



**HAL**  
open science

# Automated crack detection on metallic materials with flying-spot thermography using deep learning and progressive training

Kevin Helvig, Pauline Trouvé-Peloux, Ludovic Gaverina, Baptiste Abeloos, Jean-Michel Roche

## ► To cite this version:

Kevin Helvig, Pauline Trouvé-Peloux, Ludovic Gaverina, Baptiste Abeloos, Jean-Michel Roche. Automated crack detection on metallic materials with flying-spot thermography using deep learning and progressive training. Quantitative InfraRed Thermography Journal, 2023, pp.1-20. 10.1080/17686733.2023.2266176 . hal-04644951

**HAL Id: hal-04644951**

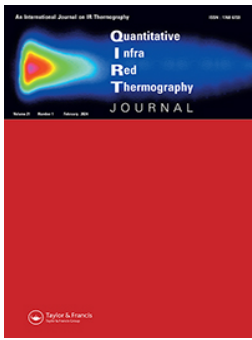
**<https://hal.science/hal-04644951v1>**

Submitted on 11 Jul 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright



## Automated crack detection on metallic materials with flying-spot thermography using deep learning and progressive training

Kevin Helvig, Pauline Trouvé-Peloux, Ludovic Gaverina, Baptiste Abeloos & Jean-Michel Roche

To cite this article: Kevin Helvig, Pauline Trouvé-Peloux, Ludovic Gaverina, Baptiste Abeloos & Jean-Michel Roche (19 Oct 2023): Automated crack detection on metallic materials with flying-spot thermography using deep learning and progressive training, Quantitative InfraRed Thermography Journal, DOI: [10.1080/17686733.2023.2266176](https://doi.org/10.1080/17686733.2023.2266176)

To link to this article: <https://doi.org/10.1080/17686733.2023.2266176>



Published online: 19 Oct 2023.



Submit your article to this journal [↗](#)



Article views: 136



View related articles [↗](#)




View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



# Automated crack detection on metallic materials with flying-spot thermography using deep learning and progressive training

Kevin Helvig <sup>a</sup>, Pauline Trouvé-Peloux<sup>a</sup>, Ludovic Gaverina<sup>b</sup>, Baptiste Abeloos<sup>a</sup> and Jean-Michel Roche<sup>b</sup>

<sup>a</sup>DTIS, ONERA, Université Paris-Saclay, Palaiseau, France; <sup>b</sup>DMAS, ONERA, Université Paris-Saclay, Châtillon, France

## ABSTRACT

In non-destructive testing for metallic materials, 'Flying-spot' thermography allows the detection of cracks thanks to the scanning of samples by a local laser heat source observed in the infrared spectrum. However, distinguishing a crack from other surface structures such as air ducts or non-planar shapes on the material surface can be challenging in an automation perspective. To address this, we propose to use deep learning techniques, which can exploit contextual information but require a significant amount of labelled data. This study presents a training method based on curriculum learning and recent denoising diffusion models to generate synthetic images. The protocol progressively increases the complexity of training images, using successively simulated data from a multi-physics finite-element software, synthetically generated data with diffusion process, and finally real data. Several detection scores are measured for various machine learning and deep learning architectures, demonstrating the benefits of the proposed approach for regular application cases and degraded experimental conditions, consisting of limited thermal enlightenment recordings.

## ARTICLE HISTORY

Received 14 July 2023

Accepted 28 September 2023



## KEYWORDS

Non-destructive testing; flying-spot thermography; deep learning; curriculum learning; denoising diffusion models

## 1. Introduction

The manual inspection of coated metallic materials can present challenges. Specifically, when it comes to surface crack detection, functional cooling structures combined with the condition of the coating can disrupt the operator, leading to false positives or cases of non-detection. This can result in costly consequences, both in terms of discarding non-defective components and wasting operator time.

Among non-destructive testing (NDT) techniques, flying-spot thermography (FST) uses an active local heat source to scan the surface of a metallic sample. Discontinuities in the heat diffusion on the surface, measured in the infrared spectrum, reveal the presence of crack-like defects. This inspection method was originally proposed for crack detection in aerospace metallic parts in the late 1960s [1]. It is effective for characterizing hard-to-detect defects without heavy

**CONTACT** Kevin Helvig  [kevin.helvig@onera.fr](mailto:kevin.helvig@onera.fr)  DTIS, ONERA, Université Paris-Saclay, 6, Chemin de la Vauve aux Granges, Palaiseau 91123, France

© 2023 Office national d'études et de recherches aérospatiales (ONERA)

instrumentation, such as micro-cracks in chipping [2]. However, the conventional examination procedure involving subtracting scanning maps and applying filters faces challenges in automation due to time-consuming processes and the need for precise manual registration and adjustments [3]. To address these limitations, we propose a 'single-pass' mode using a single forward scanning map without subtraction or classical filtering. Working on non-crossing scans eliminates prior knowledge and enables crack length tracking. This simplification of the acquisition process comes with more ambiguity in crack detection on the data, due to edges and material heterogeneity. To overcome this issue, we propose to use a deep learning technique to achieve automatic detection of defects. Here, we propose a progressive training of defect detection models using successively simulated, synthetical and experimental data.

### **1.1. Review of the literature**

Significant efforts have been made to improve the flying-spot thermography (FST) technique, including the use of flying-line thermography, which employs a line of laser spots as heat sources to accelerate scans compared to punctual flying-spot thermography [4]. Simulation works using finite element models (FEM) of flying spot and flying line have also been performed to study the thermal physics phenomena behind this examination technique and plan experiment set-ups [5,6]. Several works have further investigated this technique with the characterization of the detected defects [7] and determination of the influence of defect geometry on cracks thermal signature, like crack width [2]. Hence, FST is a promising examination technique, particularly for detecting micro-cracks on coatings, as well as for a range of potential industrial applications, such as art painting restoration [8]. While several image processing techniques have been proposed for NDT using FST [9], only a recent publication by [10] proposes to use complex recurrent architectures (RNN) based on deep learning to manage temporal features. RNN are neural networks specialized in sequential data processing, such as time series [11]. Our prior research underscores the potential of deep learning to achieve superior performance on FST data [12]. However, it has been observed that the limited quantity of data tends to disrupt the training process and diminish overall performance. Image synthesis using generative models is a common strategy in deep learning to artificially increase the total amount of data. Generative adversarial networks (GANs) have been used extensively for image generation, but they are difficult to train due to issues such as mode collapse [13]. A recent alternative to GANs is denoising diffusion models, deep neural models that generate synthetic images from random noise without the need for a discriminator [14]. This approach is spreading rapidly, with some papers indicating that these networks can outperform GANs in certain cases of image synthesis while being easier to train [15]. They have already been explored in some works highlighting their fidelity in medical examination [16], but to the best of our knowledge, not in the context of FST.

Curriculum learning, an approach to training neural networks that originates from behavioral psychology [17,18], consists in progressively introducing the machine learning system to more complex features through different training steps [19,20]. Although this approach is not yet widely used, it is promising and opens up possibilities for various applications, such as robotics [21] and natural language understanding [22].

## 1.2. Contributions

The first contribution is the implementation of an FST test bench, followed by the generation of simulated and experimental data. A neural network training protocol is then proposed, based on curriculum learning using data simulation and diffusion models as image generators. This protocol is used for automatic detection of crack-like defects on metallic samples by 'single-pass' laser thermography. We highlight the performance gains of the proposed method, compared to direct stand-alone training, as well as the improvements obtained in generalization capabilities to unknown samples. The influence of degraded experimental conditions is evaluated. The proposed curriculum learning approach hence provides a grounded framework for a proper use of synthetic data and generic simulation to train deep learning models for the automation of thermal examination techniques.

## 2. Problem statement and data

This section describes data collections for the proposed training method. [Subsection 2.1](#) presents the theoretical elements of FST, providing some key elements to calibrate the experimental and simulated set-up. The subsection compares in more details the positioning adopted here, compared with the conventional approach [3]. Then, there is an important focus on the datasets built using the simulation framework in [subsection 2.2](#), the experimental set-up built for this study in [section 2.3](#) and the synthetic data generation using diffusion models [2.4](#). A dataset summary is given in [subsection 2.5](#).

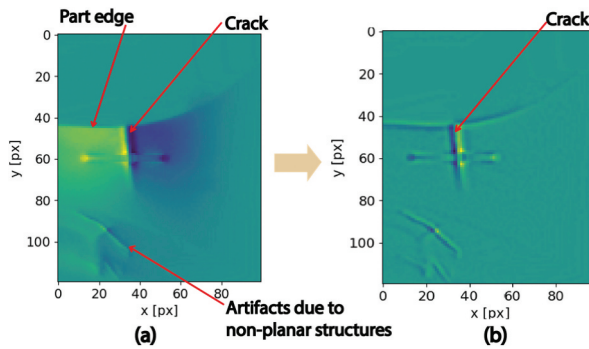
### 2.1. Problem statement

In FST, it is commonly assumed that the scan velocity remains constant during the examination. Additionally, the thermal emissivity and diffusivity of the material are quasi-constant as well. Specifically, the examinations are conducted for limited temperature variations, which are studied in the MWIR bandwidth (Middle Wavelength Infrared, 3–5  $\mu\text{m}$ ). In this context, the theoretical work of Krapez [3] introduces the Peclet number. This non-dimensional number is defined as to the ratio between convective heat transfer and thermal diffusion. According to [3], the best detection, understood as the most significant thermal discontinuity due to the crack, corresponds to a Peclet number of 1.

$$Pe = \frac{\text{Convective heat flux}}{\text{Heat diffusion}} = \frac{v_{spot} \times R_{spot}}{a}.$$

With  $v_{spot}$  [mm/s]: the velocity of the heat source scanning the surface.  $R_{spot}$  [mm]: the size of the spot due to the heat source and  $a$  [ $\text{mm}^2/\text{s}$ ]: the thermal diffusivity of the materials.

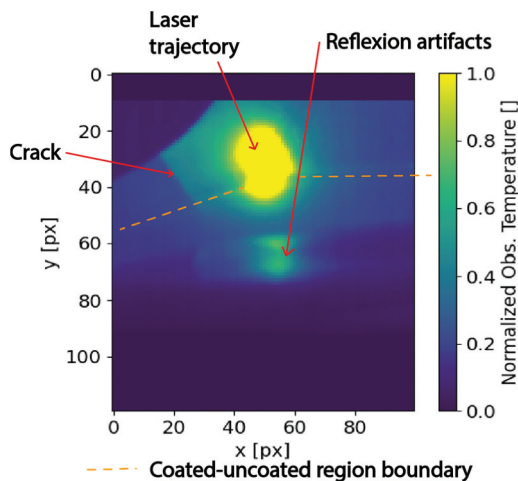
The processing method proposed in [3] for crack detection is based on subtracting two forward and backward scanning maps crossing the defect, then using Laplacian filtering. This method is applied in [Figure 1](#) on a millimetric defect, the observation scene is on the order of centimeters. This figure shows on the left the difference between both scans, on the right, the obtained map after filtering, highlighting the signal due to the crack. However, this method rises issues in an automation context: it is time-consuming for inspection and processing, requiring delicate registration of the



**Figure 1.** (a) FST thermal image before filtering obtained by following the state of the art subtraction method [3]. (b) Thermal image after Laplacian filtering: if there are still some artifacts due to the heat trajectory and non-planar surfaces, the signal due to the crack is distinguishable.

two maps before subtraction, and fine adjustments for filtering. Finally, the considered scans are passing through, assuming prior knowledge about the location of the defect and its orientation.

On the contrary, we propose to work directly on a single forward scanning map, in a so-called ‘single-pass’ mode, without subtraction nor classical filtering. We also propose to work on non-passing scans to eliminate the operator’s prior knowledge of the defect and to give the possibility to follow the crack length on the material surface. [Figure 2](#) provides an example of a thermal image used in this study. This image is obtained by averaging and normalizing all the images from a thermal recording during the scan. [Figure 2](#) illustrates various problematic surface elements for an automatic defect detection, such as material heterogeneity or heat source discontinuities. We propose here to adress the automatic detection using the ability of deep learning to learn contextual information.



**Figure 2.** Example of obtained using a ‘single-pass’ detection method, which illustrates the various surface elements that interfere with defect detection.

**Table 1.** Summary of samples used in this study. The same region is examined for each sample.

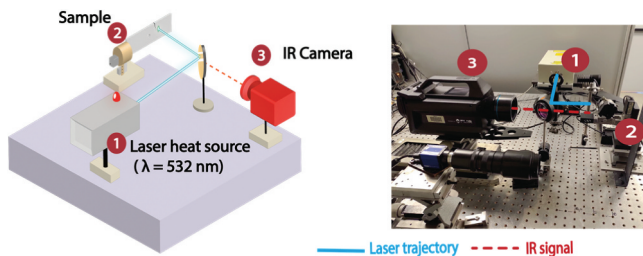
Sample code	Composition	Ratio crack length/critical length
1	Coated superalloy	0.7
2	Coated superalloy	0.9
3	Coated superalloy	1.0
4	Coated superalloy	1.2
5	Coated superalloy	1.3
6	+3 coated samples without defect	

## 2.2. Samples and experimental data

The study is performed on several coated metallic superalloy samples. The coating is a thermal barrier. Samples 1, 3, 5, and 6 are used to train the diffusion model and to fine-tune the classification networks. Fine-tuning consists in proceeding only to the training of the last layer of a pre-trained network, reducing the number of parameters trained, reducing this way the data starvation. They are also used for the direct stand-alone training. Samples 2 and 4 are used to test the generalization capabilities of the networks on unknown samples that were not included in the initial training. Sample 2 has a unique marking on the surface coating that can disturb detection. The samples and their main characteristics are summarized in Table 1. We can only provide here the ratio crack length-critical length, which gives an idea of how the crack length is distributed between the available samples. The critical crack length is the maximum allowable crack length defined by the manufacturer's specifications, referred to as the criticality in the study, is a few millimeters.

Figure 3 shows the Onera FST bench used in order to generate experimental data for crack detection. This set-up uses a laser heat source with a power varying from 0.5 to 3 W. The wavelength is 532 nm. A dichroic lens is added to reflect the laser spot on the part and to transmit the IR heat flux to the MW-IR camera, sensitive between 3 and 5  $\mu\text{m}$ . The spot covers a distance of 4.5 mm for horizontal scans, 6 mm for vertical scans, close to the defect. The scan velocity varies between 0.5 and 2.5 mm/s. Various sets of parameters are used around the operating point (Peclet tending to 1). An angular rotation is also manually applied to the samples. It varies from  $0^\circ$  (defect is vertical) to  $45^\circ$  in experiments.

The ranges of experimental settings are distributed as presented here considering the theory. 'Enlarged settings' correspond to  $Pe \in [0.35, 5.31]$ , 'Optimal settings' to  $Pe \in [0.71, 3.18]$ . 'Degraded settings' have the same Peclet numbers as in 'Enlarged settings' but with an important amount of experiments performed with limited heating duration and distant from the crack. The samples 1,3,5 and 6 are used to train machine learning architectures in several different databases. The base Ba is used with enlarged settings, to produce



**Figure 3.** FST bench of Onera.

**Table 2.** Summary of the different datasets used in this study. Origin, number of image, samples and examination settings used are indicated.

Base	Source	Nimages	Samples	Settings
Aa	Simulation	28,975	–	Enlarged settings
Ab	Simulation	6,000	–	Enlarged settings
Ba	Experimental	600	1,3,5 + 6	Enlarged settings
Bb	Diffusion models	20,000	–	–
Ca	Experimental	330	1,3,5 + 6	Optimal settings
Cb	Experimental	21	2,4	Optimal settings
D	Experimental	950	1,3,5 + 6	Degraded settings

a first amount of experimental data to train the generative model. Base Ca is used for fine-tuning trainings on the application data (Optimal settings). A dataset is also created with these samples, in order to test the performance in case of degraded experimental conditions (low thermal enlightenment), forming the database D. Samples 2, 4 are reserved to control generalization capabilities to different samples not examined for training steps (dataset Cb). Table 2 gives a summary of these different datasets, with the associated number of thermal images. All the experimental datasets used for the training steps are balanced between cracked and no-crack thermal images.

### 2.3. Simulated data

Simulated data are produced using the multi-physics finite-element software Comsol [23]. The simulation shows the trajectory of the heat source on the surface, on which a crack may or may not be present. The crack is simulated by a linear thermal resistance. The simulations do not include any surface coating nor heterogeneity. As in the experimental bench, simulated inspection parameters such as laser velocity, spot radius, and power are expanded around the optimal Peclet number to increase the amount of produced images. The heat diffusion in the surface is modeled by the convection-diffusion Equation (1), and laser heat flux  $\phi$  follows a Gaussian approximation (2) based on the studies [24,25]. The equations are given below:

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = \frac{v}{a} \cdot \frac{\partial T}{\partial y}. \quad (1)$$

With  $T$  as temperature [K],  $v$  as spot-velocity [mm/s],  $a$  as thermal diffusivity [mm<sup>2</sup>/s].

$$\phi = \frac{A \times P}{S} \times \exp\left(\frac{-(x - x_0)^2 - (y - y_0 - v \cdot t)^2}{R_{spot}^2}\right). \quad (2)$$

With  $A$  as absorption rate [%],  $P$  as power [W],  $S$  as irradiated surface area [mm<sup>2</sup>]. Variables  $x$  and  $y$  represent spatial location of the spot.

The crack is conventionally modeled as a thermal resistance of the 'air gap' type (3).

$$R_{th} = \frac{e}{\lambda_{air}}. \quad (3)$$

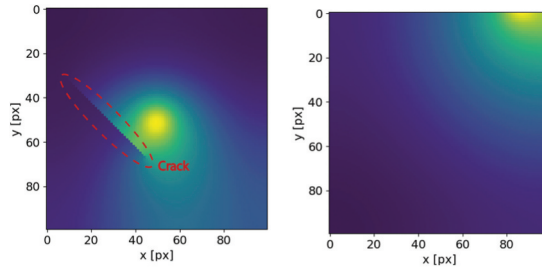
With  $\lambda_{air}$  the air thermal conductivity [ $W.m.K^{-1}$ ] and  $e$  the crack opening in this model [mm].

Several elements of the simulation framework are also available in Table 3.



**Table 3.** Settings used for the FEM simulations.

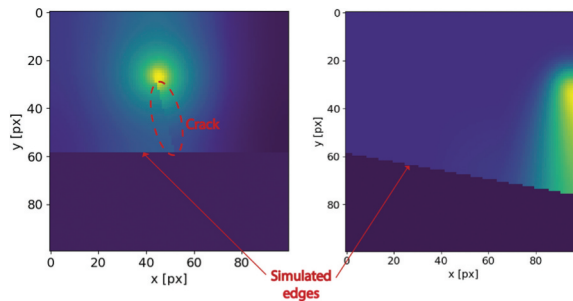
Setting	Value
Crack length [mm]	5 or 10
Orientation [degree]	0 to 90
Scan velocity [mm/s]	0.5 to 2.5
Spot radius [mm]	0.5 to 1.5
Distance crack-scan [mm]	0 to 10
Thermal diffusivity [m <sup>2</sup> /s]	7.1e-7
Mesh type	Triangular



**Figure 4.** Examples of simulated thermal images from Aa dataset. On the left, the simulation presents a defect. On the right, the image is a negative example without defect.

A first training set is generated, representing only the metallic surface, without any contours or structures. This dataset is called Aa and contains 13,000 thermal images without cracks and 15,000 with cracks. Figure 4 presents an example of a simulated image generated following this protocol for each class.

A second dataset is generated, presenting simple generic edges, such as a straight line separating the air simulated environment from the rest of the metal surface. A small rotation is added to this simulated edge, increasing contours variability. This dataset is called Ab. The variability due to the orientation and the location of the simulated metallic surface is not incorporated into this second simulation dataset: it will be introduced through common data augmentations during training such as rotation or mirror flips. This second simulated dataset contains 6,000 thermal images. This is a balanced dataset between crack and no crack images. Figure 5 presents an example of a simulated image generated following this protocol for each class.



**Figure 5.** Examples of simulated thermal images from Ab dataset. At left, the simulation presents a defect. Right image is a negative example without defect. Edges are easily visible to the naked eyes.

## 2.4. Synthetic data generation using diffusion models

For synthetic data generation, a denoising diffusion model is used [14]. The principle of this network is illustrated in Figure 6. Diffusion models are trained to convert images to Gaussian noise. Each step of this process is a denoising network, converting progressively the input image to noise. The denoising networks used in this study are U-nets, a common architecture deployed for image segmentation [26]. In a second time, trained layers are reversed to convert random noises to new synthetic images. A Pytorch adaptation of the original diffusion model developed by Ho et al. has been successfully trained without hyper-parameters adjustment [14], on the base Ba, generating the base Bb containing 20,000 thermal images, which is balanced between crack and no crack images too [27].

This database introduces the networks to more specific features able to disturb defect detection, like surface heterogeneity or complex part edges. Features vary from one thermal image to another.

Figures 7 and 8 give examples of artificial images generated using this model, respectively for without crack and with crack cases. These images highlight

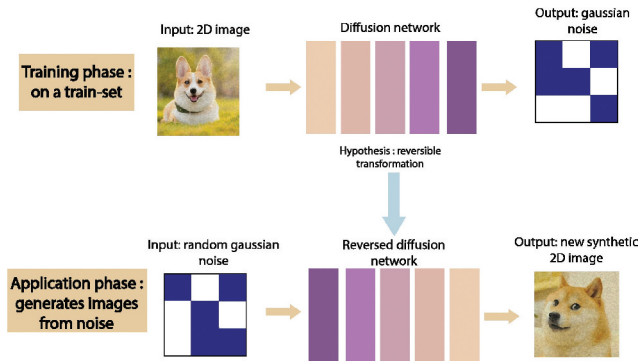


Figure 6. Denoising diffusion model principle.

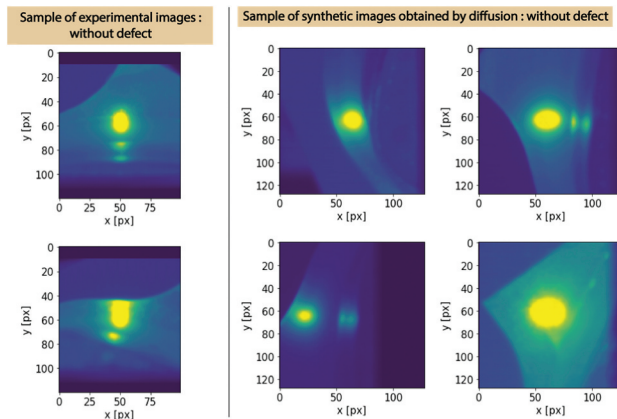
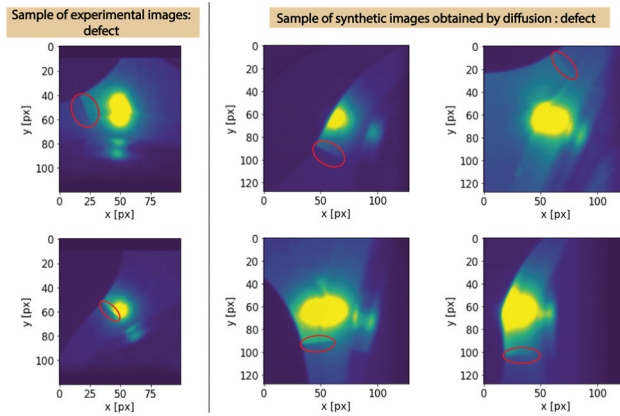


Figure 7. Sample of experimental images and synthetic images obtained using a denoising diffusion model, for uncracked images. For visual comparison.

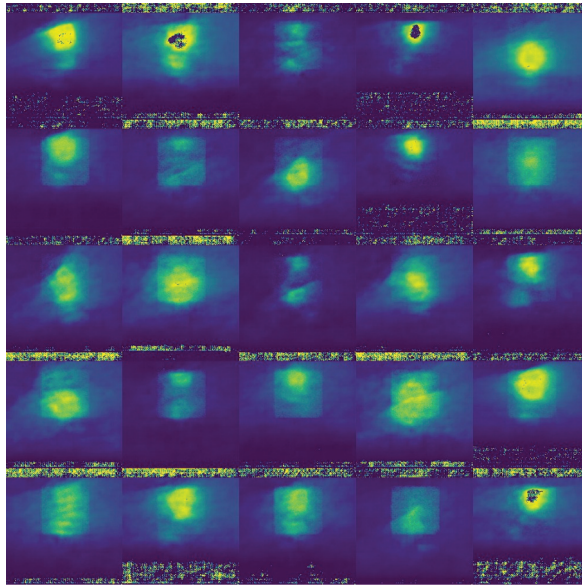


**Figure 8.** Sample of experimental images and synthetic images obtained using a denoising diffusion model, for cracked images. Defect circled in red.

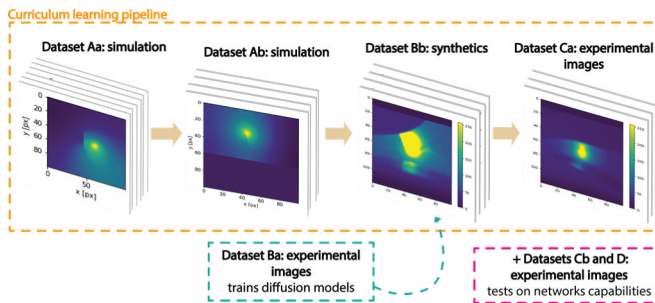
qualitatively the ability of the diffusion model to generate accurate images. Some features of these synthetic images are noticeable, seeming well imitated: border between coated and uncoated regions, the influence of this border on thermal diffusion, reflection artifacts, and samples edges. A comparison is added here between diffusion models and variational auto-encoders (VAE). These architectures are very traditional generative models: they are components to build greater synthetic generators, such as diffusion models or GANs, as for natural image synthesis [28]. The selected architecture is a beta-VAE, which is a very common and simple model for image synthesis. Pythae library is used to build the VAE architecture [29]. The Beta-VAE is trained with the same base as for the diffusion models. A small amount of data is generated as illustrated in Figure 9 which gives some generated samples using this architecture. As shown, the architecture struggles to generate synthetic data that are as accurate as diffusion-generated synthetics. The encoder and the decoder of the VAE have also needed to be built specifically to approach diffusion model synthesis performance, whereas a generic diffusion model performs very well for a large panel of publicly available datasets [15]. The properties of diffusion models may explain this difference, such as a greater expressive power, and a distributed and progressive noise-to-image transformation. If images generated using the beta-VAE must be relevant to learn some features with a less important computational cost than a diffusion model, the poor sample quality will alternate the features learning of the neural architectures.

## 2.5. Datasets summary

Table 2 provides a summary of the different databases, their origin, and their main use. The table lists the samples used, the amount of thermal images, and the experimental conditions used. The dataset Ca is split between training and test data, with a 2/3 ratio. Figure 10 gives a synthesis of the different datasets, which are used for training or for experiments on the capabilities of the networks.



**Figure 9.** Samples generated using our modified beta-VAE. Several major features such as edges emerge, but the synthesis is still confuse and noisy.



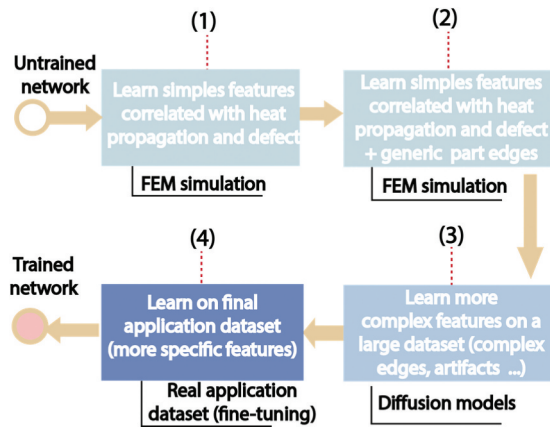
**Figure 10.** Illustration of the different datasets used for the training pipeline, with the source of each kind of data.

### 3. Method

This section describes the proposed training pipeline. The curriculum learning approach is presented in subsection 3.1, and architectures and metrics used in 3.2 and 3.3.

#### 3.1. Curriculum learning pipeline

The training protocol based on curriculum learning is described in Figure 11. The training starts with an untrained model having random weights and a generic features learning on simulated finite element data related to thermal aspects such as heat diffusion on the surface (step 1). An intermediate simulated step (step 2) is included, which allows to introduce into the training how structures such as generic, rectilinear contours will alter



**Figure 11.** Proposed training protocol based on a progressive training of defect detection models using successively simulated, synthetical and real data.

heat propagation, in addition to the defect. The network is then retrained on synthetic images produced by the diffusion models (step 3). This step allows to finalise the learning process by moving from generic features to their more specific counterparts, related to heat diffusion on a non-homogeneous material. Fine-tuning is then performed on a limited sample of experimental data (step 4). Typical augmentations such as horizontal and vertical inversions, as well as random rotations, are applied at each step of the training. The optimizer used is the Adam optimizer [30]. Learning rate is  $10^{-4}$ , with scheduling [31].

### 3.2. Architectures

Various deep learning architectures for classification are trained using the proposed training pipeline. On the one hand, a well-known convolutional classification network, a visual geometry group architecture, VGG-13, is trained [32]. Then, architectures based on the attention mechanism are studied, the vision transformers [33]. Attention mechanism allows for hierarchical learning that associates different regions of the input image [34]. This approach seems relevant for the application studied here, which presents both local illumination and multi-scale phenomena, here thermal. This ability from transformers is suggested in [35]. Therefore, two networks based on the attention mechanism are selected, the Shifted-attention-windows (Swin) and Class-attention-in-Transformers (CaiT) architectures [36,37]. The baseline is a traditional machine learning architecture that uses a Histogram-of-Oriented-Gradient (HOG) filter as a features extractor and a Support Vector Machine (SVM) as a classification head [38]. It provides a benchmark for classical machine learning methods. All the selected models are compared with their direct training on Ca, in order to measure the benefits of the proposed approach. Table 4 gives more details about the deep learning models used in this study, with the associated name in timm/pytorch-image-models library [39], used to load the different models studied here with Pytorch deep learning library.

**Table 4.** Models used in this study. The number of parameters corresponds to the number of trainable weights in the deep learning architectures.

Model name	Reference in pytorch-image-models	Number of parameters
VGG13	'vgg13'	128,957,890
CaiT	'cait_xxs24_224'	11,617,538
Swin	'swin_s3_small_224'	49,555,528

### 3.3. Metrics

For all selected methods, the metrics are evaluated on the test subset of dataset Ca. Accuracy, precision, recall and F1-score (also known as f-score) are measured [40]. Accuracy (4) is an estimation of overall model performance, representing the proportion of correctly predicted samples out of the total number of samples. Precision (5) quantifies the proportion of true positive predictions (correctly classified as presenting a defect) among all positive predictions made by the model, while recall (6) (sensitivity) estimates the proportion of true positive predictions among all actual positive instances. These metrics play a crucial role in assessing a classification model's effectiveness. The F1 score (7), or f-score, which combines precision and recall into a single metric, is particularly useful. It is calculated by taking the harmonic mean of precision and recall, providing valuable insights into the model's predictive capabilities and performance across different classes of data. The equations corresponding to each metric are given below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\text{Correct classifications}}{\text{Total}} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{F1Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

with false positives (FP) corresponding to false alarms, while false negatives (FN) correspond to missed detections. True positives (TP) are true crack thermal images, whereas true negative (TN) corresponds to true no-crack thermal images.

## 4. Results

### 4.1. General results

Table 5 summarizes the test results obtained by different selected architectures on dataset Ca for both direct training (using the training set of dataset Ca) and the curriculum learning approach proposed. The final step of the proposed training protocol (fine-tuning on the application dataset) is performed on this dataset. Overall, all the deep learning methods provide very high test-accuracy value and a significant performance gain

**Table 5.** Scores obtained using direct training and with the curriculum-based training proposal (evaluation on the test-subset of the dataset Ca).

Method	Architecture	Test-accuracy [%]	F1	Precision	Recall
<i>Baseline</i>	HOG+SVM	83	0.83	0.82	0.83
Direct training	VGG13	92	0.92	0.95	0.90
	Swin	87	0.86	0.92	0.86
	CaiT	90	0.91	0.93	0.89
<i>Curriculum learning</i>	VGG13	97	0.97	0.96	0.99
	Swin	96	0.96	0.94	0.98
	CaiT	99	0.98	0.98	0.99

compared to traditional machine learning methods, with a test-accuracy above 90% compared to 83%, respectively. Moreover, the use of curriculum training increases the performance compared to direct training: VGG13 architecture has an increase in test-accuracy of 5%, while it is about 9% for CaiT and Swin architectures. Hence, the proposed training method looks more beneficial for transformer architectures. Additionally, the performance differences between VGG13 and attention-based architectures may indicate the benefits of the attention mechanism for the application case. Architectures trained using the curriculum learning approach also appear to exhibit a reduced gap between precision and recall: the results table shows a 4% gap for CaiT architecture in direct training, whereas this gap is reduced to 1% with the proposed training method. The proposed method reduces the small bias favorizing missed detection over false alarms (Precision > Recall), which is present with conventional direct training.

## 4.2. Generalization capabilities

The generalization abilities of the networks trained with the proposed method are evaluated and compared with direct training from the Ca database. Generalization abilities to unknown samples are first evaluated, followed by the impact of degraded experimental conditions.

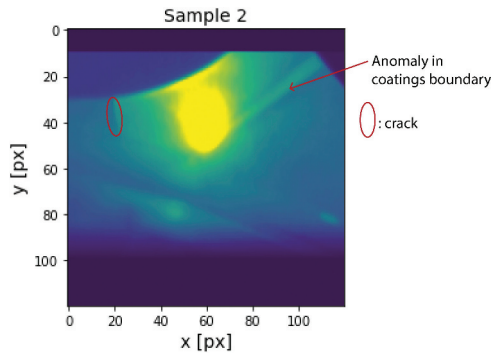
### 4.2.1. Generalization to new samples

Trained CaiT and VGG13 (using either direct or curriculum training) are tested on scans of samples 2 and 4 of [Table 1](#), not used during training, constituting the database Cb. [Table 6](#) gathers the obtained performance with test-accuracy. For sample 4, this table shows that the training process based on curriculum learning increases robustness to unknown samples, compared to direct training. The highest decrease of performance for sample 2 can be explained by an anomalous surface in comparison with other samples as shown in [Figure 12](#). This untypical coating boundary can disturb classification. Transformers seem also to have a better generalization ability than more usual convolutional networks, which is consistent with recent works on this kind of architecture [41].

**Table 6.** Test-accuracy obtained on the base Cb, which corresponds to samples not seen during training. The number of images per sample is given.

Ech.	#image	Acc. (CaiT, direct)	Acc. (CaiT, curriculum)	Acc. (VGG13, direct)	Acc. (VGG13, curriculum)
2	11	59	82	41	77
4	10	90	96	78	90

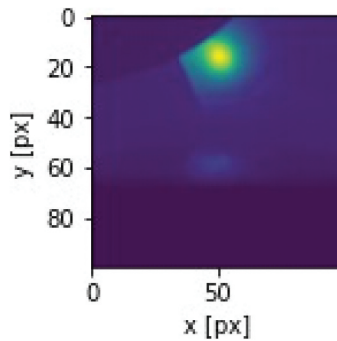




**Figure 12.** Example of thermal image from sample 2. The specificity of the surface examined may disturb detection (base D).

#### 4.2.2. Degraded thermal conditions

The selected neural networks are challenged using data from various experimental conditions, especially degraded and poor conditions, which show low crack response, using dataset D. Direct training is performed on this dataset, as well as the fine-tuning for the proposed approach. Figure 13 provides an example of these degraded thermal images. While defects can be distinguished through eyes on the thermal image, other structures like part edges are hard to see. Table 7 summarizes the accuracy obtained on a test-subset for this specific database, for direct training and the proposed training approach.



**Figure 13.** Thermal image for degraded conditions (scan velocity is 2.5 mm/s). If the defect can be detected easily in this example, contextual information such as coating boundary or edges are harder to distinguish.

**Table 7.** Tests of selected neural networks on base with degraded experimental conditions, base D.

Architecture	Test-accuracy (direct)	Test-accuracy (Curriculum)
VGG13	54 (no learning)	69
Swin	73	74
CaIT	72	78



**Table 8.** Results of the ablation study on the training pipeline: performance is evaluated for architectures selected if one or more training steps are shortcut (base C, evaluation subset).

Ablation	Train last step?	VGG13				Swin				CaiT			
		Acc.	F1	P.	R.	Acc.	F1	P.	R.	Acc.	F1	P.	R.
No step 1 and 1.5 (FEM)	Fine-tuning	96	0.97	0.97	0.97	98	0.98	0.97	0.99	95	0.95	0.93	0.97
	No step 3	96	0.97	0.94	0.99	96	0.96	0.94	0.99	94	0.95	0.91	0.99
No step 2 (diff.)	Fine-tuning	61	0.69	0.75	0.67	66	0.68	0.70	0.65	70	0.70	0.72	0.65
	No step 3	55	0.70	0.55	0.99	57	0.67	0.56	0.83	55	0.69	0.54	0.94
No step 3	–	97	0.96	0.94	0.99	94	0.95	0.89	0.99	97	0.97	0.94	0.99

Convolutional architectures seem to struggle with this dataset, as VGG trained with direct learning is unable to distinguish the defect. This issue is not observed with transformer architectures. The proposed curriculum approach improves performance, even on this suboptimal subset. These findings demonstrate the challenges that networks face when learning from degraded experimental conditions. Therefore, using the optimized parameters described in [section 2](#) is necessary for a more accurate crack detection.

### 4.3. Ablation study of the training pipeline

In this section, an ablation study is performed on the proposed training pipeline. The main principle is to miss out each element of the training procedure in order to give indications of its most contributing steps.

[Table 8](#) gives a synthesis of all the ablations performed. Impacts of these ablations in terms of test-accuracy is evaluated for all the architectures studied. F1, precision and recall are also computed.

[Table 8](#) shows poorer performance without diffusion training, with a test-accuracy between 61% and 70% (No step 2 in the Table). While the training with diffusion data and without simulations is close to the performance obtained with the complete training pipeline. From these results the most crucial element of the training pipeline proposed seems to be the diffusion model, giving a large synthetic dataset to train the networks. This step gives important enough amounts of data to perform the feature learning well, with high performance on the evaluation subset, whereas simulations give just refining increases of performance. However, pretraining on simulated data using a physically correlated model is still relevant, in order to focus neural networks on generic features associated with the defect and how it alters thermal diffusion, while reducing the total number of training epoch.

## 5. Discussions

In this section, we discuss several points of our proposed method, firstly, the number of samples and experimental conditions, then the training pipeline, and finally the choice of the generative model.

### 5.1. Data

The samples studied in this work have relatively intact thermal coatings. These samples present a mark on its surface. Adding more aged parts with highly deteriorated coatings

may introduce additional noise in the features observed by neural networks. It would be worthwhile to investigate robustness of architectures trained with the proposed method compared to direct training when more diverse and challenging samples are included. This could include a wider range of crack types or part geometries to further evaluate the capabilities of the trained models, in the perspective of building a generalist FST crack detector.

Results obtained in this study suggest a significative increase of performance for the proposed training method, as for application case (base Ca) as for tests with various conditions (databases: Cb and D). However, the influence of settings such as camera resolution is not explored here. Studying the performance of the training pipeline in even harder conditions such as introducing variation in image resolution or other thermal properties could be interesting to explore how to increase the robustness of tested architectures. Finally, the influence of the amounts of data has not been studied here. A strategy 'more is better' has been followed: the generative capabilities of the diffusion models are not limited, even if redundant features may appear, depending on the diversity of the input distribution. The question of the amount of simulated data is more interesting: if a large amount of data has been produced here, a more limited sample of simulations could work too, due to the limited complexity and quantity of features of the simulations proposed, and regarding the limited impact of these data on final performance. It can be explored in further research.

## **5.2. Synthetic generation**

The work presented highlights the ability of diffusion models to generate synthetic from relatively limited amounts of data, with a generic architecture, whereas other approaches need an important work to be adjusted to the task, and will be generic by design. Diffusion models appear to beat complex architectures like GANs [15]. We also need to address concerns about the computational cost of diffusion models. This study involved a complete training of this type of architecture from scratch, which was computationally expensive. However, the more common approach of fine-tuning a pre-trained diffusion model is clearly relevant and could be a possible improvement to greatly reduce this cost, especially in cases with more reduced datasets. Very new fine-tuning techniques dedicated to text-to-image diffusion models can perform effective image syntheses with very limited amounts of data (no more than a dozen of images), such as dreambooth approach [42].

## **6. Conclusions and perspectives**

In this article, we have presented the coupling of deep learning associated with the proposal of 'single-pass' FST for defect detection on our application samples. The paper proposed a method for automated defect detection in metallic materials using FST, based on detection networks trained through a curriculum learning approach. This proposal of training approach demonstrated performance gains compared to direct learning. The proposed training process demonstrates both general increases of performance and reduced bias between missed detection and false alarms for all the architectures studied. Furthermore, this training protocol has shown that it improved robustness of the studied deep learning models in various experimental cases.

Several possibilities are identified in order to expand this work. The comparison between direct training and curriculum learning can obviously be extended to other architectures. Playing more with the amounts of data introduced at each step of the curriculum learning could be interesting too. Evaluating the influence of crack length is another important perspective, but needs more samples to be conducted in a proper way. The construction of a large publicly available pre-training dataset is part of our ongoing work. This would allow for testing such comparisons in proper ways. On the other hand, we could turn to more advanced diffusion models [43]. To conclude, the proposed training approach is not limited to the FST inspection technique. Its application to other NDT techniques, such as flash thermography, could be implemented in the future.

### Disclosure statement

No potential conflict of interest was reported by the author(s).

### Funding

The work was supported by the Agence de l'innovation de Défense.

### Notes on contributors

**Kevin Helvig** is a Ph.D. student currently doing research at ONERA Palaiseau in France. His work is dedicated to the application of computer vision techniques to laser thermography for non-destructive materials testing, in particular exploring the coupling between active IR and visible spectrum examinations. He is graduated with an engineering degree from IMT Mines Albi, specializing in non-destructive testing and materials.

**Pauline Trouvé-Peloux** after completing her engineering training in optics at the Institut d'Optique Graduate School, Pauline Trouvé-Peloux obtained her doctorate in Information and Mathematics Science and Technologies from the Ecole Centrale de Nantes in 2012, specializing in signal and image processing. Since 2012, she has held the position of research engineer at ONERA, within the Information Processing and Systems Department (DTIS). Her research activities focus on the joint design, or co-design, of an imager through joint optimization approaches of its optics and processing parameters. The application areas of her work particularly concern compact 3D sensors for robotics or industrial inspection.

**Ludovic Gaverina** graduated with an engineering degree from Telecom Saint-Etienne and a master of research in optic, image, and computer vision from Jean Monnet University in 2013. In 2017, he received the PhD degree (title of his thesis: "Thermal characterization of heterogeneous material by flying spot laser and infrared thermography") in heat transfer from Bordeaux University under the supervision of Christophe Pradère. He currently works at ONERA, within the Materials and Structures Department, focusing on automated multiphysics non-destructive testing (NDT) techniques.

**Baptiste Abeloos** is a Research Scientist at The French Aerospace Lab ONERA, within the Information Processing and Systems Department. His PhD thesis, titled "Searches for Supersymmetry in the Fully Hadronic Channel and Jet Calibration with the ATLAS Detector at the LHC," focused on enhancing jet energy measurement accuracy and propelling the search for supersymmetry, aiding in extending the known limitations on squark and gluino masses. Currently, he works on deep learning techniques for explainability, vision-language models, and non-destructive testing.

**Jean Michel Roche** is a senior research scientist in ONERA, leader of the R&D team dedicated to non-destructive testing and structural health monitoring. He graduated from Ecole Centrale de Lyon in 2007, specializing in aeroacoustics, and successfully defended his PhD thesis in the field of the absorption of resonant liners, in 2011. Since then, he has been working on thermal-based approaches to detect defects in aeronautic structures.

## ORCID

Kevin Helvig  <http://orcid.org/0000-0001-6387-6817>

## References

- [1] Kubiak EJ. Infrared detection of fatigue cracks and other near-surface defects. *Appl Optics*. 1968 Sep;7(9):1743–1747. doi: [10.1364/AO.7.001743](https://doi.org/10.1364/AO.7.001743)
- [2] Archer T, Beauchêne P, Passilly B, et al. Use of laser spot thermography for the non-destructive imaging of thermal fatigue microcracking of a coated ceramic matrix composite. *Quant InfraRed Thermogr J*. 2019 Dec;18(3):141–158. doi: [10.1080/17686733.2019.1705732](https://doi.org/10.1080/17686733.2019.1705732)
- [3] Krapez J-C. Résolution spatiale de la caméra thermique à source volante. *Int J Therm Sci*. 1999;38(9):769–779. doi: [10.1016/S1290-0729\(99\)80033-7](https://doi.org/10.1016/S1290-0729(99)80033-7)
- [4] Mokhtari Y, Gavérina L, Ibarra-Castanedo C, et al. Comparative study of line scan and flying line active IR thermography operated with a 6-axis robot. In: *Proceedings of the 2018 International Conference on Quantitative InfraRed Thermography*; Jun 2018; Berlin (Germany): QIRT Council; 2018.
- [5] Li T, Almond DP, Rees DAS, et al. Crack imaging by pulsed laser spot thermography. *J Phys*. 2010 Mar;214:012072.
- [6] Li T, Almond DP, Rees DAS. Crack imaging by scanning laser-line thermography and laser-spot thermography. *Meas Sci Technol*. 2011 Mar;22(3):035701. doi: [10.1088/0957-0233/22/3/035701](https://doi.org/10.1088/0957-0233/22/3/035701)
- [7] Salazar A, Mendioroz A, Oleaga A. Flying spot thermography: quantitative assessment of thermal diffusivity and crack width. *J Appl Phys*. 2020 Apr;127(13):131101. doi: [10.1063/1.5144972](https://doi.org/10.1063/1.5144972).
- [8] Sfarra S, Gavérina L, Pradere C, et al. Integration study among flying spot laser thermography and terahertz technique for the inspection of panel paintings. *J Therm Anal Calorim*. 2022;147(15):8279–8287. doi: [10.1007/s10973-021-11181-8](https://doi.org/10.1007/s10973-021-11181-8)
- [9] Pech-May NW, Ziegler M. Detection of surface breaking cracks using flying line laser thermography: a Canny-based algorithm. *Eng Proc*. 2021 Nov;8(1):22.
- [10] Shi W, Ren Z, He W, et al. A technique combining laser spot thermography and neural network for surface crack detection in laser engineered net shaping. *Opt Lasers Eng*. 2021 Mar;138:106431.
- [11] Schmidt RM. Recurrent neural networks (RNNs): a gentle Introduction and overview. arXiv: 1912.05911 [Cs, Stat]. 2019 Nov.
- [12] Helvig K, Gaverina L, Trouvé-Peloux P, et al. Toward deep learning fusion of flying spot thermography and visible inspection for surface cracks detection on metallic materials. *Proceedings of the 2022 International Conference on Quantitative InfraRed Thermography*; July 2022; Paris (France): QIRT Council; 2022.
- [13] Liang KJ, Jain A, Abbeel P. Generative adversarial network training is a continual learning Problem. arXiv: 1811.11083 [Cs, Stat]. 2018 Nov.
- [14] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Number: arXiv: 2006.11239 arXiv: 2006.11239 [Cs, Stat]. 2020 Dec.
- [15] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. arXiv: 2105.05233 [Cs, Stat]. 2021 Jun.

- [16] Xie Y, Li Q. Measurement-conditioned denoising diffusion probabilistic model for Under-sampled Medical image Reconstruction. Number: arXiv: 2203.03623 arXiv: 2203.03623 [Cs, Eess]. 2022 Mar.
- [17] Pavlov PI. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. *Ann Neurosci*. 2010 Jul;17(3):136–141. 1927. doi: [10.5214/ans.0972-7531.1017309](https://doi.org/10.5214/ans.0972-7531.1017309)
- [18] Krueger KA, Dayan P. Flexible shaping: how learning in small steps helps. *Cognition*. 2009 Mar;110(3):380–394. doi: [10.1016/j.cognition.2008.11.014](https://doi.org/10.1016/j.cognition.2008.11.014)
- [19] Bengio Y, Louradour J, Collobert R, et al. Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09; New York, NY, USA: Association for Computing Machinery; 2009 Jun. p. 41–48
- [20] Hachohen G, Weinshall D. On the power of curriculum learning in training deep networks. Number: arXiv: 1904.03626 arXiv: 1904.03626 [Cs, Stat]. 2019 May.
- [21] Kilinc O, Montana G. Follow the object: curriculum learning for manipulation tasks with imagined goals. arXiv: 2008.02066 arXiv: 2008.02066 [Cs, Stat]. 2020 Aug.
- [22] Chevalier-Boisvert M, Bahdanau D, Lahlou S, et al. BabyAI: a platform to study the sample efficiency of grounded language learning. arXiv: 1810.08272 [Cs]. 2019 Dec.
- [23] COMSOL Multiphysics. Introduction To Comsol multiphysics®. Burlington, MA: COMSOL Multiphysics; 1998 [cited 2018 Feb 9].
- [24] Thiam A, Kneip J-C, Cicala E, et al. Modeling and optimization of open crack detection by flying spot thermography. *NDT E Int*. 2017 July;89:67–73.
- [25] Gaverina L, Bensalem M, Bedoya A, et al. Constant velocity flying spot for the estimation of in-plane thermal diffusivity on anisotropic materials. *Int J Ther Sci*. 2019 Nov;145:106000. doi: [10.1016/j.ijthermalsci.2019.106000](https://doi.org/10.1016/j.ijthermalsci.2019.106000)
- [26] Ronneberger O, Fischer P, Brox T, et al. U-Net: convolutional networks for biomedical image segmentation. In: Navab Neditor. Medical image computing and Computer-Assisted Intervention – MICCAI 2015, lecture notes in Computer science. Cham: Springer International Publishing; 2015. p. 234–241.
- [27] Wang P. Lucidrains/denoising-diffusion-pytorch, 2022 Oct. original-date: 2020-08-26T02:22:10Z.
- [28] Gulrajani I, Kumar K, Ahmed F, et al. PixelVAE: a latent variable model for natural images. arXiv: 1611.05013 [Cs]. 2016 Nov.
- [29] Chadebec C, Vincent LJ, Allasonniere S. Pythae: unifying generative autoencoders in python - a benchmarking use case. 2022 Oct.
- [30] Kingma DP, Jimmy B. Adam: a method for stochastic optimization. arXiv. 2014 Dec.
- [31] Ruder S. An overview of gradient descent optimization algorithms. arXiv. 2016 Sept.
- [32] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556 [Cs]. 2015 Apr.
- [33] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv: 2010.11929 [Cs]. 2021 Jun.
- [34] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Number: arXiv: 1706.03762 arXiv: 1706.03762 [Cs]. 2017 Dec.
- [35] Raghu M, Unterthiner T, Kornblith S, et al. Do Vision transformers see like convolutional neural networks? 2021; 13.
- [36] Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical Vision transformer using Shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Montreal, QC, Canada: IEEE; 2021 Oct. p. 9992–10002
- [37] Touvron H, Cord M, Sablayrolles A, et al. Going deeper with image transformers. Number: arXiv: 2103.17239 arXiv: 2103.17239 [Cs]. 2021 Apr.
- [38] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, San Diego, CA, USA: IEEE; 2005. p. 886–893.
- [39] Wightman R. Pytorch image models. 2019. Available from: <https://github.com/rwightman/pytorch-image-models>

- [40] David MWP. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv. 2020 Oct.
- [41] Paul S Chen PY, Vision transformers are robust learners. arXiv: 2105.07581 [Cs]. 2021 Dec.
- [42] Ruiz N, Yuanzhen L, Jampani V, et al. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. arXiv: 2208.12242 [Cs]. 2023 Mar.
- [43] Ho J, Saharia C, Chan W, et al. Cascaded diffusion models for high fidelity image generation. arXiv: 2106.15282 [Cs]. 2021 Dec.