A Novel Multispectral Fusion Defect Detection Framework With Coarse-to-Fine Multispectral Registration

Jiacheng Li, Bin Gao, *Senior Member, IEEE*, Wai Lok Woo, *Senior Member, IEEE*, Jieyi Xu, Lei Liu, and Yu Zeng

***Abstract*— This article introduces a new imaging approach to nondestructive defect detection by combining visual testing (VT) and infrared thermal testing (IRT) in a multispectral vision sensing fusion system. The goal is to overcome the hampering challenges faced by traditional imaging methods, including complex environments, irregular samples, various defect types, and the need for efficient detection. The proposed system simultaneously detects and classifies surface and subsurface defects, addressing issues, such as false detection due to changes in surface emissivity in IRT and the inability to detect subsurface defects in VT. A novel multispectral fusion defect detection framework is proposed, employing coarse-to-fine multispectral registration for accurate alignment of infrared and visible images with different resolutions and fields of view. Domain adaptation unifies the feature domains of infrared and visible images by replacing the phase components in the frequency domain. The framework utilizes the complementary information from infrared and visible modalities to enhance detection accuracy and robustness. Experimental validation is conducted on different specimens, confirming the effectiveness of the proposed framework in detecting and generalizing to various shapes and materials. Overall, this article presents a novel imaging system that combines VT and IRT, offering improved detection capabilities in complex environments and diverse defect scenarios. The demo code is available at: https://github.com/ljcuestc/YoloMultispectralFusion-Coarse-to-fine-Registration.gi.**

***Index Terms*— Coarse-to-fine image registration, defect detection, late fusion, multimodal, multispectral fusion.**

I. Introduction

With the rapid development of the global manufacturing industry, nondestructive testing (NDT) takes an increasingly essential role in electronic and communication equipment, instrumentation, transportation, pipeline transportation, and aerospace as an efficient, nondestructive, noncontact defects detection technology [1]. It is used to guarantee product reliability and stability in the production stage and monitor the changes of its structure and status in the operation as well as maintenance stage.

Vision-based NDT (VNDT) receives significant attention due to its visualization capability, accuracy, and convenience. VNDT includes terahertz testing (TT), X-ray testing, infrared thermal testing (IRT), and visual testing (VT), of which IRT and VT are the most representative technologies. IRT can detect subsurface defects, such as debonding, bulging, voids, and so on. Liu et al. [2] proposed a convolutional graph thermography (CGT) method for subsurface defect detection in polymer composites, which effectively reduces noise and inhomogeneous backgrounds in infrared images. Zhang et al. [3] used vibrothermography to detect impact damage in basalt fiber reinforced polymer (BFRP), carbon fiber reinforced plastics (CFRP), and linked the estimation of depth information to loadings in partial least-squares thermography. Puthiyaveettil et al. [4] designed a laser thermography system and investigated the influence of material surface absorptivity on crack detection. Ichi and Dorafshan [5] used semantically segmented IRT images to evaluate the effectiveness of IRT in the detection of subsurface deck delamination. Unlike IRT, VT can only detect surface defects. However, it can obtain high-resolution texture information to enhance the detection sensitivity and accuracy. Cheng and Yu [6] introduced a deep neural network DEA_RetinaNet for steel surface defect detection, which embeds a novel channel attention mechanism to reduce visual information loss. Li et al. [7] proposed an automatic tear measuring system for drilling-induced delamination defects in CFRP composite laminate, achieving accurate tear measurements using a double-light imaging framework under different light intensities. Ren et al. [8] discussed VT in the field of industrial defect detection systematically and applied deep learning in defect classification, localization, and segmentation. Schlosser et al. [9] proposed a novel hybrid multistage system of stacked deep neural networks (SH-DNNs), which can detect the finest structures within only a few micrometers in pixel size. However, significant technical challenges still remain with the above two spectral-based detection methods, whereas IRT results in ambiguity due to the varying emissivity, and VT cannot detect subsurface defects.

Multimodal sensing fusion can make full use of the complementary and redundant information between different modalities to improve the generalization and accuracy of the system or model. It can overcome limitations from single modality sensing. In recent years, multimodal fusion has a large number of successful applications in different

Manuscript received 22 May 2023; revised 25 November 2023; accepted 4 December 2023. Date of publication 19 December 2023; date of current version 2 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61971093, Grant 61960206010, and Grant 61527803; and in part by the Science and Technology Research Project of Pipechina under Grant WZXGL202104. The Associate Editor coordinating the review process was Dr. Boby George. *(Corresponding author: Bin Gao.)*

Jiacheng Li, Bin Gao, Jieyi Xu, Lei Liu, and Yu Zeng are with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: bin_gao@uestc.edu.cn).

Wai Lok Woo is with the Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK.

Digital Object Identifier 10.1109/TIM.2023.3344145

fields, such as autonomous intellisense and smarter healthcare. Zhang et al. [10] introduced a multimodal sensor fusion network (Robust-FusionNet) that effectively addresses the distortions caused by severe weather conditions in LiDAR point cloud data and camera images. Arnold et al. [11] built a 3-D object detection system suitable for autonomous vehicle, which realized the fusion and collaboration of multiple infrastructure sensors and can recall more than 95% of the objects in the most challenging scenario. Islam et al. and Chang et al. [12] proposed a multimodal sensor system for wound assessment and pressure ulcer care that integrated five sensing modalities, including electro-optic (EO), depth, thermal, multispectral imaging, and chemical sensing. Andreozzi et al. [13] presented a multimodal pulse waves (PWs) sensor integrating a piezoelectric electrocardiogram (FCG) sensor and a photoplethysmography (PPG) sensor, enabling simultaneous mechanical-optical measurements of PW from the same site on the body.

Similarly, multimodal fusion algorithms have been rapidly developed. Baltrusaitis et al. [14] surveyed the progress of multimodal research and summarized five challenges faced in multimodal research, namely, representation, translation, alignment, fusion, and co-learning. For the challenge of multimodal representation, Baevski et al. [15] proposed a general framework for unified representation of speech, vision, and language through self-supervised learning. This work opens up new ideas for multimodal representation. Liu et al. [16] proposed an autoencoder-based multiview missing data completion framework (AEMVC) to overcome the problem of missing data in multimodal representation. In the translation challenge, Liu et al. [17] proposed a variational multimodal machine translation model (VMMT), which can model language uncertainty in translation to eliminate the discrepancy between training and prediction in existing variational translation models. In the field of multimodal alignment, Zhang et al. [18] proposed a general multimodal detector called AR-CNN to solve the problem of position shifts between different modalities in object detection. Luppino et al. [19] presented a novel unsupervised methodology to align the code spaces of two autoencoders based on affinity information. In terms of multimodal fusion, Ma et al. [20] proposed a fusion framework of infrared and visible images called STDFusionNet on the registered public TNO and RoadScene datasets, which can preserve the thermal targets and the details of visible images. Nagrani et al. [21] proposed a novel transformer architecture [Multimodal Bottleneck Transformer (MBT)] that uses "fusion bottlenecks" for modality fusion at multiple layers and improves performance over vanilla cross-attention at lower computational cost. Facing the challenge of multimodal co-learning, Zadeh et al. [22] focused on the study of multimodal co-learning. They proved that the model after multimodal co-learning performed better in single-modal tasks based on information theory.

Despite the above, there is limited research on multimodal fusion detection systems specifically designed for near-surface defects, and the majority of existing algorithms for fusing infrared and visible images rely on publicly available registered datasets [12], [21], [22], [23], which are not directly applicable to the specific requirements of NDT. Thus, a physics-coupled multispectral vision sensing fusion NDT system is designed and a novel multispectral fusion defect detection framework with coarse-to-fine multispectral registration capability is proposed. The proposed system integrates IRT and VT to overcome the limitations of being susceptible to surface conditions of IRT and the inability to detect subsurface defects of VT. The acquisition and excitation system designed based on the physical attributes of each modality contributes to acquiring high-quality images and defect features, complementing the proposed algorithm and aiding in further improving the accuracy of both coarse registration (CR) and fine registration. The proposed system is capable of acquiring time-synchronized infrared and visible image pairs at specific trajectories and speeds for complex specimens for further processing. The proposed algorithm framework utilizes domain adaptation (DA) to unify the features extracted from the infrared and visible images and realizes the accurate registration of infrared and visible image pairs through the coarse-to-fine module. Besides, it can be seamlessly integrated with any object detection algorithm to achieve multispectral decision-level fusion detection. The framework leverages the complementarity and redundancy of infrared and visual information to improve the accuracy and robustness.

The rest parts of this article are organized as follows: Section II describes the details of the proposed algorithm. Section III elaborates on the detailed design of the proposed system as well as the samples and the implementation of the proposed algorithm. Experiments and results analysis are introduced in Section IV. Finally, conclusions and the future work are drawn in Section V.

II. Methodology

Commonly used multispectral fusion algorithms for infrared and visible image fusion are typically driven by publicly available spatiotemporal registration datasets. However, in practical defect detection scenarios, it becomes challenging to obtain the infrared and visible data streams that have already been registered in space and time. The limitations of acquisition equipment and control systems make achieving space-time synchronization difficult. Due to the lack of a unified feature domain between infrared and visible modalities, the available features for defect detection become limited. Consequently, directly obtaining spatially registered image pairs through these features proves to be challenging. Additionally, most infrared and visible fusion algorithms focus on generating fused images that align to human visual perception and often employ system of measurement, such as Chen–Blum metric (Q_{CB}), structural similarity index measure (SSIM), and mutual information (MI), to evaluate the quality of fused images [23]. In contrast, defect detection places emphasis on identifying defects and pays more attention to the missed detection rate of defects. To overcome these challenges, we propose a novel multispectral fusion defect detection framework with coarse-to-fine multispectral registration. The proposed multispectral fusion defect detection framework is shown in Fig. 1. This general framework is suitable for commonly used detectors, such as YOLO series, R-CNN series,

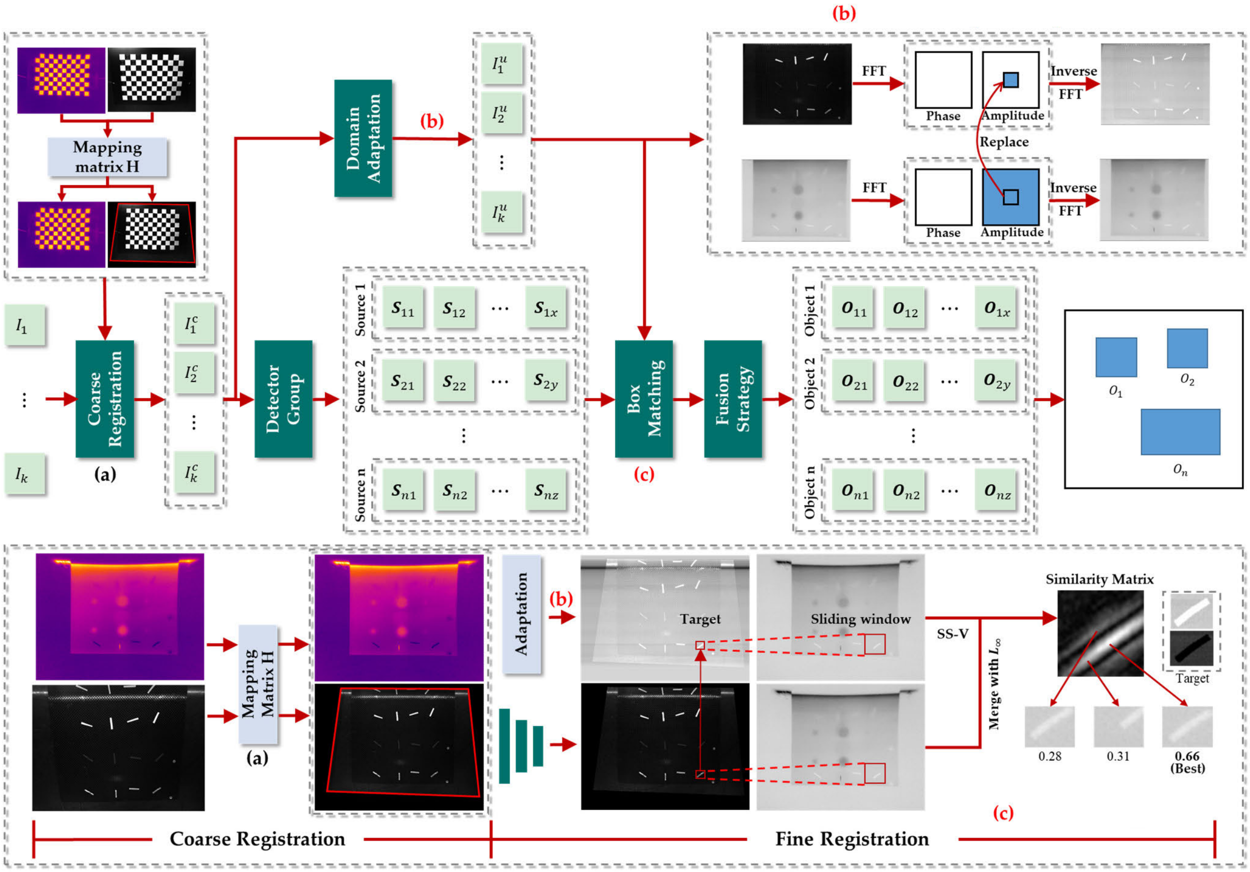


Fig. 1. Proposed novel multispectral fusion defect detection framework with coarse-to-fine multispectral registration. (a) CR with homography matrix. (b) DA for unified feature domains. (c) BM via local sliding window incorporating variance information.

and so on. Note that in our study, YOLOv5 was adopted as the baseline detector. The proposed framework can be succinctly summarized into four distinct components: CR [see Fig. 1(a)], DA [see Fig. 1(b)], box matching (BM) [see Fig. 1(c)], and fusion strategy [see Fig. 1(d)]. It is noteworthy that DA and BM are collectively referred to as fine registration.

A. Coarse Registration

Spatiotemporal registration between modalities is the basis for multispectral fusion. The region of interest is approximately assumed to be a plane due to the shape of the specimen being plane or curved with low curvature. Thus, the homography transformation can be employed to achieve CR of image pairs. Suppose the detection system obtains k raw image streams from k acquisition devices, which are denoted as $I_1, \ldots, I_k$, and $k \geq 2$. In this article, the infrared images and visible images are declared as I_1 and I_2, respectively, and $k = 2$.

Let the projection of a point $P(x_W, y_W, z_w)$ onto the plane in the infrared thermal (IR) camera pixel coordinate system and the visible camera pixel coordinate system are represented as $X_1(x_1, y_1)$ and $X_2(x_2, y_2)$, respectively. Suppose the plane is located on $z_W = 0$, the mapping of the IR camera can be expressed as follows:

$$\begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = s_1 \begin{bmatrix} f_x & \gamma & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_1 & r_2 & t \end{bmatrix}^{(3\times3)} \begin{bmatrix} x_w \\ y_w \\ 1 \end{bmatrix} \quad (1)$$

where s_1 is the scaling factor, $\begin{bmatrix} f_x & \gamma & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} = K_1$ is the intrinsic matrix of the IR camera, and $[\, r_1 \;\; r_2 \;\; t\,]^{(3\times3)} = E_1$ is the extrinsic matrix of the IR camera. In the same way, the mapping of the visible camera can be expressed as follows:

$$\begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = s_2 K_2 E_2 \begin{bmatrix} x_w \\ y_w \\ 1 \end{bmatrix} \quad (2)$$

where s_2, K_2, and E_2 are defined similarly as in (1) but for the case of visible camera. Through (1) and (2), the mapping between the point pair of the IR image and the visible image can be expressed as follows:

$$\begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = \frac{s_1}{s_2} K_1 E_1 K_2^{-1} E_2^{-1} \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} \quad (3)$$

where $(s_1/s_2)K_1E_1K_2^{-1}E_2^{-1} = H$ is a 3×3 homography matrix that represents the mapping between the point pair of the IR image and the visible image. When the relative positions of the IR camera, the visible camera, and the plane are fixed, H is a constant matrix. In this case, the spatial registration of infrared images and visible images can be achieved by solving for H. Only four pairs of points are required to calculate the unique solution of H which has eight degrees of freedom. Considering potential inaccuracies in point pairs, the optimal solution for H is derived using the least squares method on the 88 corners of the checkerboard.

B. Domain Adaptation

Domain adaptation (DA) can reduce the distribution difference between infrared and visible image pairs and unify the feature domains. Therefore, it reduces the impact of background differences on the similarity measure between infrared and visible images to obtain better registration results. According to [24], the semantic content of the image is mainly carried by the phase component of the Fourier transform. Inspired by this work, our article proposes a strategy to replace the low-frequency part of the amplitude component between the infrared and visible image pairs to reduce the distribution difference. The Fourier transform of image is defined as

$$F(I)(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} f(h,w)e^{-j2\pi\left(\frac{h}{H}u+\frac{w}{W}v\right)} \tag{4}$$

where $F(I)$ is the Fourier transform result of image I, which can be decomposed into amplitude component $F^A(I)$ and phase component $F^P(I)$.

A resizable mask is defined at the center of the amplitude component, which is expressed as follows:

$$M_\alpha(h,w) = 1_{(h,w)\in[-\alpha H:\alpha H, -\alpha W:\alpha W]} \tag{5}$$

where α is the hyperparameter, which indicates the mask size (0.01). In order to model the background distribution of the visible image I_2 that is close to the infrared image I_1, the low-frequency portion in the amplitude component of I_1 is used to replace the corresponding low-frequency portion in I_2. Note that surface defects with darker tones tend to exhibit higher absorbance, which results in their appearance as white on infrared images. As such, the inputs for DA are $255 - I_1$ and I_2. The DA result is obtained by inverse Fourier transform, which is expressed as follows:

$$\begin{aligned} I_2^u = F^{-1}\big([M_\alpha \circ F^A(255 - I_1) \\ + (1 - M_\alpha) \circ F^A(I_2), F^P(I_2)]\big). \end{aligned} \tag{6}$$

C. Box Matching

The position of the same defect in different modalities is close after CR. The reference modality and the sensed modality are introduced into the multispectral setting. Consider I_1 as the reference modality and others as the sensed modalities. The fine registration refers to the precise alignment of the bounding box in the sensed modality on the reference modality, which can be achieved by sliding a window near the corresponding position of the reference modality to find the matching box with the highest similarity.

As shown in Fig. 2, suppose the center point of B_{ij} is P, and the corresponding point of P on the reference modality is P'. Set P' as the center point of the sliding window to determine the position of the sliding window. The size of the sliding window can be calculated as follows:

$$W_w = B_w + 2 \times (N_w \times \text{Stride}) \tag{7}$$

$$W_h = B_h + 2 \times (N_h \times \text{Stride}) \tag{8}$$

where W_w and W_h and B_w and B_h are the width and height of the sliding-window and B_{ij}, respectively; N_w and N_h are the number of slides in the width direction and the height direction, respectively.

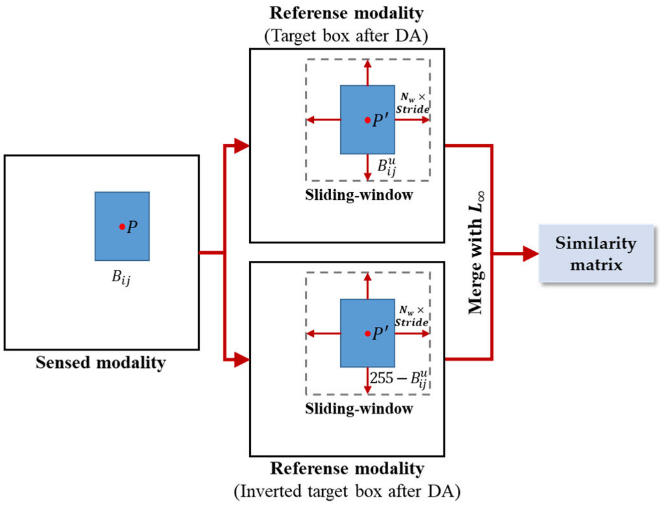


Fig. 2. Process of similarity calculation.

SSIM has been used to evaluate the similarity between infrared images and visible images [25], which includes three parts: luminance similarity score, contrast similarity score, and structural similarity (SS) score. Since the brightness and contrast of the infrared image and the visible light image are quite different, the registration mainly focuses on the structural information of the target. The SS score in SSIM is adopted to measure the similarity between boxes. The SS score is formulated as follows:

$$SS(x,y) = \frac{\sigma_{xy} + C_1}{\sigma_x\sigma_y + C_1} \tag{9}$$

where σ_x and σ_y are the variance of x and y, respectively; σ_{xy} is the covariance of x and y.

Even with the flipping of colors during the DA process, there remain several surface defects on the specimen that exhibit brighter hues relative to the background and possess elevated absorbance rates. These defects are characterized by a white appearance in both visible light and infrared images. Consequently, the target to be matched needs to slide the window on the image and the inverted image after domain adaptive processing to obtain similarity matrices, respectively. The final similarity matrix is obtained by calculating the infinite norm of these two similarity matrices

$$SS_f = \max(SS_{B\in W}(B_{ij}, B^u), SS_{B^u \in W^u}(B_{ij}, 255 - B^u)). \tag{10}$$

However, a small SS_f value will still be obtained when there is no defect target in the sliding window on the reference modality, which will interfere with the matching result. To overcome this issue, the variance (log) of the SS_f is introduced to distinguish whether or not the target is contained in the sliding window. A low variance value (threshold $<$ $\mathbf{THR_v}$) indicates that the sliding window solely encompasses the background and lacks any target.

Under the proposed SS with variance (SS-V), the matching box can be estimated as follows:

$$B_{ij}^{'} = \begin{cases} \arg\max(SS_f), & \log(\mathrm{Var}(SS_f)) > \mathbf{THR}_{\boldsymbol{v}} \\ B_{ij}, & \text{others.} \end{cases} \tag{11}$$

D. Fusion Strategy

Fusion strategy contains the strategy of bounding box fusion and class fusion.

1) Bounding Box Fusion Strategy: Inspired by non-maximum supression (NMS) [26] and weighted boxes fusion (WBF) [27], bounding box fusion strategy divides all prediction boxes into different objects by intersection-over-union (IoU) and merges the boxes of the same target into a fusion box according to specific rules. The detailed process is summarized as follows.

1) Declare empty list O for objects clusters. Add the prediction results of the reference modality $(S_{11}, \ldots, S_{1m})$ to the lists $O_1, \ldots, O_m$, respectively.
2) Iterate through the predicted boxes of the sensed modality and try to find a corresponding box in the list O. The correspondence is defined as a large overlap between boxes (IoU $> \mathbf{THR}_{\boldsymbol{c}}$). If the correspondence is found, add the prediction result S_{ij} containing the box to the list O containing the corresponding box, else create a new object list O_{m+1} and add the prediction result S_{ij} containing the box to O_{m+1}.
3) In order to minimize the fusion error, the fusion strategy is determined by whether there is a prediction result S_{ij} from the reference modality in O_k. Suppose there are N prediction results in O_k. If there is such an S_{ij}, the fusion result B_{kF} is B_{ij}; otherwise, the fusion result is the weighted average of all bounding box in O_k. The strategy can be formulated as follows:

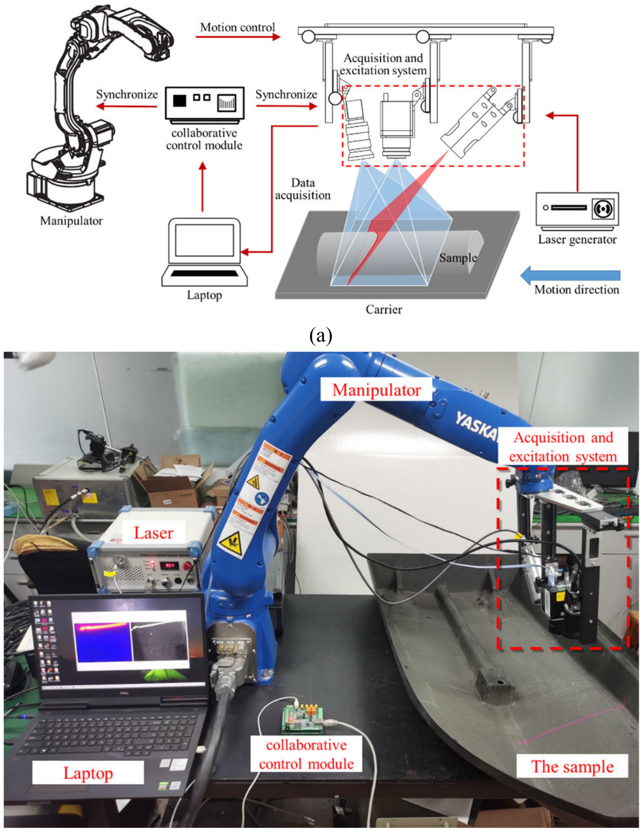
$$B_{kF} = \begin{cases} B_{ij}, & \exists B_{ij} \in S_1 \\ \dfrac{1}{N} \displaystyle\sum_{B_{ij} \in O_k} B_{ij}, & \text{others.} \end{cases} \tag{12}$$

2) Class Fusion Strategy: Class fusion strategy is divided into confidence score fusion and class fusion. The fusion result of confidence score is the average of all N scores accumulated in O_k, with the following fusion formulas:

$$CS_{kF} = \frac{1}{N} \sum_{CS_{ij} \epsilon O_k} CS_{ij}. \tag{13}$$

The class fusion strategy is designed manually according to the modality characteristics and the relationship between the modalities. Different modalities determine different fusion rules. In this article, IRT can detect surface defects and subsurface defects, whereas there exists confusion in classification. VT has a strong ability to detect surface defects, while it cannot detect subsurface defects. The purpose of fusion is to reduce the influence of the surface emissivity change of the specimen on IRT and to detect and classify surface defects and subsurface defects at the same time. Therefore, the class name fusion strategy is as follows:

$$CN_{kF} = \begin{cases} \text{surface}, & \forall CN_{ij} \in O_k,\ \exists CN_{ij} \in S_2 \\ \text{sub\_surface}, & \text{others.} \end{cases} \tag{14}$$



(a)

(b)

Fig. 3. (a) Schematic of the proposed system. (b) Proposed physic coupling multispectral vision sensing fusion NDT system.

E. Quantitative Detectability Assessment

Log-average miss rate (**MR-2**) can be used to summarize the detector performance. In previous research such as in the fields of object detection and multispectral pedestrian detection [28], [29], [30], **MR-2** has been used. The **MR-2** is computed from the miss rate (MR) against false positives per image (FPPI) (log–log). The MR and FPPI are evaluated as

$$\mathrm{MR} = \frac{\mathrm{FN}}{\mathrm{GT}} \tag{15}$$

$$\mathrm{FPPI} = \frac{\mathrm{FP}}{\mathrm{N}} \tag{16}$$

where FN is the false negative, FP is the false positive, GT is the number of the ground truth, and N is the number of the positive. Finally, **MR-2** is calculated by averaging MR at nine FPPI rates evenly spaced in log-space in the range 10^{-2}–10^{0}, and the lower score represents better performance.

III. Experimental Setup

A. Experimental Setup of the Proposed Fusion NDT System

The schematic of the proposed physic coupling multispectral vision sensing fusion NDT system is shown in Fig. 3(a). The entire system consists of a manipulator, an acquisition and excitation system, a laser generator, a collaborative control module, and a PC. The manipulator controls the movement of the acquisition and excitation system, where it can detect samples with different shapes at different speeds. The acquisition and excitation system include laser head, IR camera, and visual camera. The laser generator is emitting a line laser for excitation. When excitation system scans the sample at a certain speed with a specific trajectory, the thermal distribution and visual information on the surface of the sample will be captured by IR camera and the visual camera, respectively. The collaborative control module directly controls the movement of the manipulator while it indirectly controls the laser generator, IR camera, and visual camera through the computer, so as to realize the synchronization and coordination of the movement of the manipulator, the excitation, and the acquisition of the IR camera as well as the visual camera. Consequently, the thermal distribution reflects the information of defects, surface stains, and the texture information by the IR camera and the visual camera for further analysis. Fig. 3(b) shows the proposed system. The experimental parameters are described as follows.

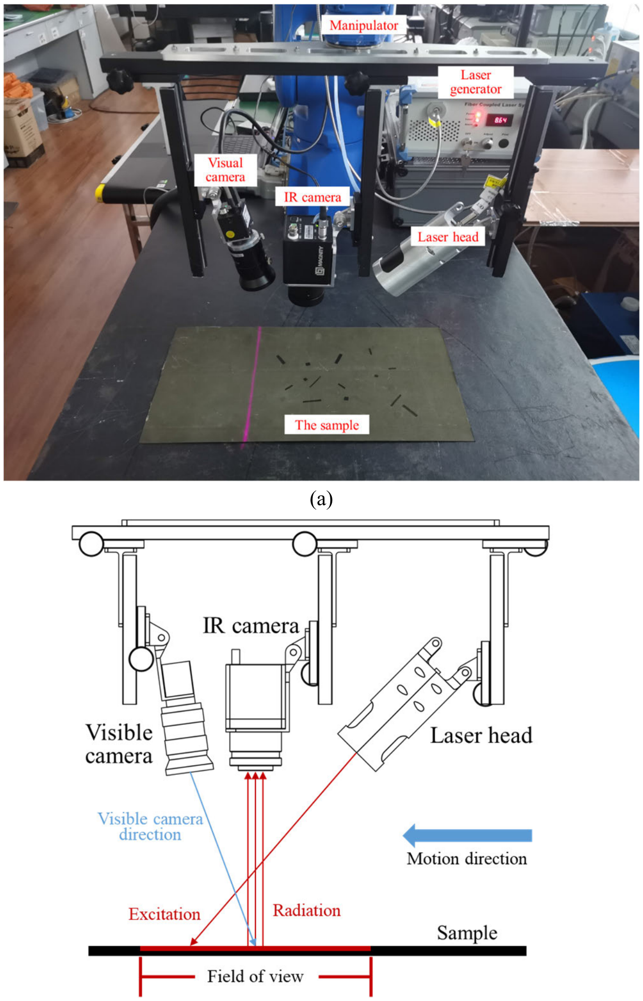
1) The working power of the laser generator is set to 20 W. The focal length of the laser head is 25 cm, and the focal area is 3 mm wide and 20 cm long.
2) The movement speed of the manipulator can be set 0–1 m/s, and the movable space range is a spherical space of 1 m. The movement accuracy is 10 μm. Due to the varying thermal conductivity of different materials, a lower thermal conductivity requires a higher excitation power. After experimental verification, the scanning speed was set to 20 and 5 mm/s for CFRP and GFRP materials, respectively, to provide optimal excitation.
3) MAGNITY MAG62 is chosen as the IR camera. The frame rate and resolution are 25 Hz and 640 × 480 array, respectively. The thermal sensitivity is 0.06 °C.
4) The visual camera is chosen as MV-SUF1200GM, which can sample at 25 Hz and has a resolution of 4096 × 3000 array. The lens is chosen as MV-LD-12-20M-A.

B. Acquisition and Excitation System

The proposed acquisition and excitation system contributes to acquiring high-quality images and defect features for further processing. The system designed based on the physical attributes of each modality is shown in Fig. 4(a) and (b). The placement of the IR camera is intricately tied to the fundamental infrared radiation theory known as Lambert's cosine law, which is expressed as follows:

$$I_\theta = I_n \cos\theta \tag{17}$$

where I_θ represents the radiation intensity in the direction θ with the normal line of the radiation surface (radiation intensity in the observation direction), and I_n represents the radiation intensity in the normal direction of the radiating surface. In order to capture the strongest radiation intensity during practical inspections, the observation direction of the IR camera should be perpendicular to the plane of the specimen.



(a)

(b)

Fig. 4. (a) Acquisition and excitation system. (b) Schematic of the acquisition and excitation system.

In addition, to maintain field of view (FOV) consistency, the direction of the visible camera should be toward the perpendicular point between the observation direction of the IR camera and the plane of the specimen, as shown in Fig. 4(b). However, the tilting of the visible camera may cause image blurring in specific regions due to variations in depth of field. When the work distance greatly surpasses the focal length, the depth of field of the camera can be calculated as follows:

$$\Delta L_1 = \frac{F\delta L^2}{f^2 + F\delta L} \tag{18}$$

$$\Delta L_2 = \frac{F\delta L^2}{f^2 - F\delta L} \tag{19}$$

where ΔL_1 and ΔL_2 represent the narrow depth of field and large depth of field, respectively, F and f represent the aperture and the focal length of the lens, respectively, L is the work distance, and δ is the diameter of the permission circle of confusion. The relevant parameters employed in the experiments are listed in Table I.

Substituting the mentioned parameters into (18), the narrow depth of field is calculated to be 42.5 mm. By employing the geometric relationships illustrated in Fig. 5, it can be calculated that the tilt angle of the visible camera should not exceed 28.9° to ensure high-quality visible images. The

TABLE I
SOME PARAMETERS OF VISIBLE CAMERA

Parameter	Value
$\boldsymbol{F},\ \boldsymbol{f}$	4, 12 mm
$\boldsymbol{L}$	150 mm
$\boldsymbol{\delta}$	0.095 mm
FOV angle	71.2° × 56.3°

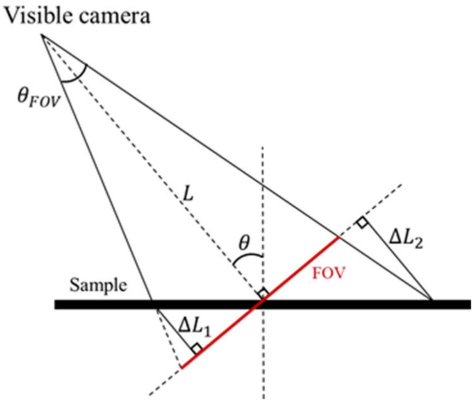


Fig. 5. Geometric relationships of the visible camera.

calculation of tilt angle is equivalent for narrow depth of field and large depth of field.

Furthermore, the surface temperature of the sample excited by the line laser is given by Cramer and Winfree [31] as

$$T(x) = \frac{q}{\pi\kappa}e^{\frac{-vx}{2\alpha}}\left[K_0\left[\frac{v|x|}{2\alpha}\right]+2\sum_{n=1}^{\infty}K_0\left[\frac{v(x^2+(2nL)^2)^{\frac{1}{2}}}{2\alpha}\right]\right] \tag{20}$$

where v is the scanning velocity, α is the thermal diffusivity, which is determined by the inherent properties of the material, and L is a constant representing the thickness of the plate. According to (20), v should be proportional to α to maintain a constant dynamic range of the surface temperature. This provides a theoretical basis for setting varying velocity for different materials of the sample during experimentation. Thus, when subsurface defects are present, there is a time delay for temperature to reach the defect and feedback to the surface. Hence, the ideal observation point should be located behind the heating point, which corresponds to the cooling phase [32]. Moreover, setting the observation point in the cooling phase can also avoid excitation interference and improve the signal-to-noise ratio.

C. CR Configuration

In this study, a novel calibration board is designed for the CR of thermal and visual cameras, as shown in Fig. 6. The novel calibration board is made of a printed circuit board with a white-on-black coating on the surface to provide a visually salient checkerboard pattern for the camera. As illustrated in Fig. 6(b), dense circuits are buried under the black squares to generate heat when energized, which provides an infrared salient checkerboard pattern for the IR camera. Simultaneously, one-to-one correspondence between black squares and buried dense coils maintains spatial consistency, which provides the same checkerboard pattern for further registration between IR camera and visual camera.

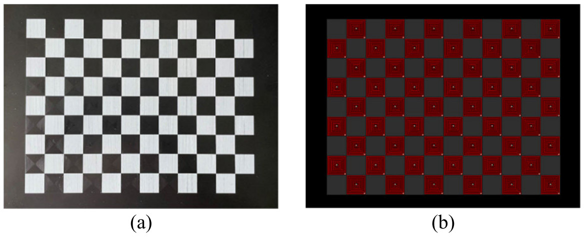


Fig. 6. Calibration board. (a) Appearance picture. (b) Internal dense circuits.

When acquiring CR data from the powered calibration board, the optimal observation time is determined based on the heat conduction process. This process can be regarded as the accumulation of the instantaneous point heat sources in space and time. Temperature variation at x of the instantaneous point heat sources is expressed as follows:

$$T(x,t) = \frac{Q}{c\rho(4\pi\alpha t)^{3/2}}e^{-\frac{x^2}{4\alpha t}} \tag{21}$$

where Q is the energy, and c, ρ, and α are the heat capacity, density, and thermal diffusivity, respectively. The temperature contrast between the reference point and its neighboring point is expressed as follows:

$$\Delta T(t) = T(0,t) - T(\Delta x,t). \tag{22}$$

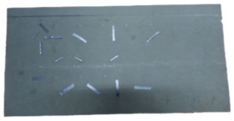
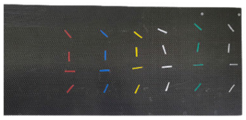
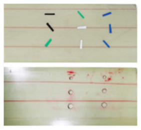
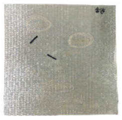
Suppose $\Delta T'(t)$ be the derivative of $\Delta T(t)$. Since all the variables are positive, it is easy to get $\Delta T'(t) < 0$, which means $\Delta T(t)$ is decreasing. Thus, the images with the largest temperature contrast at the beginning of excitation should be collected for CR.

D. Samples Preparation

In order to verify the robustness and accuracy of the proposed system and algorithm, seven standard and natural samples with different shapes and materials were chosen for testing. Four of them are standard specimens, and the other three are natural samples with natural defects, which are caused in a real industrial production environment. The first two standard specimens are flat CFRP with buried standard subsurface defects. Strips of different colors represent different surface defects. The sample no. 3 is flat glass fiber-reinforced plastic (GFRP) with through-bottom void standard subsurface defects. The sample no. 4 is curved CFRP with buried standard subsurface defects, which remains difficult for detection. The sample no. 5 and no. 6 are flat GFRP with natural defects. The sample no. 7 is the hull of a used unmanned ship made of CFRP, which is the most difficult test sample due to its complex structure.

Since different materials have different thermal conductivities, different excitation powers need to be used when detecting specimens with different materials. Concurrently, the trajectory of the manipulator will be set according to the shape of

TABLE II
DESCRIPTIONS OF SEVEN SPECIMENS

No.	Samples	Types	Defect number
1		Standard CFRP Flat	Surface: 15 Subsurface: 16
2		Standard CFRP Flat	Surface: 24 Subsurface: 30
3		Standard GFRP Flat	Surface: 9 Subsurface: 6
4		Standard CFRP Curve	Surface: 12 Subsurface: 32
5		Natural CFRP Flat	Surface: 2 Subsurface: 2
6		Natural GFRP Flat	Surface: 2 Subsurface: 25
7		Natural CFRP Special-shaped	Surface: 6 Subsurface: 4

the specimen. In the experiment, the laser power is constant, and the moving speed of the manipulator is changed to adjust the excitation power. The relevant descriptions of the different specimens and the corresponding experimental settings are shown in Table II.

E. Ground Truth of the Fusion Result

The ground truth of the fusion result reflects the defect information obtained by combining infrared and visible images, which cannot be obtained by directly labeling a single image as the ground truth from traditional target detection. Therefore, this article designs specific rules to obtain the ground truth of fusion results for both infrared and visible image pairs, which also provides a precedent of multispectral fusion detection.

The time-synchronized infrared and visible image pairs are obtained through the acquisition and excitation system, and then, the spatial registration as well as resolution unification (640 × 480) of the image pairs are realized through CR. Next, Labelme toolkit is used to label the infrared images and visible images of spatiotemporal registration to obtain $\mathrm{IR}_{\mathrm{gt}}$ and $\mathrm{VIS}_{\mathrm{gt}}$, respectively. Defects of the infrared and visible images are labeled named as "sub_defect" and "defect," respectively. Finally, $\mathrm{IR}_{\mathrm{gt}}$ is adjusted artificially against $\mathrm{VIS}_{\mathrm{gt}}$ to obtain the ground truth of fusion result (F_{gt}) according to the following rules: traverse the target box in $\mathrm{VIS}_{\mathrm{gt}}$, if the target can be detected in $\mathrm{IR}_{\mathrm{gt}}$, change the label of the corresponding target box in $\mathrm{IR}_{\mathrm{gt}}$ to "defect"; otherwise, add the target box in $\mathrm{VIS}_{\mathrm{gt}}$ to $\mathrm{IR}_{\mathrm{gt}}$.

TABLE III
SOME PARAMETERS OF THE PROPOSED MODEL

Hyper-Parameter	Value
Batch Size, epochs	64, 300
Momentum	0.937
Weight decay	0.0005
Warmup epochs, momentum, bias lr	3, 0.8, 0.1
$\boldsymbol{lr_0}, \boldsymbol{lr_f}$	0.01, 0.2
$\mathbf{THR}_{\boldsymbol{v}}, \boldsymbol{THR_c}$	5.5, 0.6

F. Implementation

The proposed algorithm is developed in Python and consists of CR and fine registration. The dataset for the CR model is derived from the dynamic scanning of calibration boards. Due to the varying scan speeds required for different materials mentioned earlier, we obtained CR data for CFRP and GFRP materials at speeds of 5 and 20 mm/s, respectively, to mitigate registration errors stemming from speed discrepancies. CR is carried out on the temporally synchronized visible and infrared datasets collected from seven samples, utilizing respective CR data to accomplish spatial registration between infrared and visible image pairs while standardizing image dimensions.

The fine registration was integrated into the base detector YOLOv5s using the PyTorch framework. Individual infrared and visible detection models are trained for each of the three natural specimen datasets due to their unique defect characteristics, whereas the datasets from four standard specimens serve to train a unified infrared and visible detection model. Identical parameters are employed for each training session across these datasets. The data are randomly partitioned into 80% for training and 20% for validation. The training adopts a warmup strategy, leverages stochastic gradient descent (SGD) as the optimizer, and implements OneCycleLR for the learning rate optimization strategy. The values of some parameters deployed during the training are shown in Table III. All the experiments are conducted on a GeForce RTX 3080 GPU with 20-GB RAM.

IV. Experimental Analysis

A. Comparison With Typical Methods

The proposed method was compared with traditional single-spectrum detection methods and the proposed decision-level fusion strategy with other similarity measures by utilizing **MR-2** as the quantitative evaluation index. Traditional

TABLE IV
COMPARISON RESULTS OF SIX TYPICAL METHODS ON SEVEN SPECIMENS. RED INDICATES THE BEST RESULT, AND BLUE REPRESENTS THE SECOND BEST RESULT

No.	Traditional NDT		Statistical-based		Deep learning-based		ours	
	IRT	VT	MI[33]	InteNCC[34]	Matchnet[35]	DeepDIM[36]	SSIM-V	SS-V
1	89.22	50.26	42.63	17.57	31.25	**11.09**	23.96	**13.18**
2	73.92	61.26	40.94	35.06	39.10	33.54	**9.71**	**9.85**
3	81.33	50.00	24.18	59.28	48.86	12.42	**10.93**	**11.42**
4	50.28	84.99	68.85	51.29	31.24	46.47	**7.10**	**6.52**
5	50.30	67.90	22.93	**17.65**	**6.98**	29.40	**6.98**	**6.98**
6	64.61	50.00	21.37	**4.26**	22.44	**4.71**	**4.71**	**4.71**
7	50.21	68.36	47.09	42.44	18.37	36.07	**16.69**	**6.36**
Mean	65.70	61.82	38.28	32.51	28.32	24.81	**11.44**	**8.43**
Time (s/frame)	-	-	7.04	0.37	135.81	507.60	6.33	6.80

TABLE V
QUANTITATIVE RESULTS OF ABLATION STUDY

	CR	BM	DA	Flat						Special			All
				1	2	3	5	6	Mean	4	7	Mean	Mean
ours				-	-	-	-	-	-	-	-	-	-
	√			**10.79**	14.24	**11.42**	8.74	**4.71**	9.98	10.84	18.37	14.61	11.30
	√	√		47.60	22.97	21.97	33.68	9.17	27.08	**5.39**	15.37	10.38	22.31
	√	√	√	13.18	**9.85**	**11.42**	**6.98**	**4.71**	**9.23**	6.52	**6.36**	**6.44**	**8.43**

single-spectrum detection methods are mainly IRT and VT. Simultaneously, MI [33], InteNCC [34], Matchnet [35], and DeepDIM [36] were chosen as the comparison methods of similarity measurement in the proposed method. They are representative of traditional statistical-based and deep learning-based methods for local registration. Table IV presents the comparative outcomes of the proposed approach with respect to other methods on all seven samples.

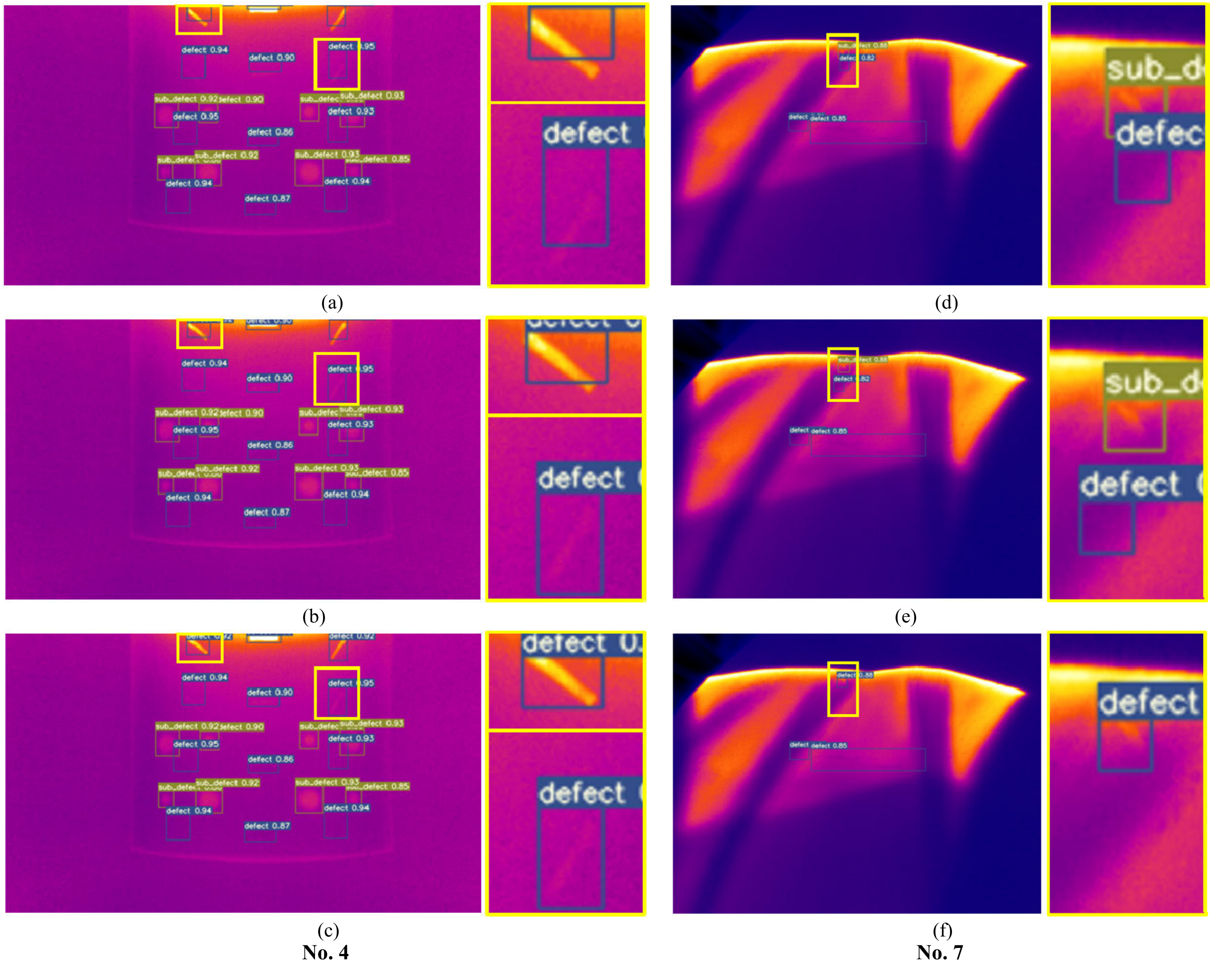
The results indicate that InteNCC, Matchnet, and DeepDIM achieved the best performance with **MR-2** of 4.26, 6.98, and 11.09 on no. 6, 5, and 1 samples, respectively. However, these methods exhibited poor performance on other specimens with an average **MR-2** of up to 20–40. In contrast, the proposed method obtained an average **MR-2** of 8.43, which represents an improvement of 16.38 over the best comparison method DeepDIM and achieved either optimal or suboptimal performance on all specimens. The robustness and accuracy of the proposed method are significantly better than the comparison methods.

The running speed of each registration algorithm was also evaluated by measuring the registration time for each image pair. The corresponding results are provided in Table IV. The proposed framework demonstrates an average processing time of 6.80 s per image pair, outperforming deep learning-based methods, which require 135.81 and 507.60 s, respectively. Nevertheless, there is still potential for improvement when compared to the processing time of the InteNCC algorithm accelerated by the integral graph, which takes only 0.37 s per image pair.

B. Ablation Study

In this section, ablation studies on the proposed method were performed on all seven samples to further analyze the individual components and their contribution to the overall performance. For each experiment, only the component under study is removed while preserving all other components unchanged. Specimens numbered 4 and 7 are special-shaped, while the others are flat-shaped. Table V presents the quantitative results of the ablation experiments, while Fig. 7 provides a detailed illustration of each step.

1) Coarse Registration (CR): Due to the inconsistent resolution and FOV between the original image pairs collected by the visual camera and the IR camera, the original image pairs cannot be used for multispectral fusion detection directly. As shown in Table V, the low **MR-2** (9.98) of the image pairs collected on the flat specimens shows that high registration accuracy can be achieved after **CR**. However, poor registration accuracy is indicated by an average **MR-2** of 14.61 for specimens with specific shape, which is shown in Fig. 7(a) and (d).



(a) (d)

(b) (e)

(c) (f)

No. 4 No. 7

Fig. 7. Details of ablation experimental results. (a) and (d) CR. (b) and (e) CR_BR. (c) and (f) Ours.

2) Box Matching (BM): Based on the CR, we further add the BM to the multispectral fusion defect detection framework and evaluate its contribution. Table V illustrates that the **MR-2** on special-shaped specimens decreased (no. 4 decreased 5.45 and no. 7 decreased 3.00), while it increased on flat specimens after adding the BM module (flat samples increased 17.10 on average). The relevant details can be seen in Fig. 7(b) and (e). The BM module demonstrates the capability to further reduce the **MR-2** and improve the registration accuracy of infrared and visible image pairs in specimens with special shapes. However, its poor performance on flat specimens negatively impacts the overall algorithm performance (increased 11.01 on average). The disparity in background and characteristics between visual and infrared images may account for this phenomenon. Visible images capture the visible light spectrum, whereas infrared images capture the infrared spectrum that conveys thermal information. These inherent differences give rise to challenges in achieving accurate registration between the two modalities.

3) Domain Adaptation (DA): In order to unify the feature domains of infrared and visible images pairs and further enhance the registration accuracy, the DA module was incorporated into the algorithm and compared the impact on the overall performance with and without it. Table V demonstrates that the **MR-2** is further reduced by 3.94 in specimens with special shapes after adding DA. Furthermore, the issue of increased **MR-2** on flat specimens caused by the BM module has been solved (the **MR-2** was reduced by 17.85 against CR + BM and by 0.75 against CR in flat specimens), which is shown in detail on Fig. 7(c) and (f).

V. Conclusion

This article has presented a novel multispectral fusion defect detection framework with coarse-to-fine multispectral registration, enabling simultaneous detection and classification of surface and subsurface defects. Additionally, a physics-coupled multispectral vision sensing fusion system has been designed to acquire time-synchronized multispectral data. The feasibility and generalization of the proposed framework and system are evaluated on seven specimens with varying materials and shapes. Ablation studies have demonstrated the effectiveness of coarse-to-fine registration in achieving accurate spatial alignment of infrared and visible detection boxes. Furthermore, DA reduces the distribution differences between infrared and visible images, unifying the feature domains. Comparative experiments have confirmed

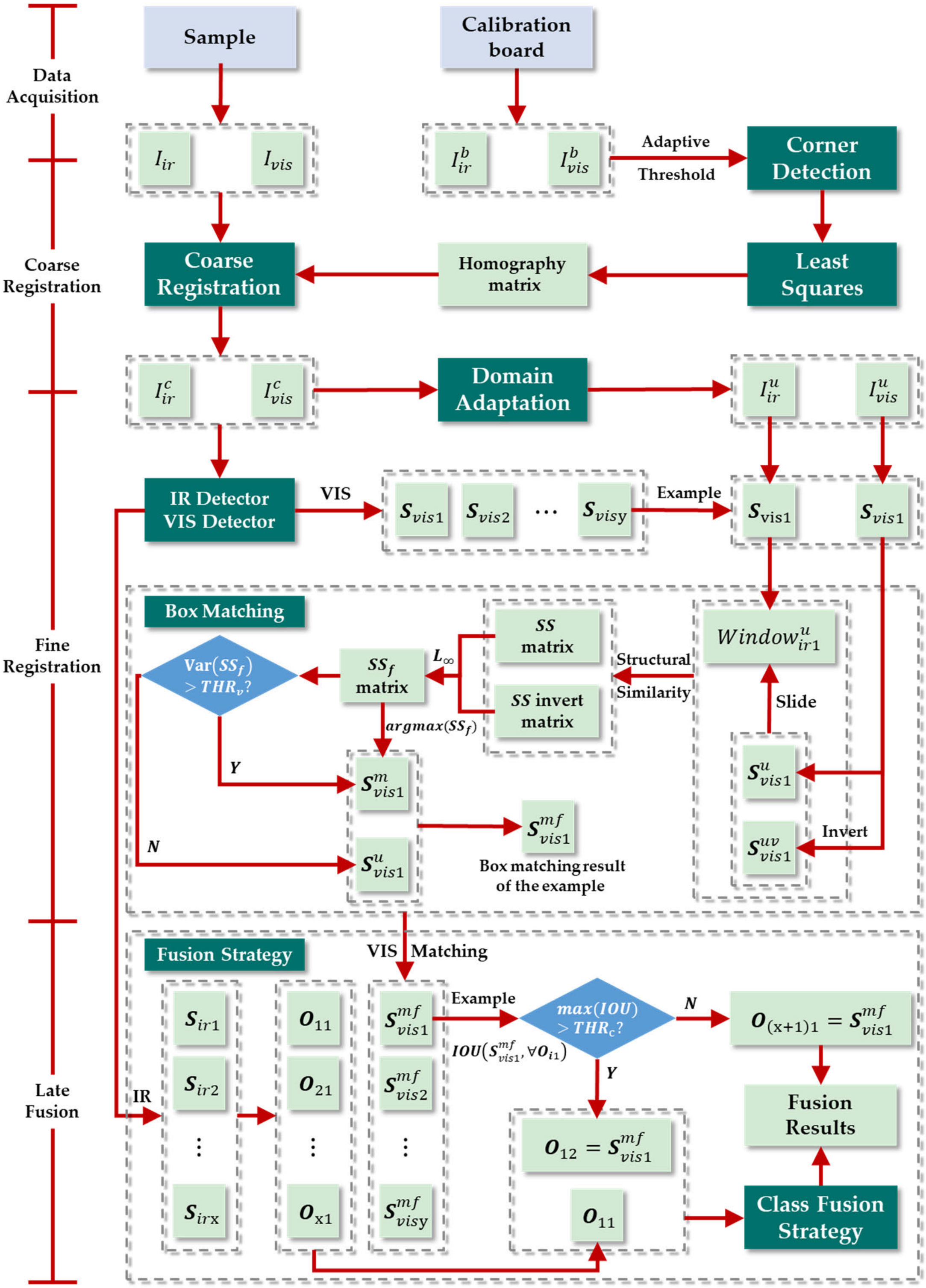


Fig. 8. Detailed flow diagram of the proposed multispectral fusion defect detection algorithm.

that the proposed framework effectively combines the advantages of IRT and VT. The proposed method achieves an impressive **MR-2** score of 8.43, outperforming the closest comparison method by a significant margin of 16.38. This notable improvement in performance showcases the enhanced robustness and accuracy of the proposed framework, demonstrating promising potential. Future work will focus on further enhancing the registration algorithm and DA module to achieve even greater improvements in speed and accuracy.

Appendix

Fig. 8 illustrates a detailed flow diagram representing the specific process and data flow of each step in the proposed multispectral fusion defect detection algorithm.

References

[1] S. Cantero-Chinchilla, P. D. Wilcox, and A. J. Croxford, “Deep learning in automated ultrasonic NDE—Developments, axioms and opportunities,” *NDT E Int.*, vol. 131, Oct. 2022, Art. no. 102703.

[2] K. Liu, Q. Yu, Y. Liu, J. Yang, and Y. Yao, “Convolutional graph thermography for subsurface defect detection in polymer composites,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[3] H. Zhang et al., “Optical and mechanical excitation thermography for impact response in basalt-carbon hybrid fiber-reinforced composite laminates,” *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 514–522, Feb. 2018.

[4] N. Puthiyaveettil, P. Rajagopal, and K. Balasubramaniam, “Influence of absorptivity of the material surface in crack detection using laser spot thermography,” *NDT E Int.*, vol. 120, Jun. 2021, Art. no. 102438.

[5] E. Ichi and S. Dorafshan, “Effectiveness of infrared thermography for delamination detection in reinforced concrete bridge decks,” *Autom. Construct.*, vol. 142, Oct. 2022, Art. no. 104523.

[6] X. Cheng and J. Yu, “RetinaNet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[7] C. Li, Y. Lei, Z. You, L. Guo, E. Zio, and H. Gao, “Vision-based defect measurement of drilled CFRP composites using double-light imaging,” *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–9, 2023.

[8] Z. Ren, F. Fang, N. Yan, and Y. Wu, “State of the art in defect detection based on machine vision,” *Int. J. Precis. Eng. Manuf.-Green Technol.*, vol. 9, no. 2, pp. 661–691, Mar. 2022.

[9] T. Schlosser, M. Friedrich, F. Beuth, and D. Kowerko, “Improving automated visual fault inspection for semiconductor manufacturing using a hybrid multistage system of deep neural networks,” *J. Intell. Manuf.*, vol. 33, no. 4, pp. 1099–1123, Apr. 2022.

[10] C. Zhang et al., “Robust-FusionNet: Deep multimodal sensor fusion for 3-D object detection under severe weather conditions,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[11] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, “Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 1852–1864, Mar. 2022.

[12] M.-C. Chang et al., “Multimodal sensor system for pressure ulcer wound assessment and care,” *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1186–1196, Mar. 2018.

[13] E. Andreozzi et al., “Multimodal finger pulse wave sensing: Comparison of forcecardiography and photoplethysmography sensors,” *Sensors*, vol. 22, no. 19, p. 7566, Oct. 2022.

[14] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[15] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, “data2vec: A general framework for self-supervised learning in speech, vision and language,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 1298–1312.

[16] Y. Liu et al., “Incomplete multi-modal representation learning for Alzheimer’s disease diagnosis,” *Med. Image Anal.*, vol. 69, no. 2, Apr. 2021, Art. no. 101953.

[17] X. Liu, J. Zhao, S. Sun, H. Liu, and H. Yang, “Variational multimodal machine translation with underlying semantic alignment,” *Inf. Fusion*, vol. 69, pp. 73–80, May 2021.

[18] L. Zhang et al., “Weakly aligned feature fusion for multimodal object detection,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9523596/metrics#metrics

[19] L. T. Luppino et al., “Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–13, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9773305

[20] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, “STDFusionNet: An infrared and visible image fusion network based on salient target detection,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[21] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, “Attention bottlenecks for multimodal fusion,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 14200–14213.

[22] A. Zadeh, P. P. Liang, and L.-P. Morency, “Foundations of multimodal co-learning,” *Inf. Fusion*, vol. 64, pp. 188–193, Dec. 2020.

[23] X. Zhang, P. Ye, and G. Xiao, “VIFB: A visible and infrared image fusion benchmark,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 468–478.

[24] M. Mancini, L. Porzi, S. R. Bulò, B. Caputo, and E. Ricci, “Boosting domain adaptation by discovering latent domains,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3771–3780.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[26] A. Neubeck and L. Van Gool, “Efficient non-maximum suppression,” in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, Aug. 2006, pp. 850–855.

[27] R. Solovyev, W. Wang, and T. Gabruseva, “Weighted boxes fusion: Ensembling boxes from different object detection models,” *Image Vis. Comput.*, vol. 107, Mar. 2021, Art. no. 104117.

[28] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[29] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, “Weakly aligned cross-modal learning for multispectral pedestrian detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5126–5136.

[30] K. Park, S. Kim, and K. Sohn, “Unified multi-spectral pedestrian detection based on probabilistic fusion networks,” *Pattern Recognit.*, vol. 80, pp. 143–155, Aug. 2018.

[31] K. E. Cramer and W. P. Winfree, “Thermal nondestructive characterization of corrosion in boiler tubes by application of a moving line heat source,” NASA Langley Res. Center, Hampton, VA, USA, Tech. Rep. NAS 1.15: 209685, 2000.

[32] D. Kaltmann, *Quantitative Line-Scan Thermographic Evaluation of Composite Structures*. Melbourne, VIC, Australia: RMIT Univ., 2008.

[33] F. Maes, D. Loeckx, D. Vandermeulen, and P. Suetens, “Image registration using mutual information,” in *Handbook of Biomedical Imaging: Methodologies and Clinical Research*. New York, NY, USA: Springer, 2015, pp. 295–308.

[34] Y. M. Fouda, “Integral images-based approach for fabric defect detection,” *Opt. Laser Technol.*, vol. 147, Mar. 2022, Art. no. 107608.

[35] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, “MatchNet: Unifying feature and metric learning for patch-based matching,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3279–3286.

[36] B. Gao and M. W. Spratling, “Robust template matching via hierarchical convolutional features from a shape biased CNN,” in *Proc. Int. Conf. Image, Vis. Intell. Syst. (ICIVIS)*. Cham, Switzerland: Springer, 2022, pp. 333–344.

Jiacheng Li received the B.Sc. degree in measurement and control technology and instrument from the University of Electronic Science and Technology of China, Chengdu, China, in 2022, where he is currently pursuing the Ph.D. degree.

His research mainly focuses on multimodal fusion, multisensor information fusion, and deep learning.

Bin Gao (Senior Member, IEEE) received the B.Sc. degree in communications and signal processing from Southwest Jiao Tong University, Chengdu, China, in 2005, the M.Sc. degree (Hons.) in communications and signal processing and the Ph.D. degree from Newcastle University, Newcastle upon Tyne, U.K., in 2006 and 2011, respectively.

From 2011 to 2013, he worked as a Research Associate of wearable acoustic sensor technology with Newcastle University. He is currently a Professor with the School of Automation Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu. His research interests include electromagnetic and thermography sensing, machine learning, nondestructive testing, and evaluation, where he actively publishes in these areas. He is also a very active reviewer for many international journals and long-standing conferences. He has coordinated several research projects from National Natural Science Foundation of China.

Wai Lok Woo (Senior Member, IEEE) received the B.Eng. degree in electrical and electronics engineering, and the M.Sc. and Ph.D. degrees in statistical machine learning from Newcastle University, Newcastle upon Tyne, U.K., in 1993, 1995, and 1998, respectively.

He was the Director of research for the Newcastle Research and Innovation Institute, and the Director of operations at Newcastle University. He is currently a Professor of machine learning with Northumbria University, Newcastle upon Tyne. He has published more than 400 articles on these topics on various journals and international conference proceedings. His research interests include the mathematical theory and algorithms for data science and analytics, artificial intelligence, machine learning, data mining, latent component analysis, multidimensional signal, and image processing.

Dr. Woo is a Member of Institution Engineering Technology. He was a recipient of the IEE Prize and the British Commonwealth Scholarship. He serves as an Associate Editor to several international signal processing journals, including *IET Signal Processing*, the *Journal of Computers*, and the *Journal of Electrical and Computer Engineering*.

Jieyi Xu received the B.Sc. degree in Internet of Things engineering from the Southwest University of Science and Technology, Mianyang, China, in 2022. She is currently pursuing the M.Sc. degree with the University of Electronic Science and Technology of China, Chengdu, China.

Her research mainly focuses on multisensor information fusion and deep learning.

Lei Liu received the B.Sc. degree in automation from the Southwest University of Science and Technology, Mianyang, China, in 2019. He is currently pursuing the M.Sc. degree in control science and engineering from the University of Electronic Science and Technology of China, Chengdu, China.

He is mainly engaged in optical pulsed thermography (OPT) combined with robotic arm to detect defects. His research interests include nondestructive testing and 3-D reconstruction.

Yu Zeng received the bachelor's degree from the Southwest University of Science and Technology, Mianyang, China, in 2021. He is currently pursuing the master's degree with the University of Electronic Science and Technology of China, Chengdu, China.

His research interests focus on infrared nondestructive testing, visual servo, robot perception, and control.