Infrared-based damage detection in thick composites via transfer learning on simulated and experimental data

Li, Muyao; Leonetti, Davide; Zappalá, Donatella; Snijder, H. H.(Bert)

DOI
10.1016/j.compositesb.2025.113112

Publication date
2026

Document Version
Final published version

Published in
Composites Part B: Engineering

Citation (APA)
Li, M., Leonetti, D., Zappalá, D., & Snijder, H. H. (2026). Infrared-based damage detection in thick composites via transfer learning on simulated and experimental data. *Composites Part B: Engineering*, *309*, Article 113112. https://doi.org/10.1016/j.compositesb.2025.113112

Composites Part B 309 (2026) 113112

Contents lists available at ScienceDirect

Composites Part B

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

Infrared-based damage detection in thick composites via transfer learning on simulated and experimental data☆

Muyao Li [a,c,*], Davide Leonetti [a,c,**], Donatella Zappalá [b], H.H. (Bert) Snijder [a]

[a] Department of the Built Environment, Eindhoven University of Technology, Eindhoven, The Netherlands
[b] Faculty of Aerospace Engineering, Delft University of Technology, Delft, The Netherlands
[c] Eindhoven Artificial Intelligence Systems Institute, Eindhoven, The Netherlands

ARTICLE INFO

Dataset link: https://doi.org/10.17632/xjf5cmkjm3.1

Keywords:
Composites
Damage detection
Step-heating thermography
Finite element analysis
Deep learning
Transfer learning

ABSTRACT

Reliable detection of subsurface defects in thick composite materials is critical for ensuring structural integrity in industrial applications such as wind turbine blades, aerospace components, and marine structures. This paper addresses dataset scarcity in AI-aided damage detection for thick composites using infrared thermography through a transfer learning framework leveraging finite element simulation data. Experimental datasets were obtained by conducting step-heating thermography experiments on glass-fiber-reinforced polymer (GFRP) and epoxy resin plates with artificial subsurface defects. Transient thermal analyses were performed on finite element models to mimic the actual step-heating thermography process, resulting in a large simulated dataset containing thermal videos representing the plate's surface thermal behavior during the heating-cooling process. Principal component thermography was used to extract features from both simulated and experimental thermal videos, compressing damage-related information in the raw data and enhancing the most informative features. Noise analysis on the experimental data revealed key differences compared to the simulated dataset. A U-Net architecture for image segmentation was implemented within the transfer learning framework, first pre-trained on simulated data and then fine-tuned with experimental data. The results revealed fundamental features shared across domains and demonstrated improved damage detectability in thick composite plates, especially for defects deeper than 15 mm. This approach demonstrates the potential of transfer learning to improve damage detection in industrial applications involving thick composite structures, such as wind turbine blades.

1. Introduction

Damage detection in composite materials is a critical aspect of structural health monitoring, particularly in industries such as aerospace, wind energy, and civil infrastructure where safety and reliability are paramount [1–3]. Composite materials, especially thick composites, present unique challenges for non-destructive testing (NDT) due to their heterogeneous nature, anisotropic properties, and the potential for complex internal damage mechanisms that may not be visible on the surface [4].

Infrared (IR) thermography has emerged as a powerful NDT technique for composite materials due to its non-contact nature, the ability to rapidly inspect large areas, and the sensitivity to subsurface defects [5,6]. Despite these advantages, traditional IR thermography methods, such as pulsed thermography, face limitations in detecting defects at greater depths, particularly beyond 10–15 mm in thick composites [7,8], because thermal waves diffuse and attenuate rapidly with depth, yielding low surface temperature contrast and poor signal-to-noise ratio. Composite components used in industrial applications can easily reach thicknesses of up to 30–40 mm, making them prone to deep delamination that is difficult to detect using conventional thermography techniques. Step-heating thermography has shown to be promising for thick composites, as it employs continuous thermal excitation to achieve greater heat penetration, allowing the detection of relatively deeper defects [9,10]. Longer heating times improve the capability to detect deeper defects. However, they may also lead to material overheating and irreversible deterioration of mechanical properties. A previous study [11] reported a decrease in the tensile strength of GFRP by 26.6% at 70 °C, compared to its value at room temperature. Detecting defects at a depth of 30 mm requires about 10 min heating with three 1000 W lamps placed at a distance of 1 m, which can raise the surface temperature to around 50 °C [12]. Thus, special care must be taken to avoid overheating the material. More advanced IR

☆ This work was supported by the Dutch Research Council (NWO).
* Correspondence to: P.O. Box 513, 5600 MB, Eindhoven, The Netherlands.
** Corresponding author.
E-mail addresses: m.li6@tue.nl (M. Li), d.leonetti@tue.nl (D. Leonetti).

https://doi.org/10.1016/j.compositesb.2025.113112
Received 2 June 2025; Received in revised form 4 October 2025; Accepted 16 October 2025
Available online 20 October 2025

List of abbreviations

Acronyms

GFRP	Glass-fiber-reinforced polymer
IR	Infrared
NDT	Non-destructive testing
HD	Heating condition
FE	Finite element
PC	Principal component
PCA	Principal component analysis
PCT	Principal component thermography
CNN	Convolutional neural networks
DL	Deep learning
IoU	Intersection over union
ReLU	Rectified linear unit
LOOCV	Leave-one-out cross-validation
CI	Confidence interval

Datasets

Lab_C	Experimental dataset of composites
Lab_R	Experimental dataset of epoxy resin
Sim_R	Simulated dataset of epoxy resin

Models

BM_C1	Benchmark U-Net trained on Lab_C with 9 PCs
BM_C2	Benchmark U-Net trained on Lab_C with 8 PCs
BM_R1	Benchmark U-Net trained on Lab_R with 9 PCs
BM_R2	Benchmark U-Net trained on Lab_R with 8 PCs
PM1	Pre-trained U-Net trained on Sim_R with 9 PCs
PM2	Pre-trained U-Net trained on Sim_R with 8 PCs
RM_R	Re-trained U-Net trained on Lab_R
RM_C	Re-trained U-Net trained on Lab_C
RM_R_C	Re-trained U-Net sequentially trained on Lab_R and Lab_C
SegNet_R	SegNet trained on Lab_R
DeepLab_R	DeepLabv3+ trained on Lab_R
SegNet_C	SegNet trained on Lab_C
DeepLab_C	DeepLabv3+ trained on Lab_C

thermography techniques, such as lock-in thermography, have demonstrated superior performance in deeper damage detection [13], but their complex methodology and strict testing conditions often make them unsuitable for industrial applications like the remote inspection of wind turbine blades. Additionally, noise in thermal measurements further complicates the identification of thermal patterns associated with deep defects [14,15].

Recent advances in artificial intelligence and deep learning (DL) have revolutionized image-based NDT by enhancing defect detection capabilities beyond what is possible with traditional signal processing techniques [16,17]. Convolutional neural networks (CNNs) have demonstrated remarkable success in various NDT applications [18–22]. In particular, the U-Net architecture, originally developed for biomedical image segmentation [23], has proven to be effective for precise damage detection and localization, especially in thin composites [18,24]. The U-Net is a symmetric encoder–decoder CNN that pairs down-sampling layers with up-sampling layers via skip connections to produce pixel-wise predictions. It captures contextual information in the contracting path and restores spatial details in the expansive path, yielding precise masks from input images [23].

However, the implementation of DL methods for NDT faces a major challenge: the scarcity of labeled data [25]. Unlike in general photographic image processing, where large datasets are readily available, NDT applications typically rely on limited experimental data due to the significant costs and time associated with specimen preparation and testing [26].

To address this challenge, researchers have explored various strategies, including data augmentation [27,28], synthetic data generation using advanced generative models [29–31], and transfer learning [21,32–34]. Among these, transfer learning is especially promising. As an advanced technique in the deep learning field, transfer learning is defined as the process of reusing the knowledge learned on a source task and dataset to improve learning on a different but related target task, typically by starting from a pre-trained model and adapting it rather than training from scratch. For example, an image classifier pre-trained on ImageNet, which contains more than 1 million photographic images, can have its final layers adapted to classify road damage using a limited amount of labeled data [33]. This approach has been successfully applied in various domains, including medical imaging [23], remote sensing [35], and structural health monitoring [36].

In the context of NDT for composites, finite element (FE) simulation offers a viable and flexible approach to generate synthetic data to train DL models [32,37]. By conducting transient thermal analysis, FE models can replicate thermal behavior under various conditions and defect scenarios, enabling the creation of diverse datasets that would be impractical and costly to obtain experimentally. However, the credibility of simulation strongly depends on demonstrating agreement between the FE model and relevant experiments within the intended domain of use. Thus, effective model validation within a specific parameter space is essential. The effectiveness of simulation-based training also depends critically on the extent to which the simulated data accurately reflect real-world thermal behaviors, accounting for failure modes, material properties, boundary conditions, and noise characteristics. Several studies have explored the integration of simulation and experimental IR data for damage detection in composites. Fang et al. [38] showed that synthetic data generated by FE simulations can be effectively merged with limited experimental data to train neural networks. Similarly, Tong et al. [39] validated a neural network trained only on simulated data using experimental IR results. It was demonstrated that synthetic IR data share some common features with real IR data. However, the literature concentrates mainly on the investigation of thin composites which are no thicker than 5 mm. The effectiveness of combining simulated and experimental IR data for thicker composites is still underexplored. Transfer learning has promising potential in the detection of deeper defects by learning from a large simulated dataset that contains features of deeper defects, transferring the knowledge to a fine-tuned model aimed at detecting defects in real IR images. The detectability of deeper defects is enhanced by the transferred knowledge.

This paper addresses the challenge of dataset scarcity in damage detection in thick composites by developing a robust transfer learning framework, which takes advantage of a large simulated dataset to enhance performance on limited experimental data. Step-heating thermography experiments and transient thermal analysis are conducted on physical plate samples and FE models, respectively. The simulation focuses on single material as in the experiments, and includes thermal maps produced under 960 distinct scenarios, such as varying defect sizes, shapes, locations, and heating conditions. The proposed framework reduces the difference between simulated and real IR data by adding synthetic noise to the simulation results and improves the

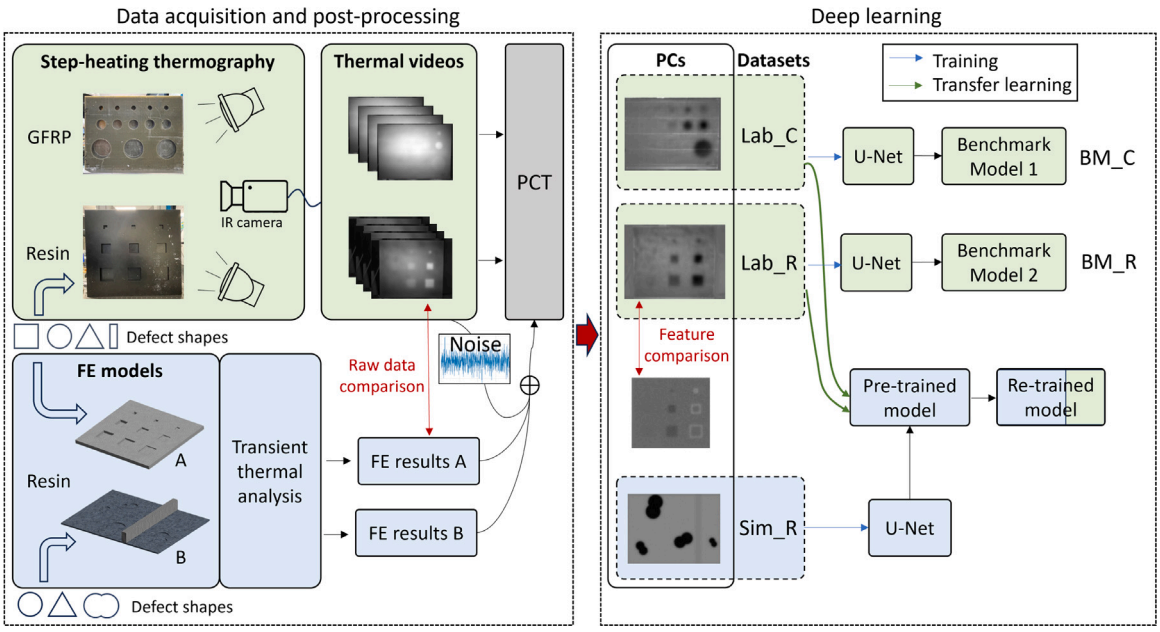

Fig. 1. Illustration of the proposed methodology in detail.

performance of the DL model in the detection of deeper defects by transferring more diverse features from simulated data to real IR data.

The rest of this paper is organized as follows. In Section 2, the methodology of this study is explained in detail: Section 2.1 provides a description of the IR experiments and the specimens, along with an introduction to the noise analysis for the experimental data. Section 2.2 gives a detailed description of the FE analysis that is used to generate the simulated dataset. Section 2.3 introduces the supervised learning procedure comprising PCT and U-Net, which is implemented for the image segmentation task. A theoretical description of transfer learning is given in Section 2.4. Following, Section 3 presents and discusses the results of the experiments and simulations, along with the performance of the DL model in the proposed framework. Finally, the conclusion of this work and the recommendations for future work are discussed in Section 4.

2. Methodology

The methodology proposed in this work is illustrated in Fig. 1. A small experimental dataset is developed by conducting step-heating thermography tests on four epoxy resin plates and one GFRP plate with different subsurface defect characteristics. This results in two datasets, namely *Lab_R* and *Lab_C*, respectively. The noise pattern in the experimental results is studied and extracted from the thermal videos.

A large simulated dataset, referred to as *Sim_R*, is developed by performing transient thermal analyses using a parametric FE model incorporating a wide variety of subsurface defects. Accurately modeling heat transfer in composite laminates is challenging due to their anisotropic nature. The commonly adopted isotropic assumption tends to underestimate in-plane conduction while overestimating through-thickness conduction, yielding temperature errors in the order of 2–3 °C [40]. To address this issue, epoxy resin is used instead of composites to preserve similar but homogeneous thermal properties. This choice allows for simpler and more reliable validation of the FE models, as the thermal behavior of epoxy resin is well-established [41].

Synthetic noise is added to the simulated results to increase the similarity to the experimental results. Both simulated and experimental thermal videos are processed with PCT to extract informative features from the raw data, which are then utilized to train the neural networks.

A U-Net model is trained on the simulation dataset *Sim_R* to obtain a pre-trained model that captures the common features that are supposed to be shared in both experimental and simulated thermal images. The model is then re-trained on experimental datasets *Lab_R* and *Lab_C* by freezing shallow layers and updating final layers. In this way, the model is adapted to real IR data but still keeping the knowledge of the common features shared by simulated and real IR data. A parametric analysis is performed by varying the number of trainable layers to evaluate its effects on the transfer learning performance. Benchmark models are developed for damage detection in both composite and resin plates, which are further used to evaluate the improvement of the re-trained models.

The U-Net architecture is implemented for the detection of defects due to its simple architecture and promising performance in image segmentation with limited training data. Previous studies [18] have demonstrated that the U-Net outperforms conventional machine learning algorithms, such as random forest and support vector machines, in detecting 2 mm-deep delamination in composite plates via pulsed thermography data. The U-Net achieves an F1-score of 0.745, whereas none of the conventional machine learning models exceed an F1-score of 0.2. Although the U-Net provides segmentation results that contain less information of sharp angles compared to other deep learning models with more complex architecture such as SegNet and DeepLabv3+, a much simpler layer structure allows an easier investigation in the influence of using a different number of trainable layers during transfer learning. A much faster training and inference process also make the U-Net one of the most suitable DL models for IR-based damage detection in composite materials.

2.1. Experimental measurements

Following the motivation in Section 1, step-heating thermography experiments are carried out in this study to generate datasets to train DL models. The research utilized the GFRP plate sample from [12], subjecting it to new tests. In addition, four epoxy resin plates were manufactured and tested with step-heating thermography.

2.1.1. Step-heating thermography test

Step-heating thermography is an NDT method that employs continuous thermal excitation to achieve deep heat penetration, allowing the

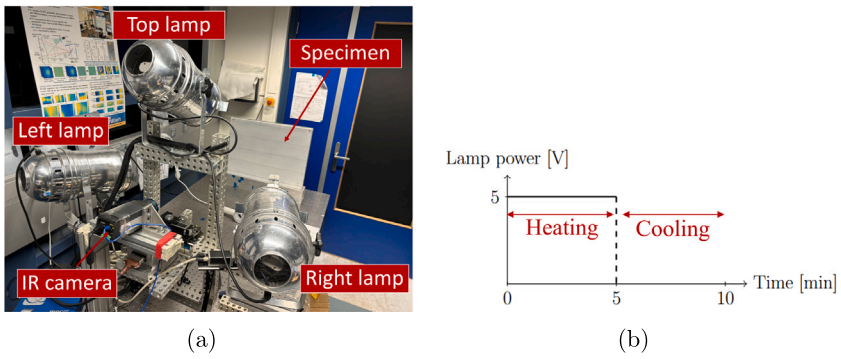

Fig. 2. (a) Experimental setup for step-heating thermography with a plate specimen. (b) Heating and cooling phases applied during each measurement cycle.

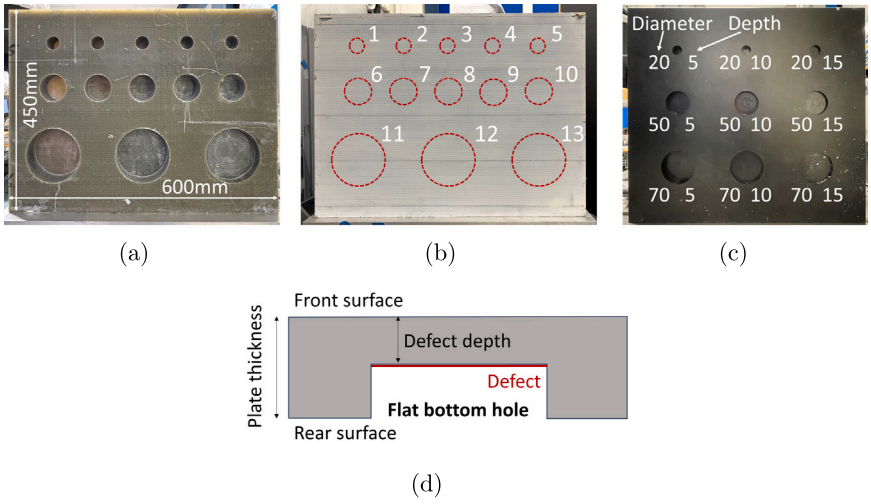

Fig. 3. Geometry and dimensions of plate specimens. (a) Rear surface of the composite plate. (b) Front surface of the composite plate. (c) Rear surface of the resin plate with circular defects. (d) Illustration of a flat bottom hole and defect depth.

detection of subsurface damage. A standard step-heating thermography setup consists of three primary components: a thermal camera, heat-generating sources, and a data acquisition and analysis system. In this research, experiments were conducted in the laser laboratory of the Aerospace Structures and Materials Laboratory at Delft University of Technology.

A schematic of the experimental setup is provided in Fig. 2(a). The surface temperature distribution was monitored using a FLIR A305sc IR thermal camera, operating at a frequency of 50 Hz with an image resolution of 640 × 480 pixels. Three halogen lamps were used as heat sources. The specimen was placed approximately 1 m away from both the heat sources and the IR camera. In the previous research of Nan [42], the heat flux on a plate specimen generated by a single lamp was calibrated with different voltages. The heat flux was observed to remain negligible at voltages below 0~3 V and plateaued above 8 V. Based on these findings, a heating voltage of 5 V was selected for each lamp, resulting in an average sample surface heat flux of 350 W/m^2.

Each measurement cycle consisted of two phases: a heating phase followed by a cooling phase, as shown in Fig. 2(b). To prevent material degradation, the heating duration was limited to 5 min, ensuring that the surface temperature of the specimen did not exceed 50 °C, well below the 60 °C threshold known to induce permanent degradation and to compromise the mechanical performance of composite materials [43]. During the 5 min heating phase, the IR camera recorded the temperature variation on the sample surface. The lamps were then turned off, while the camera continued recording for an additional 5 min to capture the cooling phase. Throughout both phases, thermal images were acquired at a rate of 2 Hz, resulting in thermal videos that contained approximately 1200 frames for each measurement.

2.1.2. Plate specimens and experimental dataset

A GFRP composite plate specimen with 13 flat bottom holes was tested with step-heating thermography experiments. This specimen, originally fabricated for the investigations conducted by [12], has overall dimensions of 600×450×51 mm ($W \times H \times D$), as shown in Figs. 3(a) and 3(b). The flat bottom holes were milled to mimic delamination defects with varying sizes and depths. The cross-sectional geometry of a single hole, illustrated in Fig. 3(d), shows the morphology of the artificial defects. In line with previous experimental work on composite materials [18,21], the depth of the artificial delamination is quantified as the material remaining thickness, measured from the front surface of the sample to the delamination plane. Table 1 provides the dimensional details of the 13 flat bottom holes introduced in the composite plate.

In addition to the composite plate, four epoxy resin plates were designed and fabricated for this study. Each resin plate has dimensions similar to the composite plate, measuring 500×450×20 mm ($W \times H \times D$), and features 9 flat-bottom holes arranged in a 3 × 3 configuration. These holes, though of the same shape, vary in size and depth, as shown in Fig. 3(c). Four different defect shapes (circular, square, triangular, and rectangular) were introduced to increase the diversity of the experimental dataset. Defects located in the same column have identical depths (5 mm, 10 mm or 15 mm).

Step-heating thermography was performed on the 5 plate specimens (1 composite plate and 4 resin plates) with the same experimental setup. Each plate was subjected to 7 distinct measurements, each corresponding to a unique heating condition (HD) defined by activating different lamps during the tests, as described in Table 2. This process generated a total of 35 thermal videos: 7 for the composite plate and

Table 1
Defect characteristics of the composite plate.

Flat bottom hole	1	2	3	4	5	6	7	8	9	10	11	12	13
Defect depth [mm]	25	20	15	10	5	25	20	15	10	5	30	20	10
Diameter [mm]	30	30	30	30	30	60	60	60	60	60	120	120	120

Table 2
Heating conditions applied during the step-heating thermography tests.

	HD1	HD2	HD3	HD4	HD5	HD6	HD7
Top lamp	✓	–	–	✓	✓	–	✓
Left lamp	–	✓	–	✓	–	✓	✓
Right lamp	–	–	✓	–	✓	✓	✓

28 for the resin plates, which together formed the two experimental datasets: *Lab_C* (for composite plates) and *Lab_R* (for resin plates).

2.1.3. Noise analysis

Noise in thermal imaging systems poses significant challenges for accurate temperature measurements and image quality. In this study, the noise in the experimental data was the main difference compared to the simulation results. In order to investigate the effects of the signal noise on the experimental results, a noise analysis was performed to study the noise patterns in the thermal signals. Previous studies [44,45] on noise in thermal imaging have shown that most thermal scenes change slowly over time because heat transfer acts as a strong low-pass process. As a result, the useful thermal signal is concentrated at low temporal frequencies, while many noise sources, such as electronics, quantization, and photon/statistical noise, contribute relatively more power at higher frequencies, up to a few tens of Hertz. In thermography tests, thermal signals are often engineered at specific modulation frequencies. For example, lock-in thermography periodically varies thermal excitation at a frequency lower than 2 Hz, while step-heating thermography uses sub-Hz modulation without any periodic excitation, resulting in an even lower frequency range than lock-in thermography.

In the image denoising field, Gaussian white noise is commonly used as synthetic noise to train data-driven denoising algorithms, demonstrating practical effectiveness despite its simplified representation. For example, in [46], an image denoiser was trained on paired clean and noisy images by modeling the noise as Gaussian. Similarly, several thermal imaging studies have assumed that noise follows a Gaussian distribution [47,48]. Thus, noise was modeled as a hybrid of multiple sources in this study, represented by a random variable following a Gaussian distribution. Since the noise exhibited frequencies significantly higher than the main thermal signal, it could be isolated using signal processing techniques. A high-pass filter was implemented to remove low-frequency components below a defined cut-off threshold, while preserving the higher-frequency elements of the signal [49]. The cut-off frequency was manually optimized to meet two essential objectives: maintaining the sharpness of the signal peaks and ensuring sufficient smoothness of the thermal data. Following these criteria, a cut-off frequency of 0.1 Hz was selected.

The high-pass filter was applied independently to each pixel in the experimental datasets *Lab_C* and *Lab_R*. For each pixel, we computed the statistical parameters of the extracted noise, specifically its mean and standard deviation. These values were then averaged across the entire datasets to characterize the overall noise pattern.

2.2. Finite element analysis

To develop supervised learning algorithms for damage detection within the transfer learning framework, finite element (FE) models were developed to generate simulation data, addressing the common challenge of limited labeled data in structural damage detection. This approach leverages the flexibility of simulation data, which allows creating extensive datasets for multiple similar structures by systematically

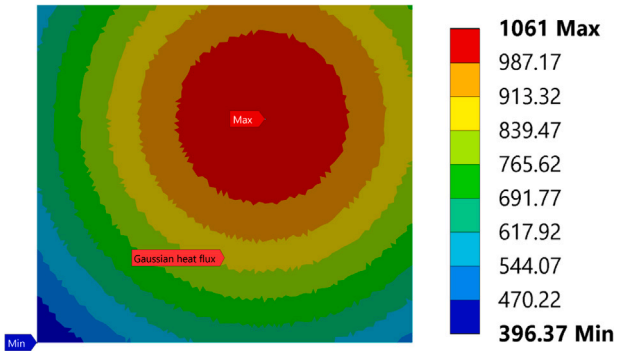

Fig. 4. Non-uniform heat flux with Gaussian distribution on the front surface of the resin plate.

Table 3
Design parameters of the FE models used for transfer learning.

Design parameters	Details
Defect shape	Circular, square, irregular
Defect depth [mm]	2–20
Plate thickness [mm]	8–25
Defect location	Randomly located
Heat flux Q [W/m^2]	200, 400, 600 and 800
Wooden web	with/without

varying the design parameters in the FE models. This section explains the development of FE models of epoxy resin plates with flat bottom holes and introduces the generated simulation dataset, which will be used in supervised learning algorithms. All FE models were developed using the ANSYS Workbench 2023 R1 [50].

2.2.1. Finite element models of epoxy resin plates

To generate a large simulation dataset, a series of FE models was developed by changing key design parameters, as shown in Table 3. The models included three different defect geometries: circular, square, and irregular. The irregular defects were modeled as two circular overlapping defects, as shown in Fig. 5. In each simulation, both the depth of the defect and the thickness of the plate were randomly selected. Each plate model contained four defects with the same shape and depth but different sizes, with defect positions randomized in each simulation. For simplification, instead of a non-uniformly distributed heat flux presented in real IR images, a uniform heat flux was applied to the front surface of the plate. To investigate the reliability of this simplification, an additional FE model with a Gaussian distributed heat flux applied to the same surface was developed, as shown in Fig. 4. The results for uniform and non-uniform heating conditions are compared and analyzed at the feature level in Section 3.2. A wooden web, shown in Fig. 5, was attached to the rear surface of the plates to replicate the structural features typical of real composite applications, such as wind turbine blades, where internal wooden supports are common. These webs alter the thermal behavior of the front surface and were added as a variation in FE models to induce defect-like thermal features. Their inclusion allowed for the investigation of the ability of the deep learning model to distinguish between actual defects and structural artifacts induced by such features.

In the transient thermal analysis, a heating-cooling process was employed, with a 5 min heating phase followed by a 5 min cooling

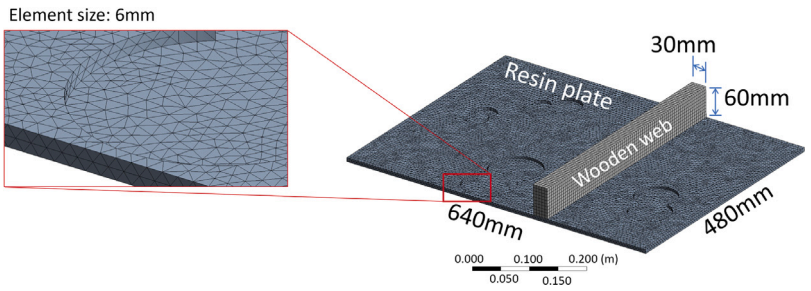

Fig. 5. Geometry and mesh of the FE model of a resin plate with irregular defects and a wood web.

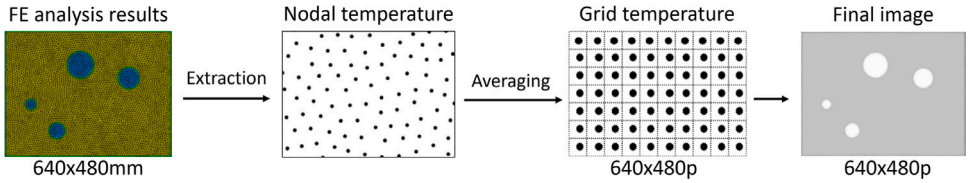

Fig. 6. Conversion from nodal temperature values to image pixel values.

Table 4
Physical and thermal properties of the epoxy resin material.

Property	Epoxy resin	Wood [50]
Density [kg/m^3]	960	160
Thermal conductivity [Wm^{-1}K^{-1}]	0.38	0.05
Specific heat [Jkg^{-1}K^{-1}]	1200	900

phase. The material properties of the epoxy resin plate were provided by the manufacturer, as listed in Table 4. Due to a relatively small increase in temperature (no more than 30 °C) observed during the thermography tests, the convection coefficient and emissivity were assumed constant for all surfaces [51]. According to heat transfer theory, the free convection coefficient for a vertical plate in air is typically in the range of 2–10 W/m^2K. Its exact value was determined by the model validation procedure described in detail in Section 2.2.3. Since white paint with an emissivity of 0.9 was applied to the plate front surface in the experiments, the same emissivity was applied in the FE models. The ambient temperature was set to 22 °C. The thermal model utilized SOLID291, a 3D 10-node tetrahedral thermal solid element. An element size of 6 mm was determined based on preliminary simulations with varying mesh resolution, with the aim of identifying the maximum size that ensured mesh convergence, in which case the simulation results were no longer dependent on the element size.

For each combination of defect shape, heat flux condition, and wood web inclusion, 40 simulations were performed, each with randomized defect depths, locations, and plate thicknesses. In total, 960 simulations were performed to build the simulation synthetic dataset, *Sim_R*, used to support the transfer learning framework.

2.2.2. Simulation data conversion

The principal output of the transient thermal analysis was the nodal temperature history. To mimic real thermal images, these nodal temperature values were converted to pixel-like grid values. More precisely, temperature data were extracted every 2 s, leading to 300 frames during the 10 min simulation. The extracted nodal temperature values were then transformed into a grid format representing image pixels, as illustrated in Fig. 6. This conversion process was designed to mimic the functionality of an IR sensor in a thermal camera, which accumulates the radiation for each pixel and converts it into a temperature value. This was achieved by first interpolating the nodal temperature data using a cubic function, and then the temperature value for each pixel was determined by averaging the function over each pixel.

After this conversion, random noise extracted from the experimental data, as described in Section 2.1.3, was added to each pixel in the simulation data. The final *Sim_R* dataset comprises 960 image series, each containing 300 frames. This format effectively replicates the structure of real thermal videos, allowing for the application of post-processing methods to time-series data.

2.2.3. Model validation

To ensure the effectiveness of the FE model, it must be validated within a parameter space defined by the experiments. In this study, the design parameters include the shape, depth, and location of the defects, as well as the applied heat flux. The FE model is validated by comparing its simulated results with the experimental measurements within the same parameter space. Once validated, additional variations of the design parameters are introduced to create a sufficient number of FE models across this space.

The comparison is carried out by investigating points at the same location on the plate surface, which are supposed to have similar temperature variation trends for simulated and experimental results. Therefore, 4 additional FE models with the same geometry as the experimental resin plates were developed for this purpose. To simplify the analysis, a uniform heat flux was applied to the front surface. The magnitude of the heat flux was determined as the average heat flux value generated by a single lamp, as mentioned in Section 2.1.1. Since the heat flux in the experiments was significantly non-uniform at the edge of the plate, two reference points were selected in the central region of the plate, as shown in Fig. 7.

The convection coefficient was determined by a curve-fitting process which compared the simulation and experimental temperature variation curves at the selected reference points on the front surface. The validation results will be discussed in Section 3.1.

2.3. Supervised learning for damage detection

This research used principal component thermography (PCT) to extract key features from thermal videos, which were then used in the training of the DL model. These PCT-derived features, being both more compact and informative-rich than thermal video data, served as efficient inputs for the DL model. Subsequently, a U-Net model was implemented to identify patterns indicative of subsurface damage. This approach combines advanced feature extraction techniques with a robust segmentation model to effectively analyze and interpret thermal imaging data for subsurface damage detection.

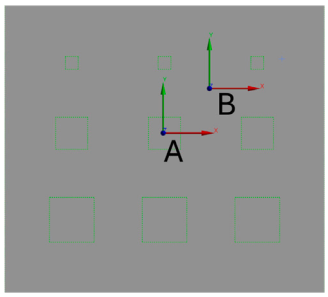

Fig. 7. Two reference points for model validation.

2.3.1. Principal component thermography

PCT is an advanced post-processing technique widely used in active thermography for NDT and defect detection in various materials [52, 53]. This method enhances defect visibility by applying principal component analysis (PCA) to a sequence of thermal images. PCT transforms the original IR data into a new set of orthogonal components, known as principal components (PCs), which are ordered such that the first few capture the most significant variations in the data, often corresponding to defects or anomalies in the inspected material.

In PCT, the dataset $X_{(N\times D)}$ is a collection of thermal images organized as a 2D matrix, where N is the number of thermal images and D is the number of pixels in each image. The elements of the matrix are the temperature values of each pixel. Thermal images are reshaped into 1D vectors and stored as the row vectors of X. PCA is then applied to X to obtain the PCs.

Given a D-dimensional dataset $X \in \mathbb{R}^D$ containing N data points, PCA aims at finding the PCs of X by looking for the eigenvectors of the covariance matrix $X^T X$. According to the theory of eigen-decomposition, a square matrix A can be decomposed as:

$$A = Q\Lambda Q^{-1} \tag{1}$$

where Q is a matrix whose columns are eigenvectors and Λ is a diagonal matrix containing eigenvalues. In PCA, given $A = X^T X$ as a symmetric matrix, Eq. (1) can be written as:

$$A = X^T X = W\Lambda W^T \tag{2}$$

where W is an orthonormal, satisfying $W^{-1} = W^T$. A typical method to obtain matrix W is singular value decomposition (SVD), which is defined as:

$$X_{(N\times D)} = U_{(N\times r)}\Sigma_{(r\times r)}V^T_{(r\times D)} \tag{3}$$

where X is the dataset, U and V are orthonormal matrices, Σ is a diagonal matrix, and r is the number of PCs to be retained. Substituting the SVD of X into Eq. (2) yields:

$$\begin{aligned} A &= (U\Sigma V^T)^T U\Sigma V^T \\ &= V\Sigma^2 V^T \end{aligned} \tag{4}$$

Comparing Eqs. (2) and (4), a correspondence between $W\Lambda W^T$ and $V\Sigma^2 V^T$ can be found, where W equals V, and Λ equals Σ^2. Therefore, the eigen-decomposition can be carried out by applying SVD, and the eigenvectors are expressed as the row vectors of $V^T_{(r\times D)}$, which represent the PCs. An illustration of the PCT procedure used in this work is shown in Fig. 8. Based on visual inspection of PC images, the first 9 PCs were used to build the training dataset in this research.

2.3.2. U-Net

The U-Net architecture consists of an encoder (contracting path) and a decoder (expanding path) working together to process images effectively. It takes a multi-channel image as input and outputs a binary segmentation map indicating the class (defective or non-defective) that each pixel belongs to. The main strength of the U-Net architecture lies in its ability to capture both context and precise localization through its symmetric design. The encoder reduces spatial dimensions while increasing feature information, while the decoder restores spatial details through up-sampling. The three key components in the U-Net architecture: encoder, decoder, and skip-connections, are shown in Fig. 9.

The encoder consists of 5 double-convolutional layers (3 × 3) followed by rectified linear unit (ReLU) activation functions that bring non-linearity into the model. Each double-convolutional layer employs two single-convolutional layers with 1 padding. The encoder implements max pooling (2 × 2) operations to reduce spatial resolution, allowing the network to capture increasingly complex features through multiple deeper layers. As depth increases, the spatial dimensions of the image decrease while the number of feature channels increases.

The decoder symmetrically mirrors the encoder and restores spatial dimensions through up-sampling operations. It combines high-level feature information with spatial details from earlier layers. This path works symmetrically with the contracting path, gradually reconstructing the spatial resolution while maintaining learned features.

One of the most critical features of U-Net is the skip-connections, which connect corresponding levels of the encoder and decoder. Each connection takes the output of each level in the encoder and attaches it to its corresponding decoder level. This mechanism allows the network to retain spatial context that might otherwise be lost due to down-sampling, thereby enabling precise localization in the final output. Additionally, skip-connections help mitigate the vanishing gradient problem during training.

The effectiveness of the U-Net has led to numerous variants and applications, making it a cornerstone model in computer vision tasks [54–56]. Its ability to work with relatively small training datasets while maintaining high accuracy has made it particularly valuable in NDT applications.

Table 5
Dataset distribution for training, validation and testing.

Dataset	Train.	Val.	Test.	Total images
Sim_R	672	144	144	960
Lab_C[a]	4		3	7
Lab_R[a]	20		8	28

[a] Cross-validation was implemented for experimental data.

2.3.3. Training

The experimental and simulation datasets were first processed with PCT. For each thermal video in the datasets, the first 9 PCs were extracted and stacked to form a 9-channel image, which was used as the input to the U-Net. All datasets used in this work, as well as the details of dataset splits, are summarized in Table 5.

Due to the very small size of the experimental dataset (only 28 images for resin plates and 7 images for the composite plate), Leave-One-Out Cross-Validation (LOOCV) was adopted in the training process, in which only one single image was left out for validation. This process was repeated by rotating the excluded image to identify the best hyperparameters, such as batch size, learning rate, and number of epochs (i.e., one complete pass of the entire training dataset through the model). Data augmentation techniques, including random cropping (350 × 350 pixels), random rotating (−90° to 90°), and random vertical flipping (probability of 0.5), were implemented to increase the variability of the training data. This is especially beneficial when dealing with small datasets.

For the simulation dataset, cross-validation was not implemented due to its large size. The U-Net model was trained multiple times on a fixed train–test split (the first row in Table 5) to find the optimal hyperparameter combination.

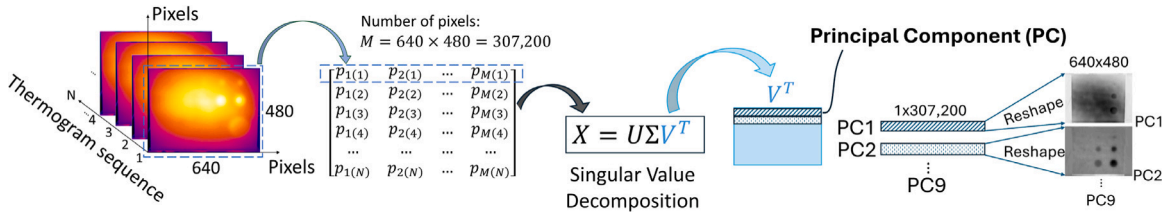

Fig. 8. Post-processing of thermal videos with PCT.

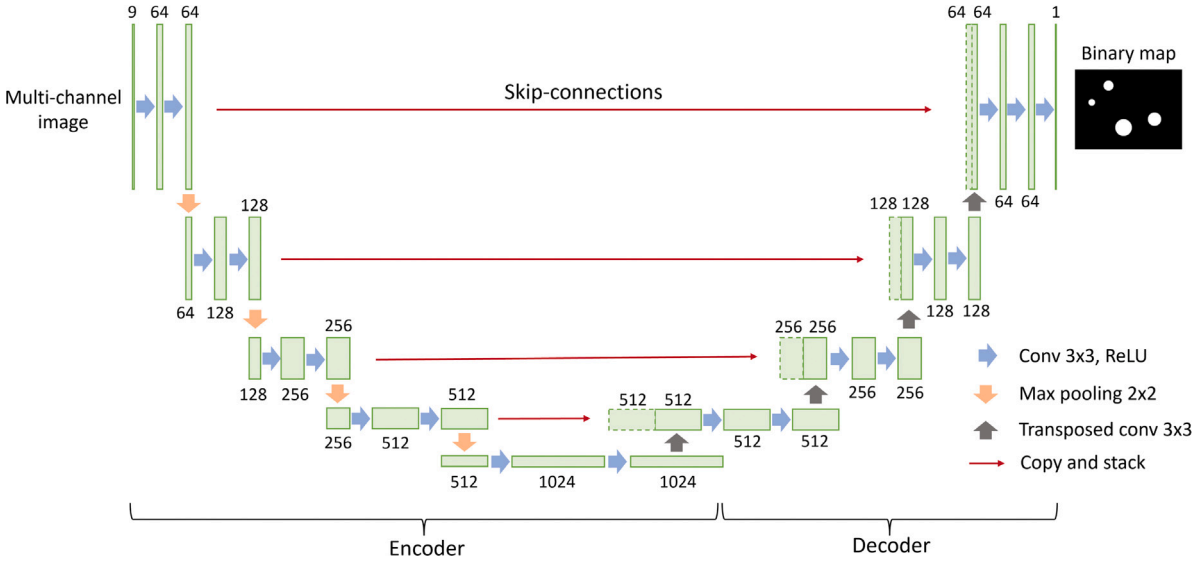

Fig. 9. U-Net architecture for image segmentation, adapted for multi-channel input images.

Supervised learning requires annotated data to provide ground truth labels. In image segmentation tasks, the data is annotated pixel-by-pixel, indicating which class each pixel belongs to. For the binary classification task in this work, each pixel was labeled as either 0 (non-defective) or 1 (defective).

All U-Net model training was implemented using PyTorch 2.0.1 and CUDA 11.7 on an NVIDIA RTX A1000 GPU.

2.3.4. Model performance evaluation

An effective model evaluation procedure requires the proper use of metrics that accurately represent the model performance. In image segmentation tasks, the most commonly used metric is the intersection over union (IoU), which evaluates the pixel-by-pixel overlap between the predicted segmentation mask and the ground truth. A score of 1 indicates perfect overlap, while 0 indicates no overlap. In addition to the IoU, the F1-score is also a useful metric for the evaluation of image segmentation models trained on imbalanced datasets, i.e. when one class contains significantly more pixels than the other. Both metrics depend on the calculation of true positives (TP), false positives (FP), and false negatives (FN) pixels. This work conducted a comprehensive model evaluation by employing all the metrics introduced below:

- TP: The number of pixels correctly classified as defective.
- FP: The number of pixels wrongly classified as defective.
- FN: The number of pixels wrongly classified as non-defective.
- Precision (P_r): The proportion of pixels predicted as defective that are actually defective, calculated as:

$$P_r = \frac{TP}{TP + FP} \tag{5}$$

- Recall (R_e): The proportion of actual defective pixels that are correctly predicted as defective, calculated as:

$$R_e = \frac{TP}{TP + FN} \tag{6}$$

- F1-score: The harmonic mean of Precision and Recall, providing a balanced measure of model performance, calculated as:

$$F_1 = \frac{2 \times P_r \times R_e}{P_r + R_e} = \frac{2TP}{2TP + FN + FP} \tag{7}$$

- Intersection over union (IoU): Pixel-by-pixel overlap between the predicted segmentation mask and the ground truth, calculated as:

$$IoU = \frac{Area\ of\ intersection}{Area\ of\ union} = \frac{TP}{TP + FN + FP} \tag{8}$$

2.4. Transfer learning

Transfer learning is a powerful machine learning technique that enables models to leverage knowledge gained from one task to enhance performance on different but related tasks. This approach mirrors human learning, where previous experience accelerates the acquisition of new, related skills. In transfer learning, knowledge from a source domain is systematically transferred to a target domain to enhance learning. A domain $\mathcal{D}$ encompasses a feature space $\mathcal{X}$ coupled with a marginal probability distribution $P(X)$, where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. The corresponding task $\mathcal{T}$ is defined by a label space $\mathcal{Y}$ and a predictive function $f : \mathcal{X} \rightarrow \mathcal{Y}$. The transfer learning architecture consists of two essential components. The source component includes domain $\mathcal{D}_S = \{\mathcal{X}_S, P(X_S)\}$ and task $\mathcal{T}_S = \{\mathcal{Y}_S, f_S(x)\}$, while the target component comprises domain $\mathcal{D}_T = \{\mathcal{X}_T, P(X_T)\}$ and task $\mathcal{T}_T = \{\mathcal{Y}_T, f_T(x)\}$, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$. The goal is to improve the target predictive function f_T in $\mathcal{D}_T$ using knowledge from $\mathcal{D}_S$ and $\mathcal{T}_S$ [57].

In practice, transfer learning is implemented by reusing a pre-trained model as the starting point for a new learning task. This significantly reduces the amount of data and computational resources

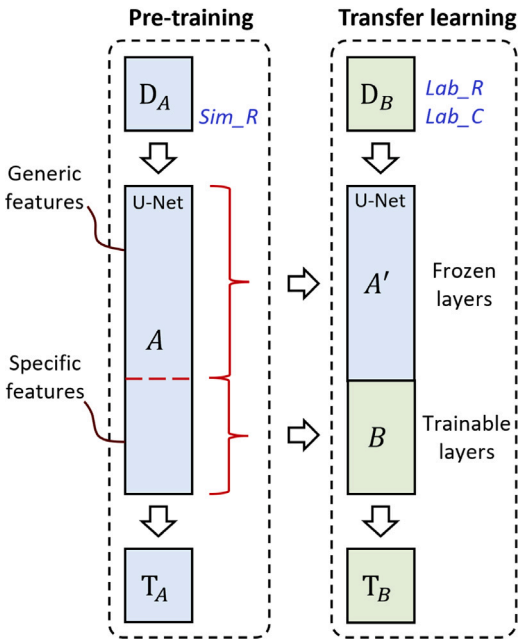

Fig. 10. Schematic explanation of transfer learning.

required for training, making it particularly valuable in deep learning applications where large datasets and extensive computational resources are not available. As shown in Fig. 10, the U-Net model A described in Section 2.3.2 was first trained and validated with the simulation dataset D_A for task T_A, i.e. damage detection in resin plates based on simulation IR data. This pre-training step enabled the U-Net model to learn useful features from simulation data. The pre-trained model was then divided into two parts: A' and B. A' contained the shallow layers which captured some generic features common to both simulated and experimental thermal images. The parameters in A' were frozen and did not update during the re-training. B included the final deep layers that contained more specific features of the simulated thermal images. These layers were set as trainable, so that the model was able to learn from experimental thermal images by updating the parameters during re-training.

3. Results and discussion

3.1. Finite element model validation and noise analysis

A representative example of the experimental data is shown in Fig. 11. This setup refers to the heating condition HD7, where the composite plate was heated from room temperature using the 3 lamps. Two reference points were identified for comparison: P_1 located in a defective region and P_2 in a non-defective region, as shown in Fig. 11(a). The temporal evolution of the surface temperature exhibited a two-stage pattern, as depicted in Fig. 11(b). The data reveal that while both defective and non-defective regions demonstrated similar temperature trends, the defective area heated more rapidly due to its reduced material thickness.

The experimental results for the resin plates were utilized to validate the FE model. The validation results for the plate with square defects are shown in Fig. 12. Two image frames, one from the numerical simulation and one from the experimental data, were extracted at 340 s. In the experimental results, the temperature distribution in the central and upper regions of the plate showed notable uniformity, despite the non-uniform heating from the three separate lamps. Therefore, reference points A, within a defective area, and B, in a non-defective area, were selected from these regions. By fitting the simulated temperature curve with the experimental temperature curve for the reference points, the mean value of the convection coefficient in the simulation was determined as 6 W/m^2K. The temperature profiles at the reference points show close agreement between experimental and simulation results, indicating a robust model design and effective validation. This validation was also performed for the other three plates. The use of a uniform heat flux in the simulations allowed accurate reproduction of the experimentally observed thermal behavior in the central area of the plates, while preserving computational efficiency.

Furthermore, the thermal measurements reveal notable noise components, as shown in Figs. 12(b) and 12(c) after zooming in, which constitute the primary distinction between the experimental measurements and the FE simulation results. Therefore, the noise in the experimental data was analyzed using a high-pass filter as described in Section 2.1.3.

Fig. 13(a) shows an example of the noise extracted from the original temperature measurements by applying a high-pass filter with a cut-off frequency of 0.1 Hz. Under the assumption that the thermal noise exhibited a Gaussian distribution, the statistical analysis shows that the noise is characterized by a mean of 0 and a standard deviation of 0.0325. Based on this, synthetic temporal noise was generated following the distribution $N(0, 0.0325^2)$, as shown in Fig. 13(b). This synthetic noise pattern was manually added to the Sim_R dataset to make the simulation results more representative of the real thermal images, as plotted in Fig. 13(c). The result shows that after adding synthetic noise, the simulated signal presents significant similarity in noise patterns compared to the experimental signal plotted in Fig. 12. The clean simulated data and the noisy simulated data were both processed using PCT in Section 3.2 to gain more insight at feature level.

3.2. Feature extraction

The PCT technique outlined in Section 2.3.1 was implemented to analyze thermal video data from both experimental tests and numerical simulations. Fig. 14 displays the first 9 PCs extracted from thermal videos of datasets containing square-shaped defects. The first column presents the PCs derived from the Lab_R experimental dataset, while the second and third columns correspond to the FE model of the resin plate under uniform and non-uniform heat flux, respectively. The fourth

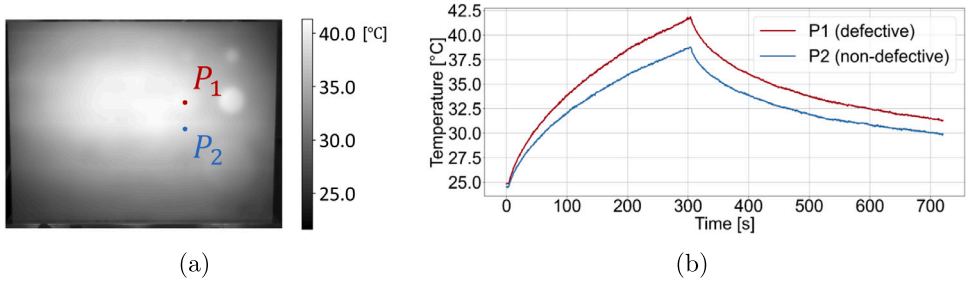

Fig. 11. (a) Thermal image of the composite plate at the end of heating (300 s). (b) Temperature history curves at points P_1 (defective region) and P_2 (non-defective region).

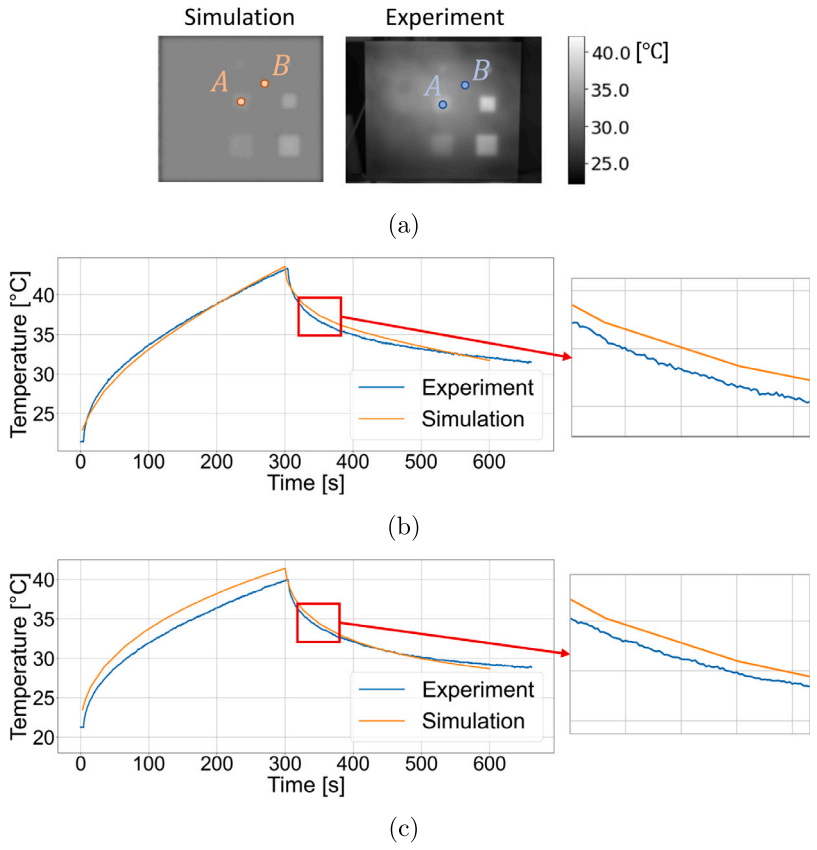

Fig. 12. Validation results of the FE model through curve fitting. (a) Reference points A and B selected from experimental (left) and simulation (right) results. (b) Temperature curve of point A. (c) Temperature curve of point B.

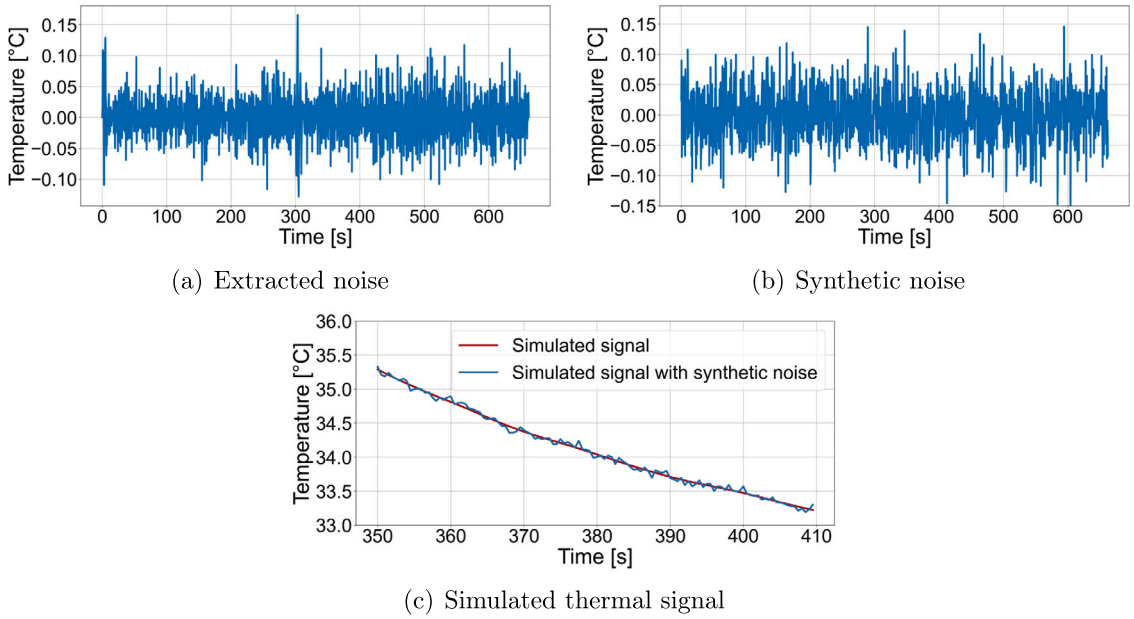

Fig. 13. (a) Noise extracted from experimental data using a high-pass filter. (b) Synthetic Gaussian noise. (c) Comparison between the original simulated signal and the simulated signal after adding noise.

column shows the PCs of the simulation with synthetic noise under uniform heat flux.

High similarity is observed among the four sets of PCs, particularly from PC2 to PC7. In contrast, PC1 exhibits significant differences across the different conditions. The PC1s in columns (a) and (c) reveal distinct non-uniform heating patterns, reflecting real experimental heating and manually applied Gaussian distributed heat flux in simulation, respectively, whereas the PC1s from the uniform heating simulation display a

much more uniform thermal distribution. This discrepancy is confined to PC1, indicating that heating information is primarily captured in the first component. These results support the rationale for using a uniform heating flux in the simulations, as the effect of heating distribution can be effectively mitigated by excluding PC1 during the analysis.

Additional differences are visible in the final two PCs. In the noise-free simulation (second column), PC8 and PC9 clearly reveal defect-related information. However, this information is significantly masked in the experimental results and in the simulated results with added synthetic noise. This demonstrates the impact of measurement noise on the detectability of the defect.

This comparison shows that the noise in the thermal images reduces the defect-related features that can be extracted by PCT. By adding synthetic noise, the simulation data were made more comparable to the real IR data, which in turn improved the effectiveness of transfer learning, as more common features were introduced between the two domains. Therefore, only the noisy simulation dataset was utilized in the following steps of the transfer learning framework. Taking into account the different features observed in PC1, the U-Net models were trained with two combinations of PCs: PC1-9 and PC2-9. This approach allowed for the investigation of the impact of including or excluding PC1 on the model performance. The deep learning experiments carried out in this study are listed in Table 6. The objectives and results of these experiments are discussed in the following sections.

3.3. Benchmark model and pre-training

Benchmark models for evaluating transfer learning were developed by training the U-Net architectures using *Lab_R* and *Lab_C*. Expt.1-4 in Table 6 outline the training strategies for different benchmark models trained on different datasets and PC combinations. Pre-trained models were obtained by training the U-Net with the same simulation dataset but different combinations of PCs (Expt.5 and 6). This study assumes that simulated and real thermal images share fundamental features, making transfer learning a viable approach. To test this hypothesis, the pre-trained models were evaluated using the *Lab_R* (Test 2 in Expt.5 and 6). Moreover, to provide a comparison with more advanced image segmentation models with more complex architectures, SegNet [58] (Expt.10 and 11) and DeepLabv3+ [59] (Expt.12 and 13) were trained on both *Lab_R* and *Lab_C* using the same data splits as the U-Net.

The U-Net models were initially trained with PC1-9. The training progress, illustrated in Fig. 15, reveals well-behaved loss curves, indicating successful convergence without signs of overfitting. The prediction instances are visualized in Fig. 16, where PC2 is plotted as an indication of the input image. It can be seen that PM1 effectively classifies defective pixels, regardless of the interference from the wood web (Fig. 16(a)). BM_C1 and BM_R1 work well in detecting shallow defects, but the information of deeper defects ($\leq$ 15 mm) is hard to recognize (Figs. 16(b) and 16(c)). The model performance is quantitatively evaluated in Table 7. When using PC1-9 (Expt.1, 2 and 5), the pre-trained models demonstrated superior performance with excellent F1-score and IoU metrics, while the benchmark models showed moderate performance. This is because PM1 and PM2 were trained and tested exclusively on simulation data, which, despite the added noise, was relatively cleaner compared to the real thermal images. As a result, the U-Net model in the simulation data successfully captured the defects.

However, when tested on *Lab_R*, PM1 failed to detect defects effectively, as shown in the third column of Fig. 17. This performance gap can be attributed to the discrepancy in PC1 between the simulation and the experimental data. As discussed in Section 3.2, PC1 captures information about the heating conditions. Therefore, when trained exclusively with *Sim_R*, the model developed a bias toward uniform heating patterns, which hindered its ability to process real thermal images with non-uniform heating distributions.

To address this limitation, a new set of models was developed by removing PC1 and using only PC2-9 (Expt.3, 4 and 6). Although this

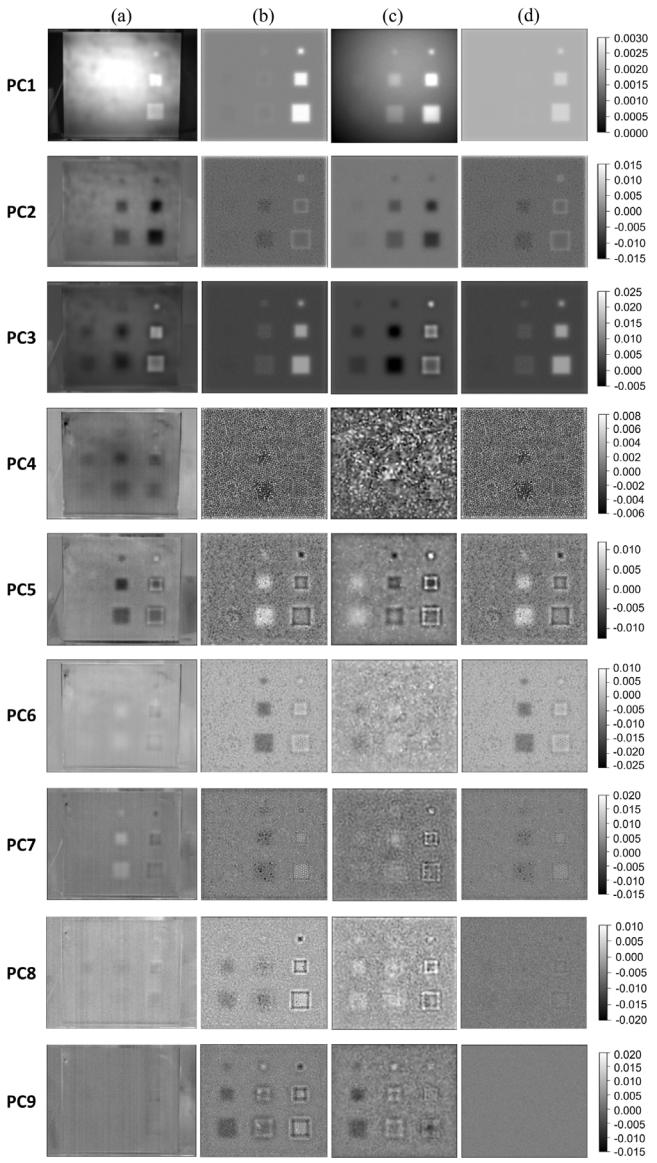

Fig. 14. First 9 principal components of the resin plate with square defects extracted from: (a) experimental dataset (*Lab_R*), (b) simulation dataset without noise, (c) simulation dataset with Gaussian distributed heat flux without noise, and (d) simulation dataset with synthetic noise.

modification resulted in a slight decrease in overall performance metrics (as shown in Table 7), due to the reduction of effective information, the new pre-trained model showed a marked improvement when tested on *Lab_R*. The fourth column of Fig. 17 shows that PM2 successfully identified several defective pixels in real thermal images. However, it is important to note that the performance of PM2 is not directly comparable to BM_R, as PM2 was trained without any exposure to experimental data.

3.4. Transfer learning

The pre-trained model PM2 was re-trained on *Lab_R* and *Lab_C* datasets following the steps explained in Section 2.4. Three re-training processes were performed using different training datasets, as summarized in Table 6: RM_R, trained only on *Lab_R* (Expt.7), RM_C trained only *Lab_C* (Expt.8) and RM_R_C trained on both *Lab_R* and *Lab_C* (Expt.9). The third process involved training RM_R on *Lab_C*. A

Table 6
Design of deep learning experiments.

Objective	Expt.	PCs	Model[a]		Dataset	Train/Val[b]	Test 1	Test 2[c]
Benchmark training	1	1–9	BM_C1		Lab_C	4	3	–
	2	1–9	BM_R1		Lab_R	20	8	–
	3	2–9	BM_C2		Lab_C	4	3	–
	4	2–9	BM_R2		Lab_R	20	8	–
Pre-training	5	1–9	PM1		Sim_R	672/144	144	8 (Lab_R)
	6	2–9	PM2		Sim_R	672/144	144	8 (Lab_R)
			Pre-trained	**Re-trained**				
Transfer learning	7	2–9	PM2	RM_R	Lab_R	20	8	–
	8	2–9	PM2	RM_C	Lab_C	4	3	–
	9	2–9	RM_R	RM_R_C	Lab_C	4	3	–
SegNet	10	2–9	SegNet_R		Lab_R	20	8	–
	11	2–9	SegNet_C		Lab_C	4	3	–
DeepLabv3+	12	2–9	DeepLab_R		Lab_R	20	8	–
	13	2–9	DeepLab_C		Lab_C	4	3	–

[a] BM: Benchmark model, PM: Pre-trained model, RM: Re-trained model, C: Composite, R: Resin.
[b] Cross-validation was implemented for experimental data.
[c] Pre-trained models were tested on experimental data without re-training.

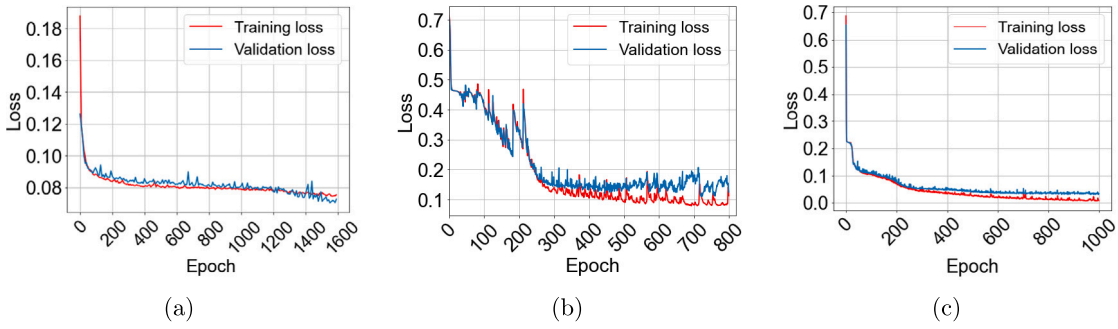

Fig. 15. Training and validation loss curves of (a) PM1, (b) BM_C1, and (c) BM_R1.

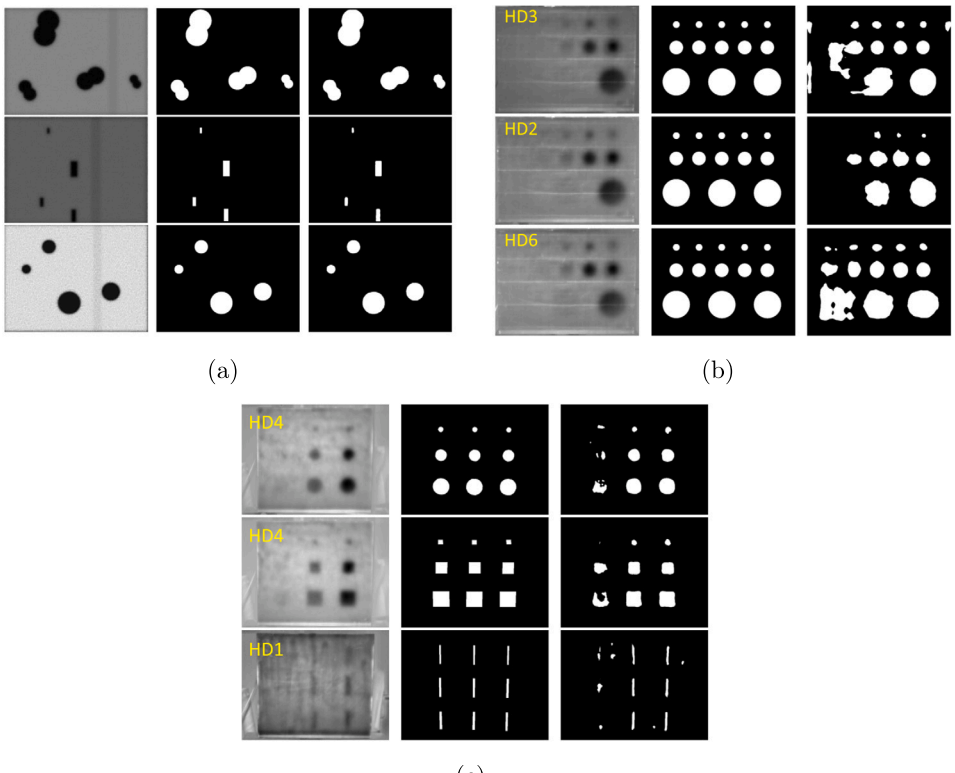

Fig. 16. Visualizations of PC2 as an indication of input images (left column), ground truth (middle column) and predictions (right column) for (a) PM1, (b) BM_C1 and (c) BM_R1.

Table 7
Results for the benchmark models and pre-trained models.

Expt.	Model	Epochs	Precision	Recall	F1-score	IoU
1	BM_C1	800	0.794	0.736	0.764	0.618
2	BM_R1	1000	0.874	0.795	0.866	0.713
3	BM_C2	800	0.781	0.743	0.762	0.615
4	BM_R2	1000	0.877	0.786	0.829	0.708
5[a]	PM1	1600	0.928/0.152	0.901/0.103	0.914/0.128	0.842/0.065
6[a]	PM2	1600	0.931/0.524	0.900/0.396	0.915/0.451	0.843/0.291

[a] Two values for each metric are exhibited as Test 1/Test 2.

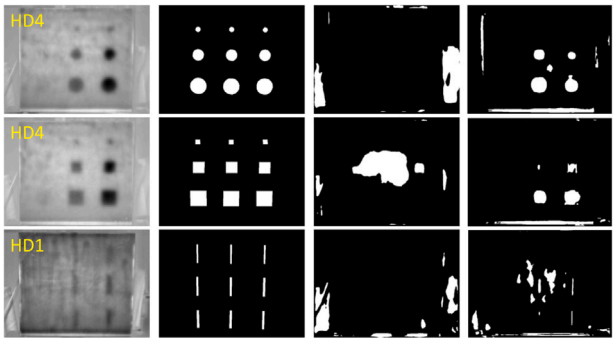

Fig. 17. The second principal component (1st column), ground truth (2nd column), and prediction results for PM1 (3rd column) and PM2 (4th column) tested on *Lab_R* dataset.

parametric analysis was performed by varying the number of trainable layers to examine its effect on the re-training results. The IoU scores of the models re-trained with different numbers of trainable layers are shown in Fig. 18(a). The results show that when a small number of layers are trainable, the re-trained models exhibit poor performance, which leads to underfitting resulting from a very limited number of trainable parameters. As the number of trainable layers increases, the three models reach and in some cases slightly outperform their benchmark models (BM_C2 and BM_R2).

An additional analysis was performed to better understand the benefits conferred by transfer learning. Specifically, PM2 was fine-tuned using varying amounts of training data from *Lab_R* and *Lab_C*, progressively reducing the dataset to identify the minimum size required to match the performance of BM_R2 and BM_C2. Fig. 18(b) presents the IoU values for re-trained models across varying numbers of training data points. For both resin and composites, model performance declined with reduced training data. Benchmark performance was attained with 15 data points for resin and 2 for composites. The proposed transfer learning framework required fewer data points to achieve equivalent performance, effectively addressing data scarcity.

The best performance of the three re-trained models is summarized in Table 8. Higher F1-score and IoU were achieved after re-training, compared with the benchmark performance shown in Table 7. Compared with Expt.3, the results of Expt.8 showed an improvement of 3.2% on F1-score and 4.4% on IoU for the composite plate. An improvement of 2.6% on F1-score and 3.9% on IoU was observed for epoxy resin plates comparing the results of Expt.4 and Expt.7. Although RM_R_C was re-trained on actual infrared images of resin plates, its ability in detecting damage in composite materials was comparable to that of RM_C. This suggests that the knowledge transferred from RM_R had only a minimal impact on the performance of RM_R_C. This limited contribution can be attributed to the greater dissimilarity between the experimental data of resin and composite materials, compared to the difference between simulated and experimental resin data. Among the evaluated architectures, the re-trained U-Net achieved a higher F1-score and IoU than SegNet. While DeepLabv3+ performed slightly better on the resin dataset, the re-trained U-Net outperformed it on the composite dataset.

Besides the observation in the final performance, a significant reduction on training epochs was achieved during the re-training process. This faster convergence indicates that the generic features learned from the simulation data can be effectively transferred to the deep learning model by training only a subset of the layers, rather than the entire model. The computational costs of the models trained on the composite dataset are reported, as shown in Table 9, in terms of total floating-point operations (FLOPs) per inference, the number of trainable parameters, the training time per epoch, and the inference latency. Compared to SegNet and DeepLabv3+, the U-Net architectures have a lower number of trainable parameters, especially for the re-trained U-Net. A significant reduction of training time per epoch is achieved thanks to fewer trainable parameters.

Figs. 19(a) and 19(b) visualize the prediction results for RM_R and RM_C, respectively. The last columns represent the results after re-training. Compared to SegNet and DeepLabv3+, the re-trained models demonstrate slightly improved performance in detecting internal

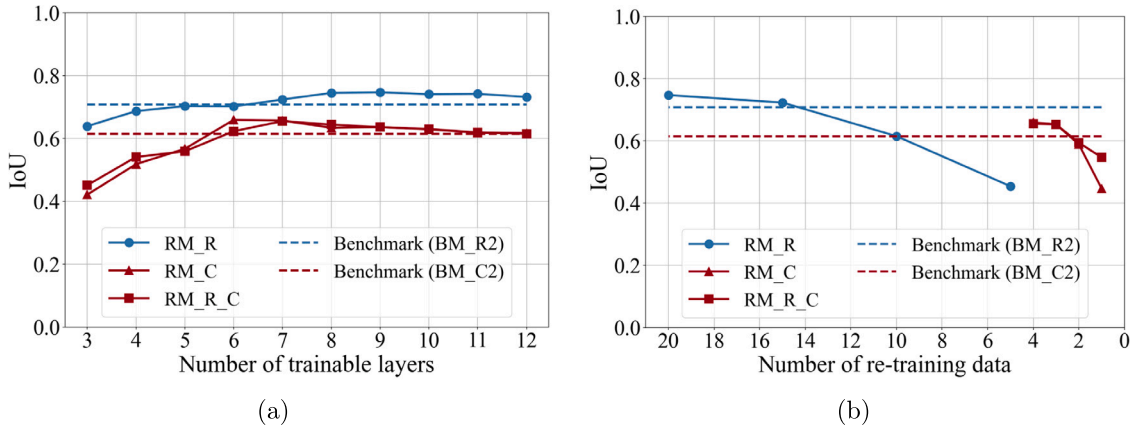

Fig. 18. Re-trained model performance with varying number of (a) trainable layers and (b) re-training data.

Table 8
Performance of the U-Net after applying transfer learning using different training datasets, and the testing results of SegNet and DeepLabv3+ on experimental datasets.

Expt.	Model	Trainable layers	Epochs	Precision	Recall	F1-score	IoU
7	RM_R	9	150	0.894	0.819	0.855	0.747
8	RM_C	6	100	0.801	0.788	0.794	0.659
9	RM_R_C	6	100	0.799	0.784	0.791	0.655
10	SegNet_R	–	1000	0.867	0.816	0.841	0.725
11	SegNet_C	–	750	0.779	0.649	0.708	0.548
12	DeepLab_R	–	1200	0.904	0.821	0.861	0.755
13	DeepLab_C	–	1000	0.747	0.787	0.766	0.621

Table 9
Computational cost metrics for the U-Net, SegNet and DeepLabv3+.

Model	FLOPs [–]	n_parameters [–]	Training time/epoch [s]	Inference latency [s]
BM_C2	1.96×10^{11}	7.7×10^{6}	8.71	1.548
RM_C	1.96×10^{11}	2.3×10^{6}	2.45	1.551
SegNet_C	1.89×10^{11}	2.9×10^{7}	10.98	1.783
DeepLab_C	2.58×10^{11}	6.1×10^{7}	20.12	2.145

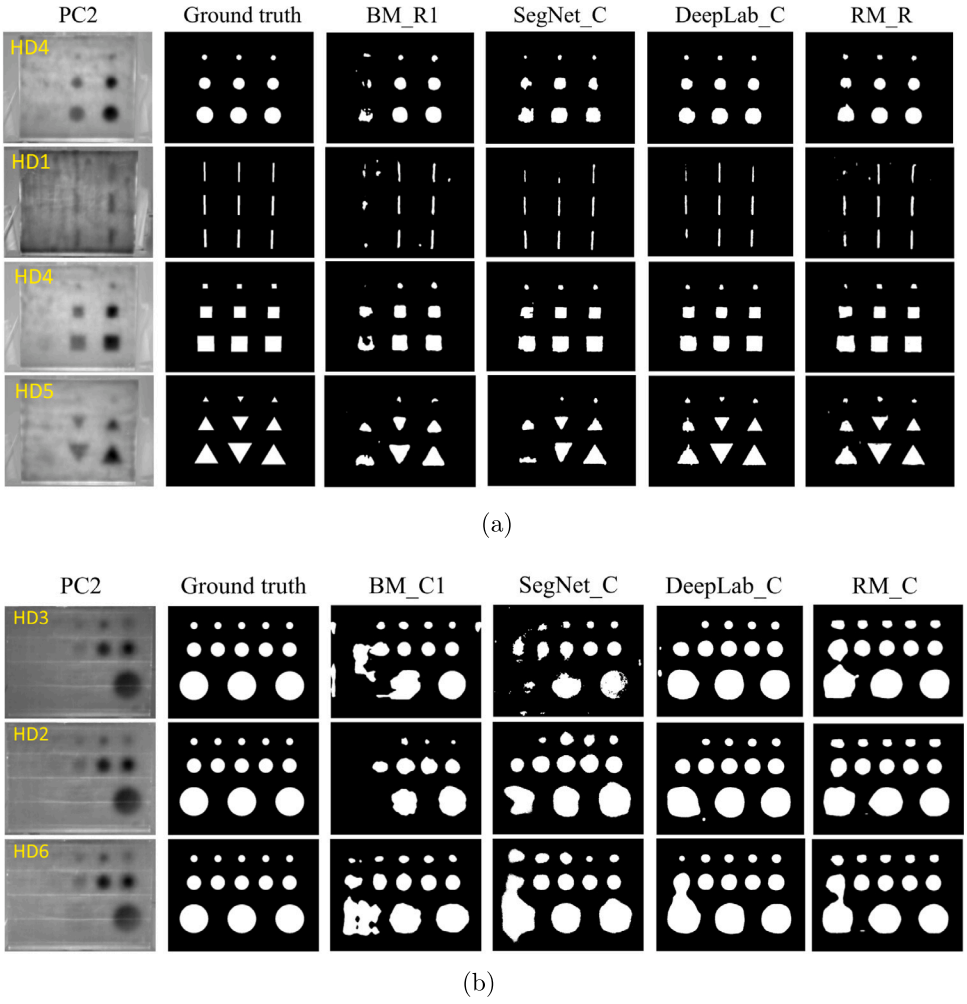

Fig. 19. Prediction results for (a) RM_R and (b) RM_C on the testing samples. Comparison is made with benchmark U-Net (BM_R1 and BM_C1), SegNet and DeepLabv3+.

defects in both resin and GFRP plates. Compared with benchmark U-Net models, a significant improvement is observed in the detection of deeper defects located on the left side of the plates. The re-trained models outperform the benchmarks because the pre-trained model provides stronger general-purpose feature representations, which reduces overfitting and accelerates convergence on limited experimental data. The re-training process starts from a better initialization of the training parameters in the pre-trained model that already encodes reusable information from the simulated data (e.g., edges, textures, thermal-related characteristics), which the re-training step then specializes for defect detection in real thermal images. This advantage becomes even more pronounced when the amount of training data is limited, as in

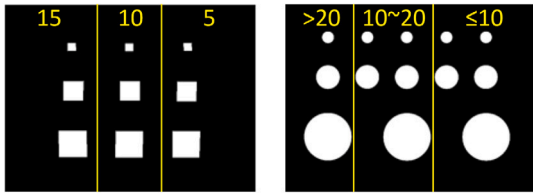

Fig. 20. Depth-based defect classes definition in (a) *Lab_R* dataset and (b) *Lab_C* dataset (Unit: mm).

Table 10
Performance metrics comparison across different defect depths for *Lab_R* and *Lab_C* datasets.

Defect depths	Lab_R			Lab_C		
	15 mm	10 mm	5 mm	> 20 mm	10–20 mm	≤ 10 mm
F1-score (benchmark)	0.694	0.913	0.934	0.582	0.818	0.909
F1-score (re-trained)	**0.844**	0.914	0.937	**0.785**	0.879	0.910
IoU (benchmark)	0.532	0.841	0.876	0.410	0.692	0.833
IoU (re-trained)	**0.730**	0.842	0.882	**0.646**	0.783	0.835

the cases of BM_C2 and RM_C. The U-Net trained exclusively on the small *Lab_C* dataset does not achieve optimal performance. However, by incorporating knowledge transferred from the simulated data, this shortcoming is effectively mitigated. This improvement is particularly significant, indicating the increased capability of the re-trained models to detect and characterize structural anomalies at greater depths.

Since the models show varying performance across different defect depths, a depth-based evaluation was performed in addition to the overall performance assessment. For *Lab_R*, the images were divided into 3 sections, each containing one of the three defect columns shown in Fig. 20(a), which contains defects at depths of 15 mm, 10 mm, or 5 mm, respectively. Performance metrics were calculated separately within each section. For *Lab_C*, three levels of defect depth were defined: $d \leq 10$ mm, 10 mm $< d \leq 20$ mm, 20 mm $< d \leq 30$ mm (Fig. 20(b)). The performance of the model was evaluated separately for each depth level. As shown in Table 10, the transfer learning process predominantly improved the detection of deeper defects. For 15 mm-deep defects in the resin plates, the F1-score increased from 0.694 to 0.844 and the IoU from 0.532 to 0.730. For the defects deeper than 20 mm in the composite plate, the F1-score increased from 0.582 to 0.785, while the IoU improved from 0.410 to 0.646. The metrics for the shallow defects showed only minimal improvement after re-training.

The results of transfer learning experiments demonstrate that the proposed approach enables the pre-trained model PM2 to learn from real IR data in a very efficient way. In addition, the pre-learned knowledge of a more diverse simulation dataset makes the re-trained models able to achieve a higher F1-score and IoU values compared to the benchmark. These improvements are mainly attributed to the improved ability of re-trained models to detect deeper defects ($\geq$ 15 mm).

3.5. Model generalization and uncertainty evaluation

The generalization capability of the developed models was assessed using the LOOCV approach during training. Each cross-validation experiment involved multiple training runs on different subsets of the data, with performance evaluated on the corresponding validation sets. The IoU was chosen as the primary metric, and its mean, standard deviation, and coefficient of variation were calculated across all validation folds. The results of this evaluation are presented in Table 11. For both resin and composite datasets, the retrained U-Net models achieved low coefficients of variation, all below 0.1. This indicates that the proposed transfer learning framework successfully enhanced the generalization of the U-Net architecture in defect detection from thermal images. In contrast, SegNet and DeepLabv3+ exhibited higher coefficients of variation, which can be attributed to the limited size of the training data. The greater complexity of their architectures also requires more training data to achieve better performance compared with U-Net. Another notable observation is that the models trained on resin data consistently exhibit slightly lower coefficients of variation compared to those trained on composite data. This difference can be attributed to the variation in dataset size available for resin and composite materials.

Uncertainty evaluation is essential in deep learning research as it provides a measure of the reliability of model predictions. To assess the uncertainty of the re-trained U-Net model RM_C, given the ultimate goal of applying the framework to defect detection in composites, a bootstrap-based confidence interval (CI) analysis was conducted. A total of 500 bootstrap resamples were generated with replacement from the testing set of *Lab_C*. For each defect depth level, 10 samples were randomly cropped and resized from the 3 available multi-channel images to form the testing set. The RM_C model was evaluated on each of the 500 resamples, resulting in 500 IoU scores, and 95% CIs of the IoUs were calculated to quantify uncertainty across different defect depth levels. The results are shown in Fig. 21, with mean IoU values of 0.632, 0.77 and 0.841 for defects depth > 20 mm, 10–20 mm, and $\leq$ 10 mm, respectively. The uncertainty of RM_C was further assessed by calculating the relative half-width (RHW) as:

$$RHW = \frac{CI_{up} - CI_{low}}{mIoU} \times 100\% \tag{9}$$

where CI_{up} and CI_{low} are the upper and lower bounds of the CI, and $mIoU$ is the mean IoU. Across the three defect depth levels, RHW values of 5.2%, 3.7%, and 3.3% were obtained, demonstrating low model uncertainty and robust performance gains for the composites, consistent with the results reported in Table 10.

4. Conclusions and recommendations

This study addresses the challenge of dataset scarcity in AI-aided damage detection for thick composites using IR thermography by developing a comprehensive simulation dataset integrated with a transfer learning framework. Step-heating thermography experiments were conducted on GFRP polymer and epoxy resin plates. In addition, FE models were validated against experimental results, confirming their accuracy in capturing the thermal behavior during the heating-cooling process. To support the transfer learning approach, the simulation dataset was expanded to incorporate various defect configurations. PCT was applied to extract the first 9 PCs from both experimental and simulation thermal videos. A noise analysis revealed that the defect-related information in the deeper PCs (PC8 and PC9) was largely obscured by experimental noise. To address this, synthetic noise was added to the simulation data to enhance comparability with real IR data.

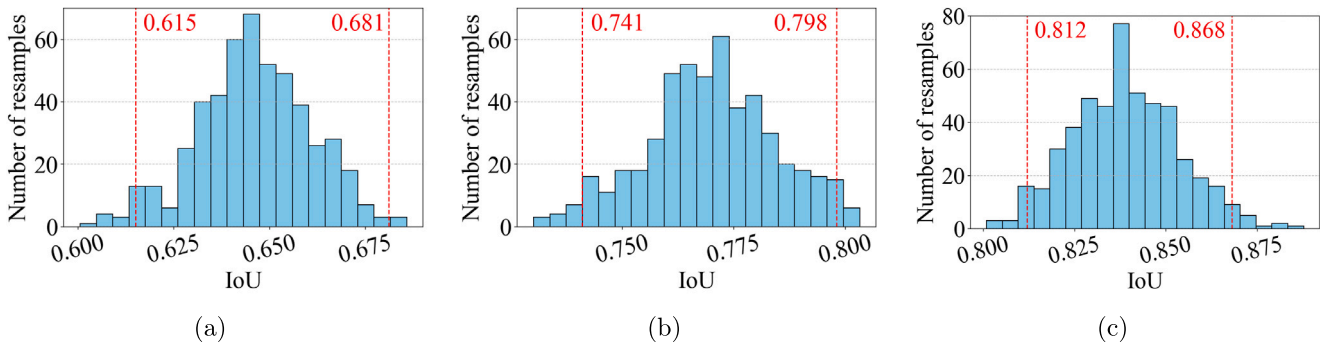

Fig. 21. Histograms of IoU scores for bootstrap-based CI estimation. Analysis was conducted for defect depth level of (a) > 20 mm, (b) 10–20 mm, and (c) ≤ 10 mm. The CIs are highlighted in red.

Table 11
Statistics of the IoU for the LOOCV of benchmark U-Net (BM_R2 and BM_C2), re-trained U-Net (RM_R and RM_C), SegNet and DeepLabv3+.

Model	Mean [μ]	Standard deviation [σ]	Coefficient of variation [σ/μ]
BM_R2	0.686	0.076	0.111
BM_C2	0.601	0.101	0.168
RM_R	0.739	0.064	0.087
RM_C	0.662	0.060	0.091
SegNet_R	0.712	0.088	0.124
SegNet_C	0.659	0.086	0.131
DeepLab_R	0.731	0.076	0.104
DeepLab_C	0.673	0.078	0.116

Initial results showed that U-Net models pre-trained only on simulated data performed poorly when tested on experimental data, primarily due to discrepancies in PC1, which captured heating condition differences. Excluding PC1 from the input significantly improved performance, confirming the hypothesis that fundamental features are shared between simulation and experimental data, and validating the feasibility of the proposed transfer learning approach.

Re-training the pre-trained model with experimental data, while updating only selected layers, resulted in faster convergence and enhanced performance metrics. Overall, the re-trained U-Net achieved higher performance and better generalization compared with SegNet and DeepLabv3+. Compared to the benchmark U-Net, the model re-trained on experimental data demonstrated higher ability in detecting subsurface defects, due to the preserved feature representations in the pre-trained model. The transfer learning approach led to improvements of 2.6% (on F1-score) and 3.9% (on IoU) for epoxy resin plates, and 3.2% (on F1-score) and 4.4% (on IoU) for the composite plate. In particular, the most significant improvements were observed in the detection of deeper defects located more than 15 mm beneath the surface. The U-Net re-trained on the composite dataset shows low uncertainty, as demonstrated by a bootstrap-based confidence interval analysis across defect depth levels, thereby confirming the robustness of the model improvements.

This research demonstrates that transfer learning provides a reliable and scalable solution to enhance the detection of deeper subsurface defects in thick composites. The U-Net architecture pre-trained on simulation data proved capable of capturing common features in thermal images used for NDT, making it adaptable to various real-world experimental datasets. While the proposed transfer learning framework demonstrates promising performance on experimental thermography data acquired under controlled laboratory conditions, its validity for industrial composite components remains to be established. Future research should focus on extending the approach to practical inspection scenarios through the incorporation of representative datasets acquired from industrial composite components with diverse materials and geometries, thus strengthening its robustness and generalization ability.

Moreover, other possible directions for future work are highlighted after understanding the limitations of this study. The simulation dataset could be improved by implementing non-uniform heat flux in FE models to better replicate actual testing conditions. Furthermore, rather than adding noise to the simulation data, applying advanced noise reduction techniques to the experimental data could help preserve defect-related information in deeper PCs. Finally, exploring alternative advanced deep learning architectures for image segmentation within the proposed transfer learning framework could offer valuable comparative insights into the relative performance of different models.

CRediT authorship contribution statement

Muyao Li: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Davide Leonetti:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Donatella Zappalá:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **H.H. (Bert) Snijder:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

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.

Acknowledgments

The authors would like to thank Dr. Roger Groves, Dr. Nan Tao and Dr. Yanan Zhang for their support in our experimental work in the TU Delft Aerospace Structures and Materials Laboratory. This research is part of the Holi-DOCTOR project, which is funded by the Dutch Research Council (NWO).

Data availability

Significant research data and algorithm have been shared through a link to Mendeley data repository https://doi.org/10.17632/xjf5cmkjm3.1.

References

[1] Ciang C, Lee J, Bang H. Structural health monitoring for a wind turbine system: a review of damage detection methods. Meas Sci Technol 2008;19(12):122001.
[2] Diamanti K, Soutis C. Structural health monitoring techniques for aircraft composite structures. Prog Aerosp Sci 2010;46(8):342–52.
[3] Maljaars J, Leonetti D, Hashemi B, (Bert) Snijder HH. Systematic derivation of safety factors for the fatigue design of steel bridges. Struct Saf 2022;97:102229.
[4] Wang B, Zhong S, Lee T, Fancey KS, Mi J. Non-destructive testing and evaluation of composite materials/structures: A state-of-the-art review. Adv Mech Eng 2020;12(4):1687814020913761.
[5] Liu J, Liu L, Wang Y. Experimental study on active infrared thermography as a NDI tool for carbon–carbon composites. Compos Part B: Eng 2013;45(1):138–47.
[6] Li Y, Yang Z, Zhu J, Ming A, Zhang W, Zhang J. Investigation on the damage evolution in the impacted composite material based on active infrared thermography. NDT E Int 2016;83:114–22.
[7] Ball RJ, Almond DP. The detection and measurement of impact damage in thick carbon fibre reinforced laminates by transient thermography. NDT & E Int 1998;31(3):165–73.
[8] Ibrahim M. Non-destructive evaluation of thick-section composites and sandwich structures: A review. Compos Part A: Appl Sci Manuf 2014;64:36–48.
[9] Ardebili A, Alaei MH. Non-destructive testing of delamination defects in GFRP patches using step heating thermography. NDT E Int 2022;128:102617.
[10] Badghaish AA, Fleming DC. Non-destructive inspection of composites using step heating thermography. J Compos Mater 2008;42(13):1337–57.
[11] Cerbu C, Wang H, Botis MF, Huang Z, Plescan C. Temperature effects on the mechanical properties of hybrid composites reinforced with vegetable and glass fibers. Mech Mater 2020;149:103538.
[12] Tao N, Anisimov AG, Groves RM. Towards safe shearography inspection of thick composites with controlled surface temperature heating. NDT E Int 2023;139:102907.
[13] Montanini R, Freni F. Non-destructive evaluation of thick glass fiber-reinforced composites by means of optically excited lock-in thermography. Compos Part A: Appl Sci Manuf 2012;43(11):2075–82.
[14] Deane S, Avdelidis NP, Ibarra-Castanedo C, Zhang H, Yazdani Nezhad H, Williamson AA, Mackley T, Maldague X, Tsourdos A, Nooralishahi P. Comparison of cooled and uncooled IR sensors by means of signal-to-noise ratio for NDT diagnostics of aerospace grade composites. Sensors 2020;20(12):3381.
[15] Vavilov V, Chulkov A, Shiryaev V, Kuimova M, Zhang H. Noise suppression in pulsed IR thermographic NDT: Efficiency of data processing algorithms. NDT E Int 2024;148:103240.
[16] Niccolai A, Caputo D, Chieco L, Grimaccia F, Mussetta M. Machine learning-based detection technique for NDT in industrial manufacturing. Mathematics 2021;9(11):1251.
[17] Yousefi B, Kalhor D, Usamentiaga Fernández R, Lei L, Castanedo CI, Maldague XP. Application of deep learning in infrared non-destructive testing. In: QIRT 2018 proceedings. 2018.
[18] Pedrayes OD, Lema DG, Usamentiaga R, Venegas P, García DF. Semantic segmentation for non-destructive testing with step-heating thermography for composite laminates. Measurement 2022;200:111653.
[19] Luo Q, Gao B, Woo WL, Yang Y. Temporal and spatial deep learning network for infrared thermal defect detection. NDT E Int 2019;108:102164.
[20] Liu K, Zheng M, Liu Y, Yang J, Yao Y. Deep autoencoder thermography for defect detection of carbon fiber composites. IEEE Trans Ind Informatics 2022;19(5):6429–38.
[21] Saeed N, King N, Said Z, Omar MA. Automatic defects detection in CFRP thermograms, using convolutional neural networks and transfer learning. Infrared Phys Technol 2019;102:103048.
[22] Kompanets A, Duits R, Pai G, Leonetti D, (Bert) Snijder HH. Loss function inversion for improved crack segmentation in steel bridges using a CNN framework. Autom Constr 2025;170:105896.
[23] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference. Springer; 2015, p. 234–41.
[24] He Y, Mu X, Wu J, Ma Y, Yang R, Zhang H, Wang P, Wang H, Wang Y. Intelligent detection algorithm based on 2D/3D-UNet for internal defects of carbon fiber composites. Nondestruct Test Eval 2024;39(4):923–38.
[25] Miorelli R, Skarlatos A, Vienne C, Reboud C, Calmon P. Deep learning techniques for non-destructive testing and evaluation. Appl Deep Learn Electromagn: Teach Maxwell's Equations To Mach 2022;99–143.
[26] Jia J, Li Y. Deep learning for structural health monitoring: Data, algorithms, applications, challenges, and trends. Sensors 2023;23(21):8824.
[27] Virkkunen I, Koskinen T, Jessen-Juhler O, Rinta-Aho J. Augmented ultrasonic data for machine learning. J Nondestruct Eval 2021;40(1):4.
[28] Cormerais R, Longo R, Duclos A, Wasselynck G, Berthiau G. Data augmentation and artificial neural networks for eddy currents testing. In: Electromagnetic non-destructive evaluation. IOS Press; 2020, p. 245–52.
[29] Jiangsha A, Tian L, Bai L, Zhang J. Data augmentation by a cyclegan-based extra-supervised model for non-destructive testing. Meas Sci Technol 2022;33(4):045017.
[30] Tian L, Wang Z, Liu W, Cheng Y, Alsaadi FE, Liu X. A new GAN-based approach to data augmentation and image segmentation for crack detection in thermal imaging tests. Cogn Comput 2021;13:1263–73.
[31] Luleci F, Catbas FN, Avci O. Generative adversarial networks for labeled acceleration data augmentation for structural damage detection. J Civ Struct Heal Monit 2023;13(1):181–98.
[32] Tavares A, Di Lorenzo E, Cornelis B, Peeters B, Desmet W, Gryllias K. Machine learning approaches to damage detection in composite structures combining experimental and simulation domains. Mech Syst Signal Process 2024;215:111412.
[33] Arya DM, Maeda H, Ghosh SK, Toshniwal D, Mraz A, Kashiyama T, YSIIof Technology Roorkee, India, TUof Tokyo, Japan, E. Amazon Luxembourg. Transfer learning-based road damage detection for multiple countries. 2020, ArXiv abs/2008.13101, https://api.semanticscholar.org/CorpusID:221377141.
[34] Chamangard M, Ghodrati Amiri G, Darvishan E, Rastin Z. Transfer learning for CNN-based damage detection in civil structures with insufficient data. Shock Vib 2022;2022(1):3635116.
[35] Ma Y, Chen S, Ermon S, Lobell DB. Transfer learning in environmental remote sensing. Remote Sens Environ 2024;301:113924.
[36] Soleimani-Babakamali MH, Soleimani-Babakamali R, Nasrollahzadeh K, Avci O, Kiranyaz S, Taciroglu E. Zero-shot transfer learning for structural health monitoring using generative adversarial networks and spectral mapping. Mech Syst Signal Process 2023;198:110404.
[37] Seventekidis P, Giagopoulos D, Arailopoulos A, Markogiannaki O. Structural health monitoring using deep learning with optimal finite element model generated data. Mech Syst Signal Process 2020;145:106972.
[38] Fang Q, Ibarra-Castanedo C, Maldague X. Automatic defects segmentation and identification by deep learning algorithm with pulsed thermography: Synthetic and experimental data. Big Data Cogn Comput 2021;5(1):9.
[39] Tong Z, Cheng L, Xie S, Kersemans M. A flexible deep learning framework for thermographic inspection of composites. NDT E Int 2023;139:102926.
[40] Brachna R, Kominek J, Guzej M, Kotrbacek P, Zachar M. Numerical computation of anisotropic thermal conductivity in injection molded polymer heat sink filled with graphite flakes. Polymers 2022;14(16):3284.
[41] Wan X, Demir B, An M, Walsh TR, Yang N. Thermal conductivities and mechanical properties of epoxy resin as a function of the degree of cross-linking. Int J Heat Mass Transfer 2021;180:121821.
[42] Tao N, Anisimov AG, Groves RM. FEM-assisted shearography with spatially modulated heating for non-destructive testing of thick composites with deep defects. Compos Struct 2022;297:115980.
[43] Jin F, Park S. Thermal properties of epoxy resin/filler hybrid composites. Polym Degrad Stab 2012;97(11):2148–53.
[44] Chatterjee K, Tuli S, Pickering SG, Almond DP, pulsed Acomparisonofthe. Lock-in and frequency modulated thermography nondestructive evaluation techniques. Ndt E Int 2011;44(7):655–67.
[45] Chrzanowski K. Critical review of present-day methodology of thermal imager noise characterization. Metrol Meas Syst 2025;1–21. http://dx.doi.org/10.24425/mms.2025.154668.
[46] Milanfar P, Delbracio M. Denoising: A powerful building-block for imaging. Inverse Probl Mach Learn 2024.
[47] Van Trees HL, Detection, estimation, theory modulation. Part III: radar-sonar signal processing and Gaussian signals in noise. John Wiley & Sons; 2001.
[48] Vizioli L, Moeller S, Dowdle L, Akçakaya M, De Martino F, Yacoub E, Uğurbil K. Lowering the thermal noise barrier in functional brain mapping with magnetic resonance imaging. Nat Commun 2021;12(1):5181.
[49] Orfanidis SJ. Introduction to signal processing. Prentice-Hall, Inc.; 1995.
[50] ANSYS, Inc. ANSYS mechanical, release 2023 r1, help system. Canonsburg, PA: ANSYS, Inc.; 2023, aNSYS®, https://www.ansys.com.
[51] Bejan A. Convection heat transfer. John wiley & sons; 2013.
[52] Rajic N. Principal component thermography. Tech. rep., DSTO; 2002.
[53] Rajic N. Principal component thermography for flaw contrast enhancement and flaw depth characterisation in composite structures. Compos Struct 2002;58(4):521–8.
[54] Amin A, Ma H, Hossain MS, Roni NA, Haque E, Asaduzzaman S, Abedin R, Ekram AB, Akter RF. Industrial product defect detection using custom u-net. In: 2022 25th international conference on computer and information technology. IEEE; 2022, p. 442–7.
[55] Vasquez J, Furuhata T, Shimada K. Image-enhanced u-net: optimizing defect detection in window frames for construction quality inspection. Buildings 2023;14(1):3.

[56] Wang H, Li X. Expanding horizons: U-net enhancements for semantic segmentation, forecasting, and super-resolution in ocean remote sensing. J Remote Sens 2024;4:0196.
[57] Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data 2016;3:1–40.
[58] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder–decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39(12):2481–95.
[59] Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder–decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision. 2018, p. 801–18.