 0.05, weight decay of 0.0005, and batch size of 64.

The corresponding warm start epochs are 5 with initial warmup learning rate of 0.001, and the total epochs are set to 31. We use poly scheduling to decay the learning rate during the training process, as shown in Eq. (4).

$$lr = lr_{base} \cdot \left(1 - \frac{iter}{total\_iter}\right)^{0.9}, \tag{4}$$

where lr_{base} is the initial learning rate, $iter$ is the current iteration, and the total iteration can be calculated as 3816.

Given a labeled classification dataset $\mathcal{D} = \{(\mathbf{x}_i^{fringe}, \mathbf{x}_i^{light}, y_i)\}_{i=1}^{N}$, where y_i represents the hand-annotated label for the i-th image $(\mathbf{x}_i^{fringe}, \mathbf{x}_i^{light})$. $\mathbf{x}_i^{fringe}$, and $\mathbf{x}_i^{light}$ denote the captured fringe pattern and the light intensity contrast map, respectively. The optimization target is to train the defect classification model, and despite different methods, the standard cross-entropy (CE) loss is adopted as the loss function, which can be written as $\ell_{ce}(p_i, y_i)$, where p_i stands for the softmax probabilities generated by the model for the i-th input. For the feature fusion strategy, the whole loss function can be formulated as:

$$\mathcal{L} = \frac{1}{N} \sum_{\left(\mathbf{x}_i^{fringe}, \mathbf{x}_i^{light}, y_i\right) \in \mathcal{D}} \ell_{ce}\left(g \circ f(h_1(\mathbf{x}_i^{fringe}), h_2(\mathbf{x}_i^{light}), y_i\right), \tag{5}$$

where the different inputs $(\mathbf{x}_i^{fringe}, \mathbf{x}_i^{light})$ are firstly fed into separate encoders (h_1, h_2), and then f for multi-modal feature fusion. $g \circ f$ denotes the composition function of the decision module g and the feature fusion module f.

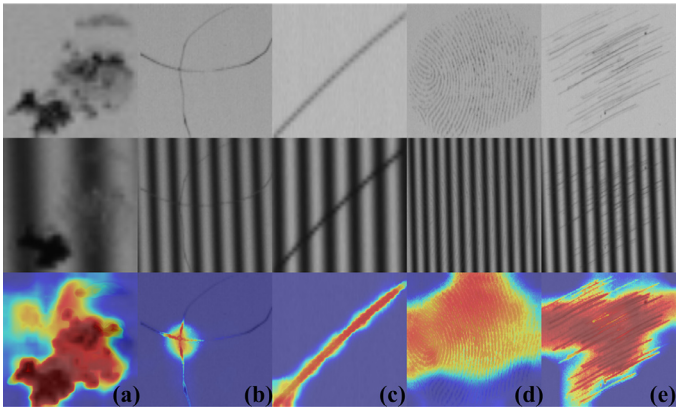

Fig. 15. Illustration of CAM for (a) dirt, (b) staggered scratch, (c) hair, (d) fingerprint, and (e) scuffing.

Several augmentation methods, such as horizontal flip, rotation, and color jitter, are applied to prevent overfitting in the training stage. All experiments are implemented based on the Python language and the framework of Pytorch (Facebook). The GPU of the operating system is Nvidia GeForce GTX 1080Ti, which has an 11-GB dedicated display memory size.

3.1. Qualitative and visualization analysis

After network training, the network activation of different defects is firstly evaluated. As shown in Fig. 15, the class activation mapping [27] (CAM) of five kinds of defect instances is illustrated. From top to bottom are the light intensity contrast map, the captured fringe pattern, and the CAM produced by the proposed model, respectively. The CAM shows that the defect areas will have a more obvious activation than the non-defect area, which demonstrates that the proposed network has a great ability in feature extraction. It is worth noting that, for the staggered scratch defects, the proposed network focuses and has a significant activation on the staggered area, as shown in Fig. 15(b), which further verifies that the proposed model can focus on the special features of each kind of defect.

To further evaluate the defect classification ability of the proposed fusion network, the visualization of the feature spaces is obtained based on the t-SNE [28]. For each kind of defect, 150 samples are randomly se-

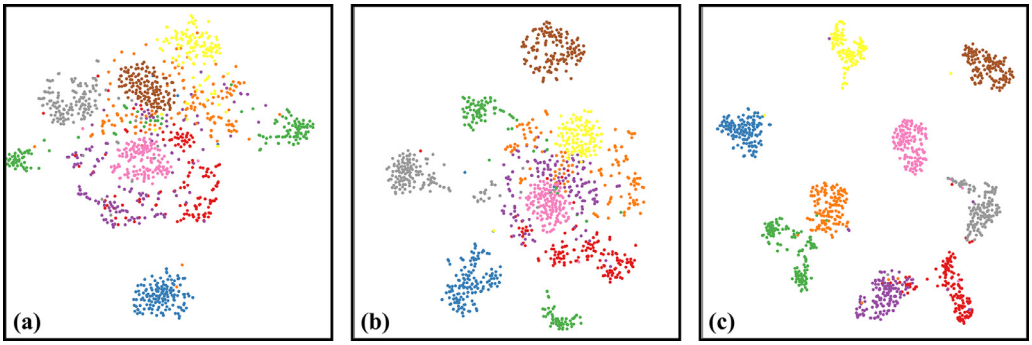

Fig. 16. Illustration of t-SNE for the single-modal network based on (a) only the fringe pattern, (b) only the light intensity contrast map, and (c) the multi-modal fusion network.

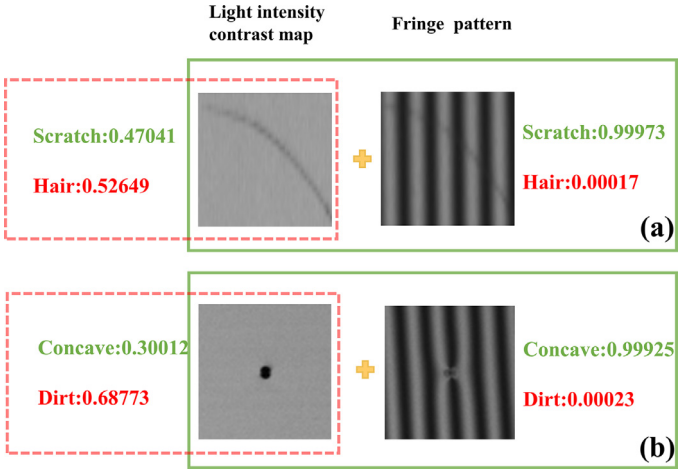

Fig. 17. Hard case illustration of (a) scratch, and (b) concave defect.

lected for visualization. The t-SNE of the single-modal network and the multi-modal network is shown in Fig. 16. As demonstrated in Fig. 16(a) that the model, only uses the fringe pattern, generating confusing decision boundaries of features, which will lead to wrong classification results. The single-modal network based on the light intensity contrast map improves the distinguishability of features to a certain extent, as shown in Fig. 16(b). Nevertheless, some features from different defects are still mixed. While the multi-modal network, which combines two different modalities of information, has much more clear ones in that the features of the same defect are minified, and the distance among different defects is enlarged. Fig. 16 demonstrates the effectiveness of the proposed feature fusion network from the feature point of view.

Fig. 17 shows two instances of the hard case in the dataset. When using only the light intensity contrast map, the network gets comparable predicted probabilities for different kinds of defects, which draws wrong classification results. Compared with the hair, the scratch area can still partly reflect the fringe pattern, and concave defects will have a phase change around the defect area when compared with the textural defect dirt. Hence, the fringe pattern provides additional information for defect classification. As shown in Fig. 17, based on the proposed multi-modal fusion network, the classification probabilities of the instance of scratch and concave defects are 99.97% and 99.92%, respectively. The proposed multi-modal fusion network shows a better ability to handle hard cases than the single-modal network.

3.2. Quantitative analysis

For quantitative performance analysis based on the proposed network, accuracy is used to evaluate different methods, and F1-Score is employed to measure the performance in the specific category. A true positive (TP) represents the defects that are correctly classified. A false positive (FP) means non-defective areas are mistakenly classified as defects, while a false negative (FN) is defective areas are mistakenly detected as non-defective areas, and a true negative (TN) denotes that the non-defective areas are correctly detected. Moreover, the precision rate is generally used to evaluate the global accuracy of the model, while the recall rate is the fraction of the correctly recognized true positives over the total number of actual positives. The precision and recall are described as Eqs. (6) and (7).

$$precision = \frac{TP}{TP + FP}. \tag{6}$$

$$recall = \frac{TP}{TP + FN}. \tag{7}$$

F1-Score is the harmonic average of precision and recall, and can be expressed as:

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall}. \tag{8}$$

Accuracy is the ratio of the number of correctly classified samples to the total number of samples, and can be described as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{9}$$

According to the discriminability of different classes, the defects are divided into two groups (difficult ones and simple ones), and then the two categories are evaluated separately. The simple categories contain Fingerprint, Staggered Scratch, and Scuffing, which can be accurately classified by human-vision based on only the light intensity contrast map. The difficult categories contain Dirt, Hair, Fiber, Scratch, Concave, and Convex, which need additional information to improve the classification accuracy.

In particular, all results are averaged over ten random seeds to draw reliable conclusions. Table 5 shows the accuracy comparison with different methods on the validation set. The mean and standard deviation are calculated over ten random seeds. As demonstrated in Table 5, the light intensity contrast map achieves a significantly better performance than the captured fringe pattern, which improves the accuracy by 2.1%, 1.9%, and 3.0% in three categories, respectively. The results prove the superiority of the light intensity contrast map. Due to the additional information, the input fusion method outperforms the single-modal network in difficult categories. However, compared with the network based on the light intensity contrast map, the input fusion network performs worse in simple categories, and the accuracy decreases by 0.4%, which exposes the shortness in information fusion of the input fusion network. It is worth noting that, the feature fusion method obtains the best results under all settings, which indicates the effectiveness of multi-modal information fusion.

Table 6 shows the comparison of the F1-Score for all kinds of defects, and the bold defects indicate the difficult categories. As demonstrated

Table 5
Accuracy comparison with different methods.

Methods	Accuracy		
	All categories	Difficult categories	Simple categories
Captured fringe pattern only	94.6 ± 0.2	94.2 ± 0.2	96.4 ± 0.5
Light intensity contrast map only	96.7 ± 0.1	96.1 ± 0.2	99.4 ± 0.2
Input Fusion	96.6 ± 0.1	96.2 ± 0.2	99.0 ± 0.2
Feature Fusion	**97.0 ± 0.1**	**96.5 ± 0.1**	**99.5 ± 0.1**

Table 6
F1-Score comparison with different methods.

Defect	F1-Score			
	Captured fringe pattern only.	Light intensity contrast map only	Input Fusion	Feature Fusion
Dirt	90.5 ± 0.4	93.6 ± 0.5	93.5 ± 0.3	**94.1 ± 0.4**
Fingerprint	99.3 ± 0.2	99.4 ± 0.2	99.2 ± 0.2	**99.6 ± 0.1**
Hair	99.1 ± 0.2	98.3 ± 0.4	99.4 ± 0.2	**99.6 ± 0.2**
Fiber	92.3 ± 0.2	95.0 ± 0.3	95.1 ± 0.3	**95.3 ± 0.2**
Scratch	91.0 ± 0.8	96.2 ± 0.5	95.8 ± 0.5	**96.4 ± 0.3**
Staggered Scratch	91.0 ± 1.2	**99.1 ± 0.3**	98.7 ± 0.5	98.9 ± 0.2
Scuffing	99.4 ± 0.2	99.8 ± 0.1	99.5 ± 0.1	**99.9 ± 0.1**
Concave	98.7 ± 0.2	98.9 ± 0.1	98.8 ± 0.2	**99.0 ± 0.1**
Convex	98.0 ± 0.5	98.6 ± 0.2	98.7 ± 0.2	**99.0 ± 0.2**

Table 7
Ablation study on the channel reduction hyper-parameter$r$.

r	Accuracy		
	All categories	Difficult categories	Simple categories
1	96.9 ± 0.1	96.4 ± 0.1	99.3 ± 0.2
2	96.9 ± 0.1	96.4 ± 0.1	99.4 ± 0.1
4	96.9 ± 0.3	96.4 ± 0.3	99.4 ± 0.2
8	**97.0 ± 0.1**	**96.5 ± 0.1**	**99.5 ± 0.1**
16	96.8 ± 0.2	96.4 ± 0.2	99.3 ± 0.2

Table 8
Ablation study on different components in feature fusion module.

Channel reduction module	Adaptive feature selection module	Accuracy		
		All categories	Difficult categories	Simple categories
		96.8 ± 0.2	96.3 ± 0.2	99.3 ± 0.2
	√	96.9 ± 0.1	96.4 ± 0.1	99.3 ± 0.2
√		96.9 ± 0.1	96.4 ± 0.1	99.3 ± 0.3
√	√	**97.0 ± 0.1**	**96.5 ± 0.1**	**99.5 ± 0.1**

in Table 6, the feature fusion method achieves the best F1-Scores in all difficult categories and almost all simple categories, further proving the robustness and superiority of the proposed multi-modal feature fusion network.

Table 7 is the ablation study result for verifying the effectiveness of channel reduction. Hyper-parameter r represents the rate of channel reduction to F_{Fringe}, and $r = 1$ means the feature will not be reduced. As shown in Table 7, proper use of channel reduction can help achieve better results. In the experiments, the hyper-parameter r is set with eight as the default setting unless specifically stated.

Table 8 shows the experiments to ablate each component of the feature fusion module step by step. As demonstrated in Table 8, compared to the baseline, the channel reduction module and channel selection module can bring additional accuracy gain, verifying the effectiveness of the proposed feature fusion module.

Following this result, these components are applied in all experiments in Tables 5 and 6. Finally, when adding all components together, the proposed multi-modal feature fusion network achieves results under all protocols with the accuracy of 97.0%, 96.5%, and 99.5% in all categories, difficult categories, and simple categories, respectively.

4. Limitations and future work

In this study, we proposed a multi-modal network to improve the feature extraction based on the combination of the light intensity contrast map and the captured fringe pattern. The experimental results demonstrate that the defect classification accuracy is improved by the proposed method to a certain extent but not significantly improved. The purpose of the study is not to propose a state-of-the-art multi-modal network for defect classification but to provide a new approach for the defect classification of specular surfaces. Hence, there are still two aspects that can be further researched. Firstly, the proposed method only applies one captured fringe pattern as the information supplement, which still has information lost. Extracting and utilizing the information from the temporal fringe pattern sequence needs to be further studied. Then, the benchmark dataset collects the fringe images from two cameras, and the 2D defect classification only needs the information from one camera. In the future, the 3D defect classification for specular surfaces combining the information from the two cameras needs to be researched.

5. Conclusion

Defect classification for specular surfaces is of vital importance in the manufacturing industry. Conventional machine-learning-based defect classification methods lack generalization ability, while deep-learning-based methods are almost binary classifications or classifications of obvious defects. In this study, the deflectometry system is established for generating the benchmark defect dataset for specular surfaces, and the dataset is released for the first time in the deflectometry area. The dataset contains nine kinds of geometrical and textural defects. Furthermore, some of the defects may have low classification accuracy with only one light intensity contrast map. To improve the classification accuracy, the coupling images, which consist of the light intensity contrast map and the captured fringe pattern, are utilized as the input of the network to provide abundant information for defect classification. The multi-modal fusion network is constructed for better feature extraction and fusion, which consists of the dual branch backbone, feature fusion module, convolutional block attention module, and the classification head. After trials with ten random seeds on the benchmark dataset, compared with the single-modal network and the input-fusion network, the proposed multi-modal feature fusion network performs the best, and the final classification accuracy is $(97.0 \pm 0.1)\%$ in all categories, which

demonstrates the effectiveness and robustness of the proposed multi-modal defect classification approach.

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

CRediT authorship contribution statement

Jingtian Guan: Conceptualization, Methodology, Formal analysis, Data curation, Investigation, Software, Validation, Writing – original draft, Writing – review & editing, Funding acquisition. **Jingjing Fei:** Methodology, Formal analysis, Investigation, Software, Validation, Visualization, Writing – review & editing. **Wei Li:** Formal analysis, Investigation, Software, Validation, Supervision, Writing – review & editing. **Xiaoke Jiang:** Conceptualization, Supervision, Writing – review & editing. **Liwei Wu:** Supervision, Project administration, Writing – review & editing. **Yakun Liu:** Data curation, Writing – review & editing. **Juntong Xi:** Conceptualization, Supervision, Project administration, Funding acquisition, Writing – review & editing.

Data availability

I have share the link of the dataset at the manuscript. The dataset is publicly available for download at https://tanasguan.github.io/SpecularDefect9/.

Funding

This study was funded by the National Natural Science Foundation of China (52175478, 52205533), Science and Technology Commission of Shanghai Municipality (21511104202), Ministry of Industry and Information Technology of the People's Republic of China (CJ04N20), and Ministry of Education–China Mobile Research Foundation (CMHQ-JS-201900003).

References

[1] Forte P, et al. Exploring combined dark and bright field illumination to improve the detection of defects on specular surfaces. Opt Lasers Eng 2017;88:120–8.
[2] Tao X, et al. A novel and effective surface flaw inspection instrument for large-aperture optical elements. IEEE Trans Instrum Meas 2015;64(9):2530–40.
[3] Huang L, et al. Review of phase measuring deflectometry. Opt Lasers Eng 2018;107:247–57.
[4] Zhang Z, et al. Recent advance on phase measuring deflectometry for obtaining 3D shape of specular surface. Proc SPIE 2020;11552:185–94.
[5] Zhang Z, et al. Phase measuring deflectometry for obtaining 3D shape of specular surface: a review of the state-of-the-art. Opt Eng 2021;60(2):020903.
[6] Tarry C, Stachowsky M, Moussa M. Robust detection of paint defects in molded plastic parts. In: Proceedings of the Canadian conference on computer and robot vision; 2014. p. 306–12.
[7] Tandiya A, et al. Automotive semi-specular surface defect detection system. In: Proceedings of the 15th conference on computer and robot vision; 2018. p. 285–91.
[8] Ziebarth M. Empirical comparison of defect classifiers on specular surfaces. In: Proceedings of the 2013 joint workshop of fraunhofer IOSB and institute for anthropomatics; 2014. p. 155–69.
[9] Kofler C, Spöck G, Muhr R. Classifying defects in topography images of silicon wafers. In: Proceedings of the winter simulation conference; 2017. p. 3646–57.
[10] Zhang C, Chang C, jamshidi M. Concrete bridge surface damage detection using a single-stage detector. Comput Aided Civ Infrastruct Eng 2019;35(4):389–409.
[11] Wei B, et al. A new method using the convolutional neural network with compressive sensing for fabric defect classification based on small sample sizes. Text Res J 2019;89(17):3539–55.
[12] Tao X, et al. Detection of power line insulator defects using aerial images analyzed with convolutional neural networks. IEEE Trans Syst Man Cybern Syst 2002;50(4):1486–98.
[13] H. Schreiber, and J.H. Bruning, "Phase Shifting Interferometry," in Optical Shop Testing, pp. 547–666, (2007).
[14] Maestro-Watson D, et al. Deep Learning for Deflectometric Inspection of Specular Surfaces. In: Proceedings of the 13th international conference on soft computing models in industrial and environmental applications. Springer; 2018. p. 280–9.
[15] Maestro-Watson D, et al. Deflectometric data segmentation for surface inspection: a fully convolutional neural network approach. J Electron Imaging 2020;29:041007.
[16] Zhou Q, et al. Deepinspection: deep learning based hierarchical network for specular surface inspection. Measurement 2020;160:107834.
[17] Guan J, et al. Defect detection method for specular surfaces based on deflectometry and deep learning. Opt Eng 2022;61(6):061407.
[18] Qi Z, et al. Phase-modulation combined deflectometry for small defect detection. Appl Opt 2020;59(7):2016–23.
[19] Zhou M, et al. Mutual information-driven pan-sharpening. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2022. p. 1798–808.
[20] Liang Y, et al. Multimodal material segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2022. p. 19800–8.
[21] S. Woo et al., "CBAM: convolutional block attention module," 10.48550/arXiv.1807.06521.
[22] Lin TY, et al. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2017. p. 936–44.
[23] He K, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 770–8.
[24] T. Liang et al., "BEVFusion: a simple and robust LiDAR-camera fusion framework," 10.48550/arXiv.2205.13790.
[25] Zhang Z. A flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell 2000;22(11):1330–4.
[26] Guan J, et al. An improved geometrical calibration method for stereo deflectometry by using speckle pattern. Opt Commun 2022;505(15):127507.
[27] Zhou B, et al. Learning Deep Features for Discriminative Localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 2921–9.
[28] Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.