Measurement 200 (2022) 111653

Contents lists available at ScienceDirect

Measurement

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

Semantic segmentation for non-destructive testing with step-heating thermography for composite laminates

Oscar D. Pedrayes [a,*], Darío G. Lema [a], Rubén Usamentiaga [a], Pablo Venegas [b], Daniel F. García [a]

[a] Department of Computer Science and Engineering, University of Oviedo, Campus de Viesques, Gijón, 33204, Asturias, Spain
[b] Aeronautical Technologies Centre (CTA), Juan de la Cierva 1, Miñano, 01510, Basque Country, Spain

ARTICLE INFO

MSC:
0000
1111

Keywords:
Quality control
Destructive
Defects
Flaws
Semantic segmentation
Deep learning

ABSTRACT

In this paper, semantic segmentation networks such as UNet and DeepLabV3+ are evaluated and compared against Random Forest and Support Vector Machines in the field of step-heating active infrared thermography for subsurface defect detection and localization. To collect information from an entire digital recording sequence into a particular image, post-processing methods such as PCT, PPT, Kurtosis, Skewness and TSR are used. Two datasets are created, one with 3-channel images using PCT, and one using all the above post-processing methods to condense the heating and cooling processes into 30-channel images. This evaluation study shows that DeepLabV3+ is able to detect most defects in specimens with a similar structure to training samples without false positives even for defects of different depth and area. UNet requires the use of 30-channel images to achieve results closer to DeepLabV3+. Random Forest and Support Vector Machines are unable to compete with the recent methods as they are unable to detect defects correctly.

1. Introduction

Quality control is of significant interest in industry, causing continuous efforts to improve on previous methods. Non-destructive testing (NDT) is a set of analysis methods to inspect, test, and evaluate materials, components, or systems without harming the object. This approach has advantages over destructive testing (DT): it can be used to analyze every item instead of one for each batch and the repeatability of the tests make it possible to repair the products, which leads to lower cost since items need not be replaced after testing. Additionally, since the object is not damaged during the test, NDT tests can also be applied to detect failures as a maintenance method during the lifetime of the product, improving long-term use, and safety [1].

NDT methods can be divided into Contact and Non-Contact. Contact methods are ultrasonic testing, eddy current testing, magnetic testing, and penetrant testing. Non-Contact methods are air-coupled ultrasonic, radiography testing, thermography, shearography, and visual inspection [2].

Today, the most common analysis of defects is done manually by an expert in the field. The results of non-contact NDT inspections, and of some automated contact techniques as well, are generally represented through images. In these situations, the experts usually use image post-processing techniques to make their job faster. Even so, this approach is still costly and time-consuming compared to the potential of a solution based on deep learning.

In recent years, deep learning approaches have made significant advances in the field of infrared thermography [3–7]. Infrared thermography does not need coupling media facilitating the production and speed of scans. This approach has no harmful side effects (such as radiation in X-ray evaluation), improving safety and inspection rates in prolonged use cases. Infrared thermography can be grouped into passive and active. Passive infrared thermography uses the differences in the temperature of the product under natural conditions, that is, without applying heat to the object [8]. Active infrared thermography evaluates temperature differences during and after a heating process. It is important to mention that there is no agreement on the appropriate stimulation and post-processing methodology for a given material and flaw type [9]. This heating process can be done using photographic flashes, halogen lamps, ultrasonic transducers, or other methods [10]. Depending on the method used to heat the item, active infrared thermography can be classified in [6]:

1. Pulsed Thermography (PT): the object is heated for a short time, typically with a flash lamp or a coil for Eddy Current Pulse Thermography [11].
2. Step-Heating Thermography (SHT): the object is heated for longer periods than PT, reaching deeper defects.

* Corresponding author.
E-mail address: UO251056@uniovi.es (O.D. Pedrayes).

https://doi.org/10.1016/j.measurement.2022.111653
Received 11 March 2022; Received in revised form 10 July 2022; Accepted 20 July 2022
Available online 26 July 2022

3. Lock-in Thermography (LT): the object is heated by a modulated heatwave. The temperature changes are compared to the original heatwave revealing defects.

This paper evaluates multiple state-of-the-art methods for image segmentation. Image segmentation is the task of grouping the pixels of an image by creating a segmentation mask. High-level segmentation algorithms generate an easily interpretable classification such as bicycle or road using low-level features, including contrast levels, edges, textures, etc. The most common approaches to image segmentation are shown below:

- Threshold segmentation is the simplest method and consists of classifying pixels with respect to a threshold value.
- Edge-based segmentation is one of the most common approaches. This method identifies edges of different objects in an image using differences in texture, contrast, gray level, color, saturation and other features.
- Region-based segmentation algorithms find groups of pixels by locating seed points. The seed points increase or decrease in size and can merge together to produce different regions.
- Watershed segmentation treats the image as if it were a topographic map. It considers the brightness of a pixel as its height and finds the lines that run along the top of those ridges.
- Clustering algorithms divide the image into clusters of pixels that have similar characteristics. It separates the data elements into clusters where the elements in one cluster are more similar compared to the elements present in other clusters.
- Convolutional networks generate low-level feature maps in an automated fashion. This means that these features do not need to be easily understood by humans. After generating the low-level feature maps, neural networks recognize the relationships between the different features to classify the pixels of the image. Neural networks for the classification of each pixel in an image are known as semantic segmentation networks. If the distinction between multiple instances or objects of the same class is added, it is known as instance segmentation. And if both ideas are combined, so that there are classes without instances and classes with instances, it is known as panoptic segmentation.

Semantic segmentation is one of the most recent methods for image segmentation and has proven to be of great use in other fields such as autonomous driving [12] or crop classification [13]. This approach seems the most suitable given the growing trend for using deep learning models with active infrared thermography [6,7]. Semantic segmentation networks are evaluated for defect localization in composites using step-heating thermography. Given its popularity and the flexibility of its structure to adapt to changes in the required inputs, the semantic segmentation network UNet [14] is used for this evaluation. In addition, the semantic segmentation network DeepLabV3+ [15] is also evaluated, given its more recent and complex architecture. Then, as a basis for comparisons, the older methods Random Forest (RF) and Support Vector Machines (SVM) are used as well.

In a thermographic NDT inspection, the raw results consist of a sequence of thermal images that contain the temperature evolution history of each pixel in the observed scene. There is an explicit limitation in VRAM when using convolutional neural networks, requiring information from every frame about the heating and cooling processes to be compiled into different channels of a particular multichannel image. Post-processing techniques are used to accomplish this: Principal Component Thermography (PCT), Pulsed Phase Thermography (PPT), Kurtosis, Skewness, and Thermographic Signal Reconstruction (TSR). These methods help compile information from the whole sequence into different channels of a particular image and improve the signal-to-noise ratio (SNR). These methods will be discussed in more depth in the "Post-processing methods" subsection. In this way, the network can recognize patterns of all the frames from a sequence simultaneously.

The most obvious advantage of UNet is that it is capable of processing images with more than three channels. In this study, UNet is evaluated with images consisting of 30 channels using the methods described. However, to provide a fair comparison, images with three channels are also tested. This allows for a comparison with another more recent semantic segmentation network DeepLabV3+ [15], and other older methods such as Random Forest (RF) [16] and Support Vector Machines (SVM) [17].

Recent works tend to use simple or manually created convolutional network architectures [18,19] and older methods for object detection such as FasterRCNN [20]. There are few papers that use more recent, complex semantic segmentation architectures [4,21]. Those that do use a more modern architecture typically use UNet or one of its variations, but it appears that none of them explore the use of more than three channels against the use of only three channels per image. Moreover, the use of DeepLabV3+ in the field of defect detection is scarce [22], and to the best of the authors' knowledge, non-existent on the subject of subsurface defect detection.

The composite laminate evaluated in this work is a carbon-fiber-reinforced polymer (CFRP) laminate. Known for its strength-to-weight ratio and rigidity, it is often used in aircraft, cars, or bicycle frames [23]. NDT methods are preferred for CFRP since this material is costly, and an impact can create delamination inside the material, provoking subsurface damage invisible on the surface. Creating large datasets in NDT thermography is a costly and time-consuming process. For this reason, many papers use only a few specimens for their studies [24–26], so approaches that do not require large datasets (as is the case with UNet), or the need to use a pretrained model (as is the case with DeepLabV3+), are required. In this study, only one specimen is used to generate the datasets for training. By rotating the specimen 10°, up to 36 different digital recordings are generated with non-repeating data. Since the specimen has a different illumination and background for each digital recording, and it has to be heated and cooled again, the resulting data can be considered new and non-repeating, unlike other methods that consist of rotating the images. To further validate the trained models, another two new specimens on which to perform the testing are added.

The dataset containing the training and testing samples for semantic segmentation is described in the "Dataset" subsection and is released for public usage in the following DOI: https://doi.org/10.5281/zenodo.5426792.

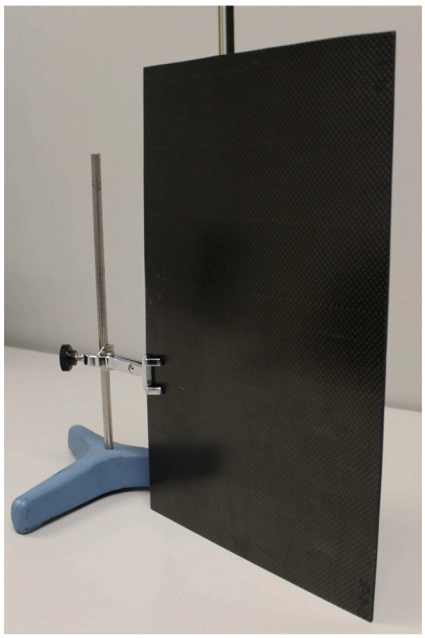
2. Materials and methods

2.1. Carbon-fiber-reinforced polymer laminate

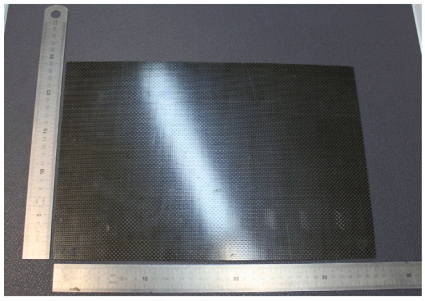
CFRP is a composite material composed of a reinforced carbon fiber, and a matrix to bind the reinforcements together. Fig. 1 shows: a general photograph of the specimen to be used for training (Fig. 1(a)); a photograph showing its measurements (Fig. 1(b)); and a photograph showing the location of the defects (Fig. 1(c)). The 360 mm × 240 mm specimen is 2.5 mm thick, following a 12-ply structure as seen in Fig. 2.

The specimen has artificially induced flaws. In this specimen there are two different types of defects: Polytetrafluoroethylene (PTFE) thin films, and steel chips defects. There are 12 defects, 9 are PTFE thin films and 3 are steel chips defects. The PTFE films simulate delaminations, which are common defects in composite materials produced by the separation of adjacent plies, while the steel inserts simulate accidental inclusion of small pieces of cutting tools used during the manufacturing process of the material.

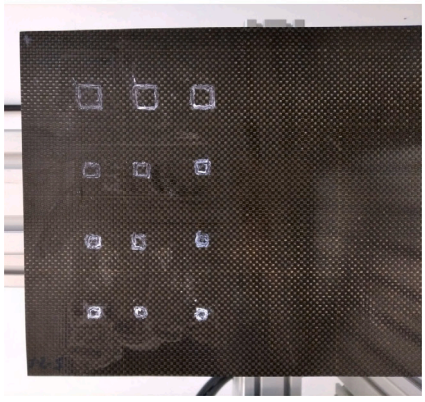
There are three different sizes of PTFE (12 mm×12 mm, 7 mm×7 mm, and 5 mm×5 mm) each at 3 different depths (0.63 mm, 1.46 mm, and 2.08 mm), and only one size of steel chip (5 mm×5 mm) located at 0.63 mm, 1.46 mm, and 2.08 mm. The bottom three defects are steel chips defects and the rest are PTFE defects. The height of the defects was measured with a calibrated caliber obtaining a value of



(a) Image



(b) Measurements



(c) Defects

Fig. 1. Photographs of the specimen.

0.06 mm. In Fig. 2, the location of the defects in the specimen is shown. The depths and layers of the defects are presented from shallowest to deepest, from left to right: the first column of defects has a depth of 0.63 mm, the second column 1.46 mm and the third column 2.08 mm.

A greater surface area of the defect implies a greater heat flow affected by the presence of the defect; and consequently, it implies a greater variation of temperatures in the areas near the defect. On the other hand, the shallower the depth, the lower the lateral heat dissipation effect. As a consequence of the presence of a defect, the heat flow towards the surface will be less degraded, making the thermal effect on the surface more evident.

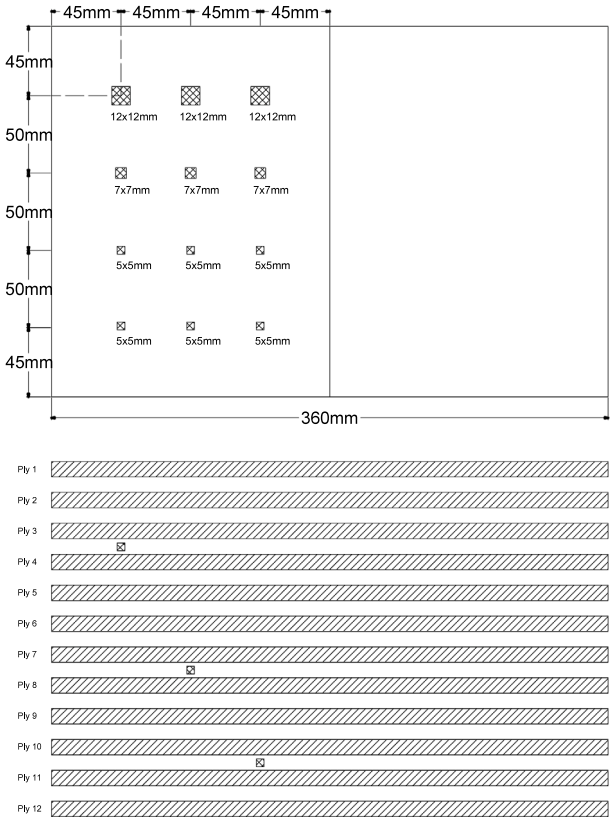


Fig. 2. Arrangement of the defects and dimensions of the specimen.

2.2. Infrared thermography using step heating

The CFRP laminate is heated using two halogen lamps (eurolite PAR-64 Profi floorspot model of 1 000 W) for ten seconds. After the ten seconds, the two lamps are turned off to let the object cool down for another ten seconds. This process is recorded (using an IR detector NETD of less than 55 mK, and optics of 25 mm F/1 lenses) for a total of twenty seconds at 50 FPS at a resolution of 640 × 480 pixels, resulting in a total of 1,000 frames for each digital recording. The camera used to record the digital recordings is a Xenix Gobi 640 GigE model with a spectral range between 8-14 μm and a pixel resolution of 480 x 640. In Fig. 3 a diagram of the setup for the recordings with the location, distance and angle of the infrared camera, halogen lamps, and specimen is shown.

The heating time necessary to reveal the presence of defects was roughly defined by numerical simulation in a preliminary stage, and subsequently, the definite heating time was verified by experimental assessment. This time span produced the maximum number of defects to be detected preventing the sample from overheating.

Subsurface defects heat and cool down at different rates than the rest of the object. The active stimulation is applied to exploit this feature as a way to obtain the maximum possible contrast between the defects and the rest of the object. For each digital recording, the CFRP laminate is cooled down to room temperature before the process starts, to avoid heating the object at different temperatures.

Fig. 4 shows this heating and cooling process for a pixel with defect and a nearby pixel without defect (see Fig. 5). In addition, these same reference points are added but with the specimen rotated 120° (see Fig. 6). No absolute values are needed, only the differences between

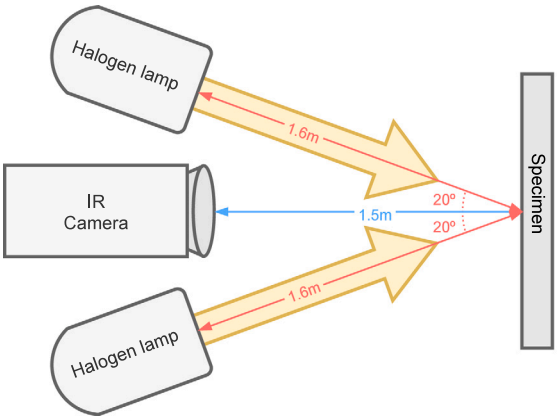


Fig. 3. Diagram of the setup of the recordings.

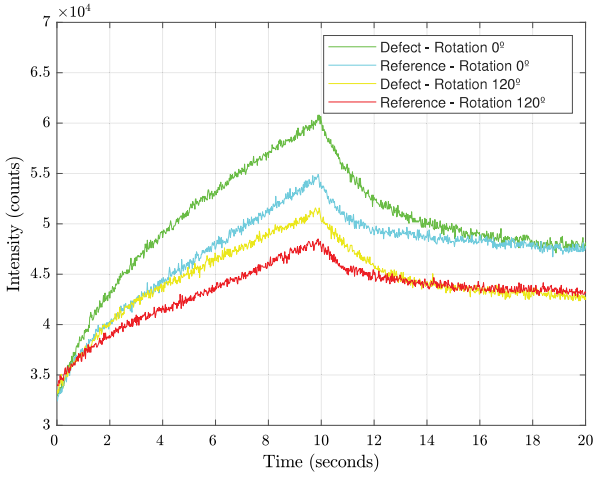


Fig. 4. Heating and cooling signal intensities for the time sequence of the CRFP laminate. Signal color correspond to those of Figs. 5 and 6.

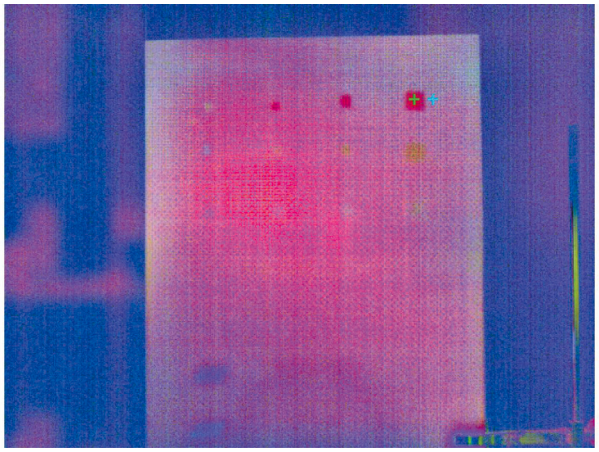


Fig. 5. Pixels used for reference in Fig. 4 for the specimen rotated 0°.

nearby pixels are required to locate the defects. When rotating the specimen it can be observed that the response is not the same. This is because the thermal energy is not transmitted uniformly throughout the specimen, which is of great interest in order to create a dataset that allows the network to generalize.

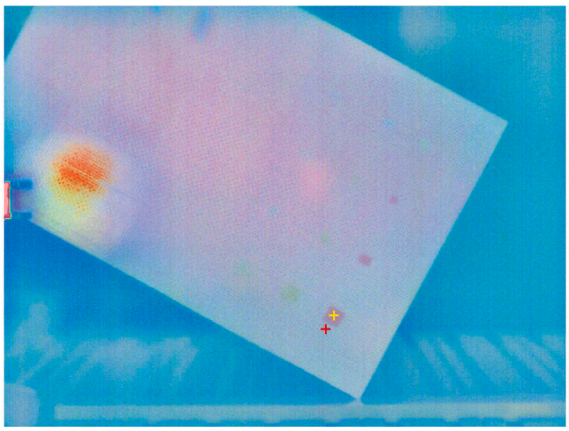


Fig. 6. Pixels used for reference in Fig. 4 for the specimen rotated 120°.

2.3. Post-processing methods

Image post-processing methods are used to summarize the information of a full digital recording into a particular multichannel image. This data compression is necessary to be able to use UNet and DeepLabV3+ due to their computational cost. To study the effect of this compression, two different approaches are evaluated.

The first approach converts the heating process of the digital recording into an image with 3 different channels using only the Principal Component Thermography (PCT) [27] method. This approach is tested with all the methods (Random Forest, Support Vector Machines, UNet, and DeepLabV3+).

The second approach takes advantage of both the heating and cooling sequences and uses 15 channels for each, resulting in images with 30-channel. Each channel stores post-processed images generated by the following methods: PCT [28], PPT [29,30], Kurtosis [31], Skewness [32] and TSR [33,34], as detailed in their respective subsections. This approach can only be tested with UNet.

2.3.1. Principal Component Thermography (PCT)

PCT is applied to each pixel time history, calculating a linear transformation to the initial data from the eigenvectors of the associated covariance matrix. Using this method a distinction between defect and non-defect is more easily visible.

In this study, for the 3-channel images, each post-processed image corresponds to components 1st, 3rd and 4th of the PCT of the heating sequence (500 frames). The second component is not used for the 3-channel images since the signal to noise ratio is higher in the 3rd and 4th components [23].

For the 30-channel images, the first four channels of the PCT are used in both the heating (500 frames) and cooling (500 frames) sequences separately, thus giving a total of eight channels.

2.3.2. Pulsed Phase Thermography (PPT)

PPT is a method to calculate the phase of thermographic data per pixel time history based in the Discrete Fourier Transform (DFT) algorithm [35]. The DFT algorithm is usually used in image post-processing to filter out periodic noise. It can be used to obtain an image that only represents the edges. For the 30-channel images, the phase of the minimum frequency is used, obtaining a post-processed image for the heating process and another for the cooling process. Eq. (1) is used to calculate PPT, where T is the temperature, n the frequency increment, N the number of frames, i the imaginary number, Re_n is the real part of the DFT, and Im_n the imaginary one.

$$F_n = \sum_{t=1}^{N-1} T(t) e^{\frac{2\pi i t n}{N}} = Re_n + Im_n \tag{1}$$

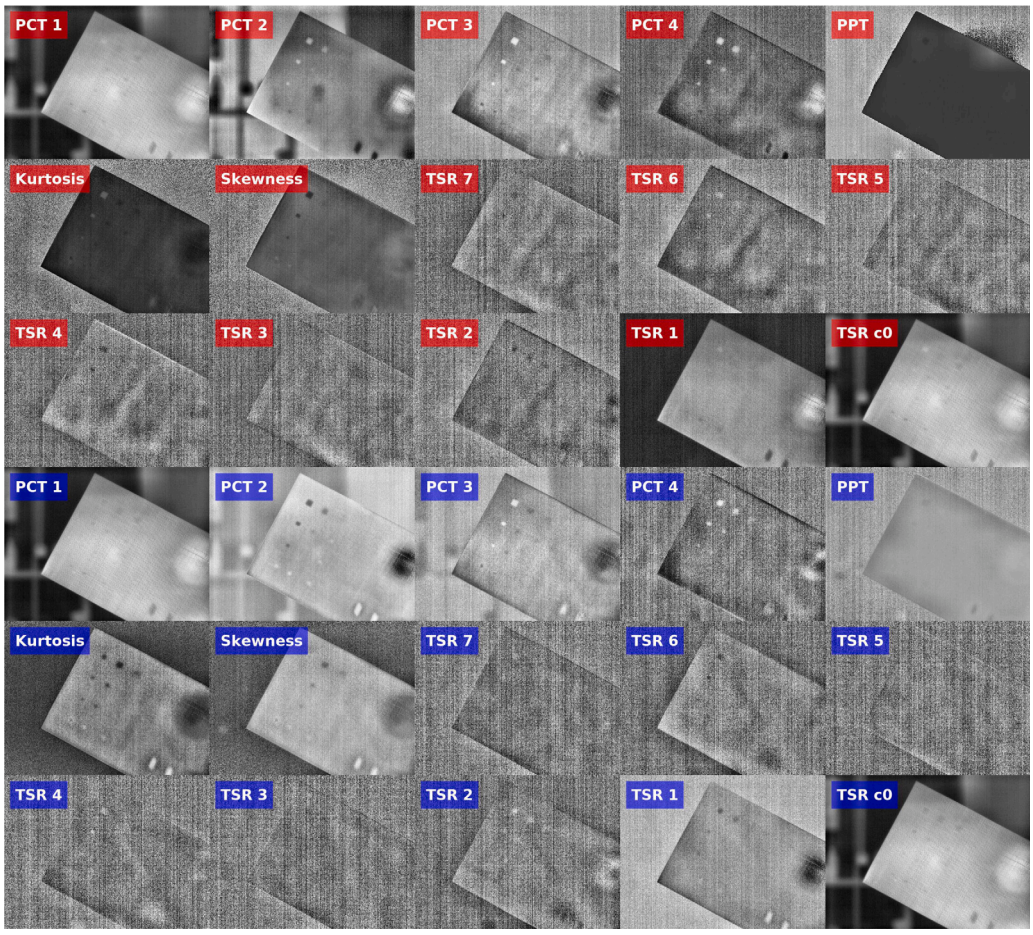


Fig. 7. Example of a 30-channel images. Images with a red label are obtained from the heating sequence. Images with a blue label are obtained from the cooling sequence..

Eq. (2) is used to calculate the phase.

$$\phi = atan\left(\frac{Im_n}{Re_n}\right) \tag{2}$$

2.3.3. Kurtosis

Kurtosis measures the degree of peakedness of a distribution. If the distribution is the same as the normal distribution it has a value of zero, if it is higher it has a positive value, and if it is lower a negative value. In this case this measure is calculated per pixel time history using the heating and cooling sequences, obtaining two channels for the 30-channel images.

Eq. (3) is used to calculate Kurtosis. T is the temperature data from the pixel time history, $\bar{T}$ is the mean of the temperature data, s the standard deviation, and N the number of frames.

$$Kurtosis = \frac{\sum_{i=1}^{N}(T_i - \bar{T})^4/N}{s^4} \tag{3}$$

2.3.4. Skewness

Skewness measures the lack of symmetry. A positive skew means that the longest tail of the distribution is at the right of the histogram and the reverse for the negative skew. A distribution that is fully symmetric has a value of zero. The skewness is calculated per pixel using every frame in the heating process or cooling process. For the 30-channel images, this results in two channels, one for the heating process and another for the cooling process.

Eq. (4) is used to calculate Skewness. T is the temperature data from the pixel time history, $\bar{T}$ is the mean of the temperature data, s the standard deviation, and N the number of frames.

$$Skewness = \frac{\sum_{i=1}^{N}(T_i - \bar{T})^3/N}{s^3} \tag{4}$$

2.3.5. Polynomial fit

Polynomial fit, also known as Thermographic Signal Reconstruction (TSR) when calculated using logarithmic expressions, is a method for estimating thermal diffusivity by removing noise from a thermal signal based on a sequence. This method is calculated per pixel time history and is commonly used for defect detection. It is considered that a degree of 7 generally provides optimal results for defect detection in laminates [36]. This generates one post-processed image for the coefficients of each degree plus its coefficient zero. Taking this into consideration, for the 30-channel images, eight channels for the heating process and another eight channels for the cooling process are created.

Eq. (5) is used to calculate the Polynomial fit. T is the temperature pixel time history, n is the degree and t is the time or frame of the thermogram.

$$T(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n \tag{5}$$

2.4. Dataset

Using the method described in Section 2.2, 36 digital recordings are generated. All the digital recordings record the same CFRP laminate using different rotations, which alters lighting, lamp reflections and background among other things. This process is done to obtain more data for training and to improve variability. For each digital recording the CFRP laminate is rotated 10°. From each digital recording, two images (one with 3 channels and the other one with 30 channels) are generated using the post-processing methods mentioned in Section 2.3. In Fig. 7 an example of every post-processed image from one of the 30-channel images is shown. The objective of this study is not visualization

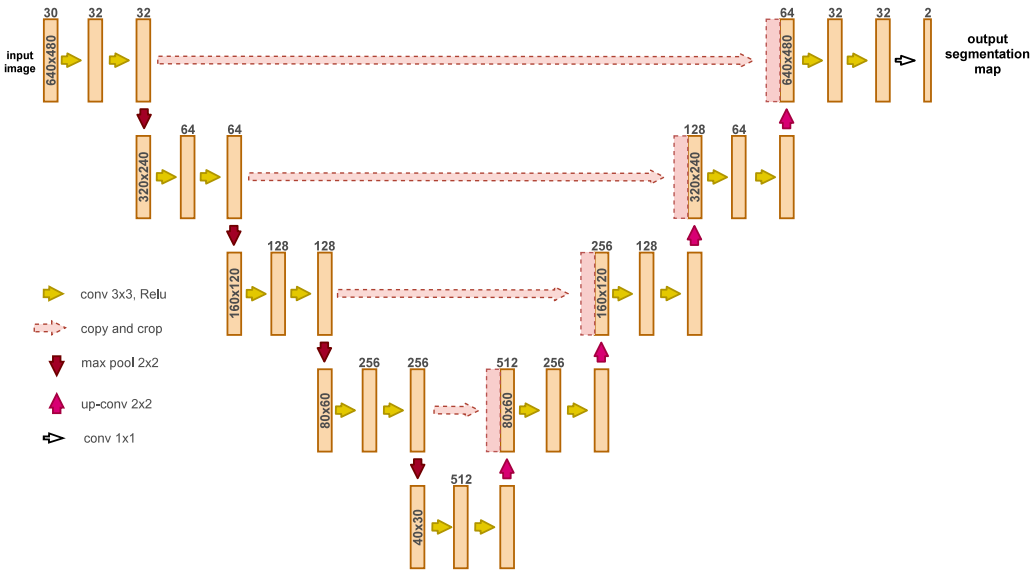


Fig. 8. The UNet architecture used for optimal results. (This graphic is inspired from the UNet architecture paper [14]).

but detection and localization. Fig. 7 merely provides an understanding of the inputs to be fed into the neural networks.

Ground truth masks are generated by experts in the field, who verify that the defects are correctly classified.

Two datasets are created, one that uses images with 3 channels and another that uses images with 30 channels. In this way, a comparison can be made to determine if more information results in better accuracy. Each dataset consists of 36 images. The first 30 are used for training and the last 6 for testing and visualization. Both datasets can be found at the following DOI: https://doi.org/10.5281/zenodo.5426792.

Classes are divided in "defect" and "other". The objective is binary classification so the "defect" class is the target class, and the "other" class is the non-target class that refers to everything else, including the rest of the specimen and the background of the digital recording.

2.5. Analysis of the evaluated architectures

2.5.1. Unet

UNet is one of the first and most referenced networks in semantic segmentation with over 29,000 cites of its original paper in Google Scholar. Its original purpose was for binary classification to segment cells in biomedical imagery and to train and produce precise predictions with as few training images as possible [14]. The name "UNet" comes from its u-shaped architecture as the result of a symmetric encoder–decoder. UNet was quickly adapted to work with all kinds of imagery and class number as it offers a high degree of flexibility thanks to its simple layout. This has caused the rapid development of new variations. An overview of the UNet architecture used in this evaluation study can be seen in Fig. 8.

2.5.2. Deeplab

DeepLab is a semantic segmentation architecture made by Google. DeepLabV1 [37] presents atrous convolutions to tune the resolution at which features are calculated. DeepLabV2 [38] details Atrous Spatial Pyramid Pooling (known as ASPP) to increase the accuracy of predictions at different scales. DeepLabV3 [39], tunes the ASPP module and uses a Batch Normalization module to simplify the setup of the data eliminating the need for a manual normalization. DeepLabV3+ [15] is the fourth and most recent version of DeepLab. It converts its architecture to a encoder–decoder architecture. There is an auto machine learning version called AutoDeepLab [40] which is based on the DeepLabV3+ architecture. An overview of the DeepLabV3+ architecture used in this work can be seen in Fig. 9. The DCNN module is the backbone network used and it usually is a variation of ResNet, Xception or MobileNet.

2.6. Network parameters

This section provides a brief description of the network parameters used to modify the architectures.

UNet and DeepLabV3+ have some common network parameters: the input size, which controls the resolution and channels of the input images, the number of classes to use in the experiment, and the use of padding to fill each convolution to keep the resolution of the final feature map the same size as the input.

UNet has two controllable network specific parameters consisting of the depth of the architecture, which is based on the number of max pooling layers, and the number of filters at each level, which is controlled by the number of filters at the first level and then multiplied by two at each level.

DeepLabV3+ has a specific controllable network parameter called output stride. This parameter controls the separation between each step of a convolution. It is calculated as the result of the division between the input image resolution and the final feature map. For example, an input image that has a resolution of 512×256 pixels and a final feature map of 32×16 would result in an output stride of 16.

In all cases, the initialization of the convolutional filter weights follows the Kaiming He et al. [41] algorithm.

2.7. Training parameters

This section lists the training parameters used to train the models. Optimal training parameters are re-evaluated for every change in the network parameters described in Section 2.6.

As a first step, the optimal batch size, learning rate and number of epochs are investigated. This process is repeated for each solving algorithm available, which in this case are Adam or Stochastic Gradient Descent with Momentum (SGDM). On the other hand, the value of the L2 regularization is studied separately, to apply a penalty to the loss function in order to decrease the complexity of the model and reduce overfitting.

Then, if the Precision and Recall metrics are unbalanced, different class balancing approaches are evaluated. Methods such as inverse frequency weighting (IFW), mean frequency weighting (MFW) and manually chosen custom weights are evaluated.

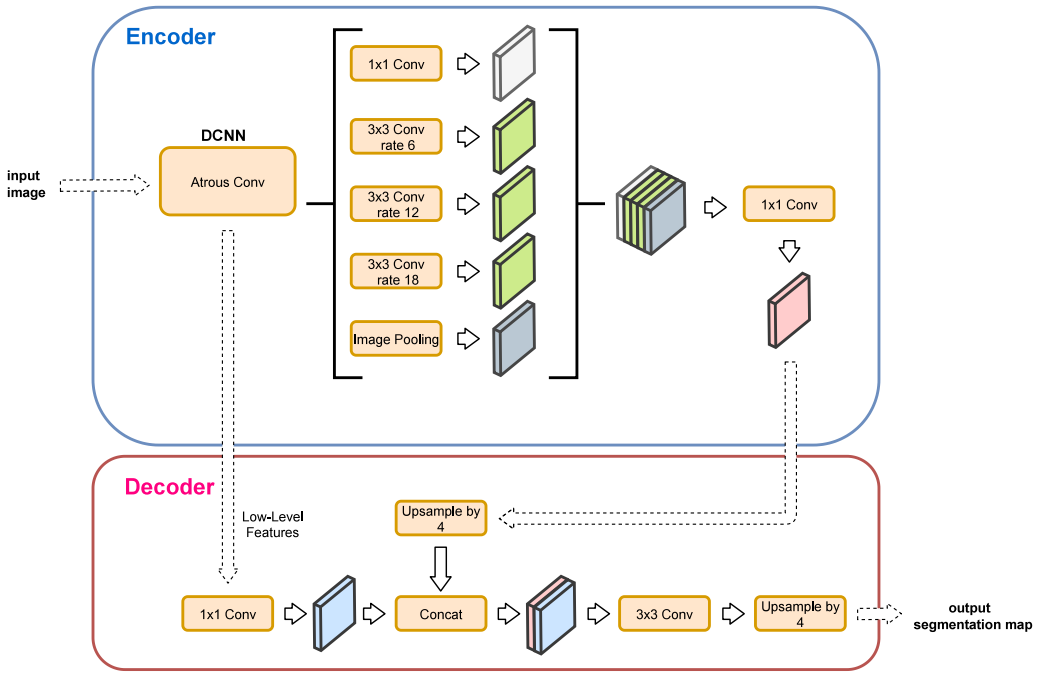


Fig. 9. The DeepLabV3+ architecture used in the experiments. This graphic is inspired from the DeepLabV3+ paper [15].

In this work, the use of a gradient clipping value, to constrain the maximum possible value of the gradient, is not necessary, since the exploding gradient problem is not present in the training process.

Finally, to add new data samples, improve variability and reduce overfitting, data augmentation methods are applied to the training set. These methods consist of enlarging or flipping the images. In addition, the dataset is shuffled before each epoch to minimize overfitting.

2.8. Performance metrics

This section provides a brief description of the metrics used [42] to evaluate the performance of the trained models.

- True positive (TP): correctly classified pixels.
- True negative (TN): pixels correctly classified as belonging to other classes.
- False positive (FP): pixels classified wrongly as the target class.
- False negative (FN): pixels wrongly classified as belonging to other classes.
- Precision (P): Percentage of correctly classified pixels from the total number of predictions for a particular class.

$$P = \frac{TP}{TP + FP} \tag{6}$$

- Recall (R): Percentage of correctly classified pixels from the total number of pixels for a particular class.

$$R = \frac{TP}{TP + FN} \tag{7}$$

- F-score (F_1): Value that combines Precision and Recall making it easier to compare models. A good model should have a balance between Precision and Recall. This metric should not be used alone as it does not indicate whether the two metrics are balanced. This metric is equivalent to the Dice Coefficient with two classes.

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{8}$$

- Intersection-Over-Union (IoU): Value that measures the similarity between ground truth and prediction. This metric is equivalent to the Jaccard Index.

$$IoU = \frac{Area\,of\,Overlap}{Area\,of\,Union} = \frac{TP}{TP + FN + FP} \tag{9}$$

2.9. Training procedure

The network and training parameters must be tuned to reach the best possible results. In this evaluation, these hyperparameters are calibrated manually for each network, obtaining the best configuration for each parameter one by one. The effects of changing multiple parameters at the same time has not been studied in depth. However, a manual process that would research every combination of parameters is not feasible. In this regard, there is still leeway to improve results but the time required is far too great for a small improvement in accuracy.

To obtain realistic results, the datasets are divided in training and testing images. From the total of 36 images, the first 30 are used for training and the last 6 for testing and visualization.

To select the best experiments both Precision and Recall are evaluated. When both metrics are high and are balanced it is considered as a good result. If they are unbalanced, the accuracy of the model is compromised. A high Precision and a low Recall means that the model is predicting few pixels but those that are predicted are correct. If the Recall is high and the Precision low, it means that the model is predicting more pixels than there are in the ground truth. Only the metrics of the target class are provided because the non-target class is irrelevant.

To offer a better representation and facilitate the understanding of the metrics, visualization examples of the six testing images are provided for the best experiment of each architecture. This can help to give a better idea of how the model is predicting the defects.

The hardware used to train the models of the experiments consist of a GPU NVIDIA RTX 2080 Ti and a I7-9700 K CPU.

3. Results and discussions

3.1. Random forest and support vector machines

Random Forest runs several decision tree algorithms. Each decision tree gives a classification and the choice with the most “votes” is the

Table 1
Metrics for the experiments with Random Forest and Support Vector Machines.

Method	Precision	Recall	IoU	F_1
RF	.103	.132	.061	.116
SVM	.037	.604	.036	.069

final prediction. Support Vector Machines search for a hyperplane with the widest margin between the two classes that best separates two different classes of data points.

Experiments are carried out with Random Forest and Support Vector Machines as a basis for making comparisons. Both use the same feature vector, which is calculated using thirteen features. The first three features consist of the red, green and blue (RGB) values of each pixel of the input image, which correspond to the first, third and fourth components of PCT.

The fourth feature is the local binary pattern (LBP), a texture descriptor used in computer vision, calculated by thresholding the neighborhood of every pixel into a binary number using a 3 × 3 grid, and converting the result to a decimal number [43]. To calculate the neighborhood components, a radius of 24 is used totaling in 192 neighbors. LBP is applied to a gray scale version of the 3-channel images in order to obtain more spatial context. LBP is calculated with Eq. (10), where P is the total number of neighbors, R the radius, c is to the central pixel and g is the value of a pixel.

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)^{2^p} \text{ with } s(x) = \begin{cases} 1, & \text{if } x \geq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The last nine features consist of multiple Haralick texture features, which are texture descriptors used in computer vision for image classification. All Haralick features are based on the gray-level co-occurrence matrix which shows the frequency at which each gray level occurs in a pixel at a fixed geometric location with respect to other pixels. The features used are: angular second moment, contrast, correlation, sum of square: variance, inverse difference moment, sum average, sum entropy, and entropy. Equations for all the Haralick texture features are in [44]. Haralick features are applied to the gray scale version of the 3-channel images in order to obtain more spatial context.

A subsampling to the feature vector of each image is done in order to reduce training time and memory usage. This generates 1,000 observations per image with 13 features per observation. With 30 images to train, 30,000 observations are used for training.

For Random Forest the number of estimators and their maximum depth is manually optimized. In addition, different class weights are tested for both Random Forest and Support Vector Machines. The optimal experiments for each methods are listed in Table 1. The best Random Forest experiment uses 1000 estimators and a maximum depth of 10. In the case of Support Vector Machines a radial basis function kernel is used, and the gamma value is calculated as the inverse of the multiplication of the number of features by the variance. In both cases the class weight for the non-target class and for the target or defect class is balanced using the proportional inverse of the class frequencies. Results from these experiments can be seen in Table 1.

According to Table 1, both experiments obtain low metrics: below 12% in F_1-Score. To prove that these values are too low, visualizations of Random Forest and Support Vector Machines are shown in Figs. 10 and 11 respectively.

In SVM, the edges of the specimen are classified as defects, this is due to the great variance between the specimen and the background. In RF, although this can be observed in some cases, it is much less obvious. Moreover, both models predict many more pixels from the most shallow defects as these are the ones with the most variance.

There is a circular area detected at the bottom of both RF and SVM. This area is the reflection of the heating lamps. By observing Fig. 11, it is clear that SVM is very sensitive to these artifacts, more so than RF. These reflections can be avoided by positioning the camera properly, although this is not always possible in real inspections due to lack of space. It is very common to find reflections in real inspections. Therefore, it seems reasonable to include them in the study and analyze the robustness of the models in their presence. Although the effects can be minimized using Plexiglas filters.

Theoretically, better results could be achieved by improving the feature vector. The selection of features has the most significant impact on how well these methods perform. However, in this study, common features for image segmentation are used [43,44].

Table 2
Network parameters for UNet.

Network parameters		
Parameter	3-channel images	30-channel images
Input size	640 × 480 × 3	640 × 480 × 30
Classes	2	2
Depth	4	4
Filters on first level	32	32
Padding	Yes	Yes

Table 3
Training parameters for UNet.

Training parameters		
Parameter	3-channel images	30-channel images
Solver	Adam	Adam
Epochs	1000	1000
Batch size	8	4
Learning rate	0.001	0.001
Class weighting	0.35–0.65	0.20–0.80
Gradient clipping	No	No
L2 regularization	0.0001	0.0001
Data augmentation	Mirror in X/Y	Mirror in X/Y
Shuffle	Yes	Yes

Table 4
Metrics for the experiments with UNet.

Experiment	Precision	Recall	IoU	F_1
3-channel images	.689	.717	.542	.703
30-channel images	.764	.726	.593	.745

3.2. UNet

This section presents the optimal segmentation results with UNet for 3-channel images and 30-channel images. Both experiments have the same optimal hyperparameters (Tables 2 and 3) with the exception of batch size and class weights. Since the images with 3 channels take less memory than the images with 30 channels, the maximum batch size can be increased from four to eight images. In the case of the class weights, the optimal weights differ between datasets from a value of 0.65 to 0.80.

The depth of the UNet architecture coincides with the original implementation but the number of filters on the first level has been reduced by two. This affects the whole architecture dividing the numbers of filters by two. Using fewer filters means faster training times and increased batch sizes. There is no need for gradient clipping since there is no exploding gradient problem. L2 regularization works best when using the 0.0001 default. All the training data is shuffled before every epoch to prevent overfitting.

The metrics from the testing of the "30-channel images" and "3-channel images" experiments can be seen in Table 4.

Table 4 shows a great difference between using 3 and 30 channels. In this case the 30-channel images experiment has an almost 5% higher F_1-Score. Both experiments surpass 70% in F_1-Score and obtain a balance between Precision and Recall.

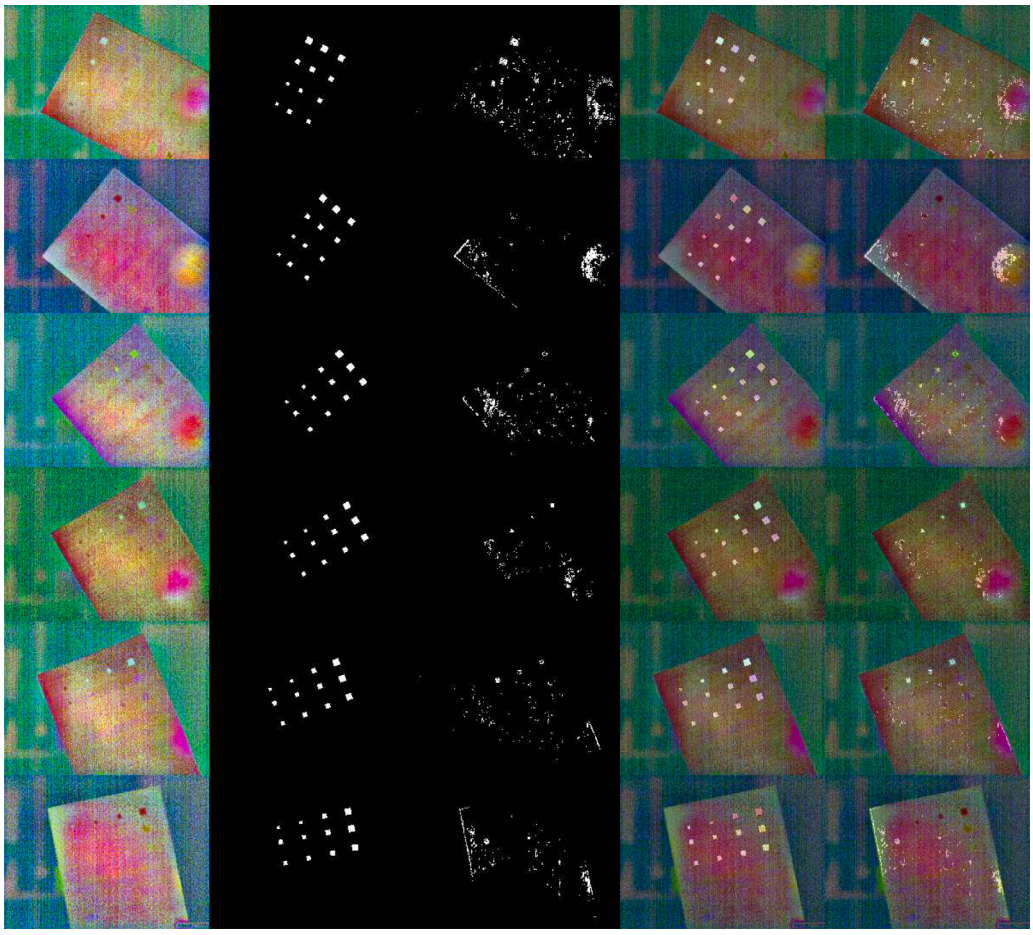


Fig. 10. Visualization of the predicted results for Random Forest. (**1st col.**) Original images, (**2nd col.**) ground truth masks, (**3rd col.**) predictions, (**4th col.**) original images and ground truth masks, (**5th col.**) original images with predictions.

The 3-channel images experiment takes 00 h:31 m:31 s to train with the specified hardware, whereas the 30-channel images experiment takes 02 h:15 m:15 s. The extra channels make the architecture more complex.

To accompany these results a visualization of the testing images can be seen in Figs. 12 and 13. In these figures a great difference between models can be seen. The experiment for 3-channel images (Fig. 12) detects all defects although the ones with more depth have a much smaller area than the ground truth and there are some false positives. Fig. 13 has much less noise but it has trouble detecting all the defects in some of the images.

3.3. DeepLabV3+

This section presents the best experiment with DeepLabV3+ with 3-channel images. Table 5 shows the network architecture parameters. In this case the backbone architecture that performs the best is Xception65. Xception71 has more layers and therefore more VRAM is needed for the same batch size. A smaller batch size, even when using a network with more layers, performs worse. For this same reason, an output stride of 16 is preferred.

Table 6 shows the training parameters. In this case, DeepLabV3+ has a more complex architecture than UNet so the maximum batch size possible for eleven gigabytes of VRAM is four images. The value of the optimal class weight for the 3-channel images is the same as UNet. There is no need for gradient clipping since there is no exploding gradient problem. Furthermore, this architecture works better with a smaller learning rate than UNet. The best L2 regularization value coincides with that recommended by the developers. The whole training set is shuffled before every epoch to prevent overfitting.

Table 5
Network parameters for DeepLabV3+.

Network parameters	
Input size	640 × 480 × 3
Classes	2
Backbone	Xception65
Output stride	16
Padding	Yes

Table 6
Training parameters for DeepLabV3+.

Training parameters	
Solver	Adam
Epochs	1000
Batch size	4
Learning rate	0.0005
Class weighting	0.35–0.65
Gradient clipping	No
L2 regularization	0.00004
Data augmentation	Scale 0.5–2.0 with 0.25 steps
Shuffle	Yes

To achieve faster training times and allow the model to generalize better, the training starts from a pre-trained model on the ImageNet dataset [45].

The metrics from the testing of the "3-channel images" experiment can be seen in Table 7.

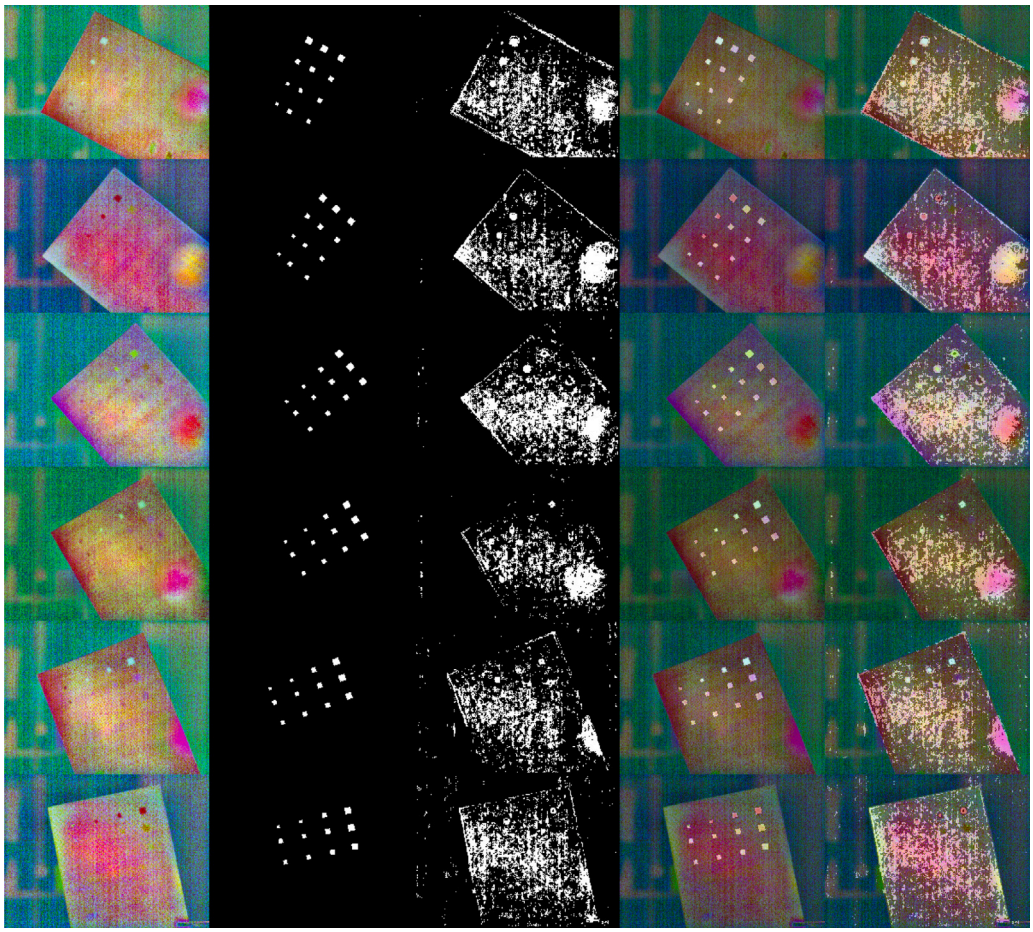


Fig. 11. Visualization of the predicted results for Support Vector Machines. (**1st col.**) Original images, (**2nd col.**) ground truth masks, (**3rd col.**) predictions, (**4th col.**) original images with ground truth masks, (**5th col.**) original images with predictions.

Table 7
Metrics for the experiment with DeepLabV3+.

Experiment	Precision	Recall	IoU	F_1
3-channel images	.760	.786	.629	.773

In Table 7 the results show high values for the metrics, above 77% in F_1-Score and with Precision and Recall balanced. This experiment obtains even better results than the 30-channel images experiment with UNet, which is impressive given the difference between the 30-channel images and 3-channel images experiments in UNet.

This experiment with DeepLabV3+ takes 01 h:18 m:27 s, more than 2.5 times longer than UNet under the same conditions. However, is still almost two times faster than the 30-channel experiment with UNet.

To accompany these results a visualization of the testing images can be seen in Fig. 14. In these figures a great difference between models is observed with respect to those of UNet. This model detects almost every defect and has virtually no noise. It has most trouble detecting 5 mm×5 mm defects at maximum depth. However, in the majority of the testing images all the defects are found.

3.4. Discussion

Neither Random Forest nor Support Vector Machines can detect defects in CFRP laminates using the image post-processing methods described. The metrics (Table 8 and Fig. 15) and visualization images (Figs. 10, 11, 12, 13 and 14) make it clear that these methods are not reliable enough, at least with the features selected, to detect defects with high confidence. They do not generalize well enough. Thermographic data generally has high levels of noise and low levels of contrast. These characteristics give high variance to the features for the same defect, making them difficult to detect for conventional models such as RF and SVM.

In the case of UNet, results are much improved. With the 3-channel images the metrics might be considered low. However, the defects are all distinguishable in the visualization images although there is some noise in the predictions. When it comes to the 30-channel images, the result metrics show a clear improvement. The noise of predictions is vastly reduced and the visualization images show that almost all the defects are found.

DeepLabV3+ performs better than UNet even when only the 3-channel images can be used. This evaluation provides the best results, nearing 80% of F_1-Score. The visualization images are clearer than those produced by UNet and almost all of the defects are found. DeepLabV3+ is more computationally complex than UNet under the same conditions, requiring more than twice as much training time. However, DeepLabV3+ is still almost twice as fast as UNet with 30 channels.

For semantic segmentation models, unlike SVM and RF models, lamp reflection is not classified as a defect. This is desirable in real inspections, where reflections are often unavoidable. This indicates that the manually created features are not enough to learn that the reflection is not a distinguishing feature of the defects. However, UNet and DeepLabV3+ are able to “learn” that the reflections are not a distinct part of the defects. This is possible because by rotating the specimen, the reflection is not always in the same part of the specimen.

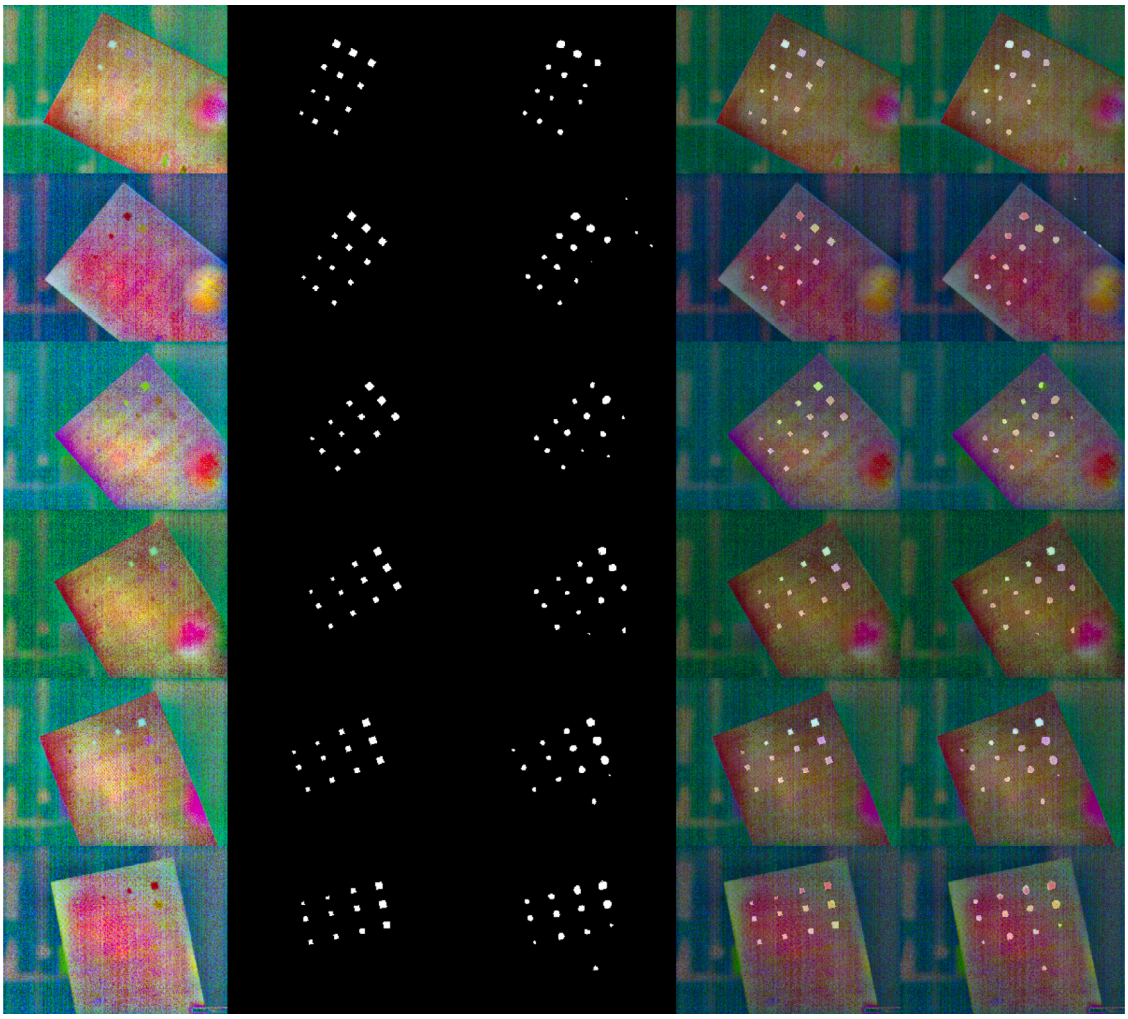


Fig. 12. Visualization of the predicted results for UNet evaluated with 3 channels. (**1st col.**) Original images, (**2nd col.**) ground truth masks, (**3rd col.**) predictions, (**4th col.**) original images with ground truth masks, (**5th col.**) original images with predictions.

Table 8
Metrics for all the methods.

Experiment	Precision	Recall	IoU	F_1
RF (3-channel)	.103	.132	.061	.116
SVM (3-channel)	.037	.604	.036	.069
UNet (3-channel)	.689	.717	.542	.703
UNet (30-channel)	.764	.726	.593	.745
DeepLabV3+ (3-channel)	.760	.786	.629	.773

4. Other samples

This section evaluates new specimens with different internal structures. The objective of these evaluations is to observe how far the semantic segmentation models generalize. For this purpose, the predictions of these specimens are run with the previous models, trained with the specimen presented in Section 2.1.

The defects of these specimens are artificially generated, however, they may be slightly offset from the original scheme. For this reason, an ultrasonic inspection is performed to find and check the real positions of the defects. The ground truth of these two new specimens is generated by a manual procedure. First, a probe is passed over the surface of the specimen, scanning the signal it receives in a similar way to an oscilloscope. In this way, it is possible to detect signal changes that are indicative of a defect. This defective area is marked with a pencil on the specimen itself. Finally, using the thermographic image and the RGB image in which the pencil marks can be observed, a ground truth mask is generated manually by observing and overlapping both images.

Table 9
Metrics for specimen 2.

Experiment	Precision	Recall
UNet (3-channel images)	.56	.47
DeepLabV3+ (3-channel images)	.83	.41

The first specimen has a similar structure to the training specimen. However, this specimen has half the depth (1.125 mm) and a smaller number of layers (6 plies). In this case, the depth of the defects is 0.75 mm, 0.56 mm, and 0.19 mm from left to right. The top three defects are steel chips defects and the rest are PTFE defects. (See Fig. 16).

As can be seen in Fig. 17, all the defects of the part are successfully detected with DeepLabV3+. It appears that the smaller defects have a predicted area greater than the area of the ground truth defects. UNet is also able to detect almost all of the defect but the predicted image has more noise. In Table 9 metrics for this evaluation are obtained. These metrics present lower precision than expected due to this increase in the area of small defects.

The second specimen has a very different structure from the training specimen. This specimen not only has greater depth (20 plies and a total depth of 3.825 mm), but also, ply 7 is of greater depth and

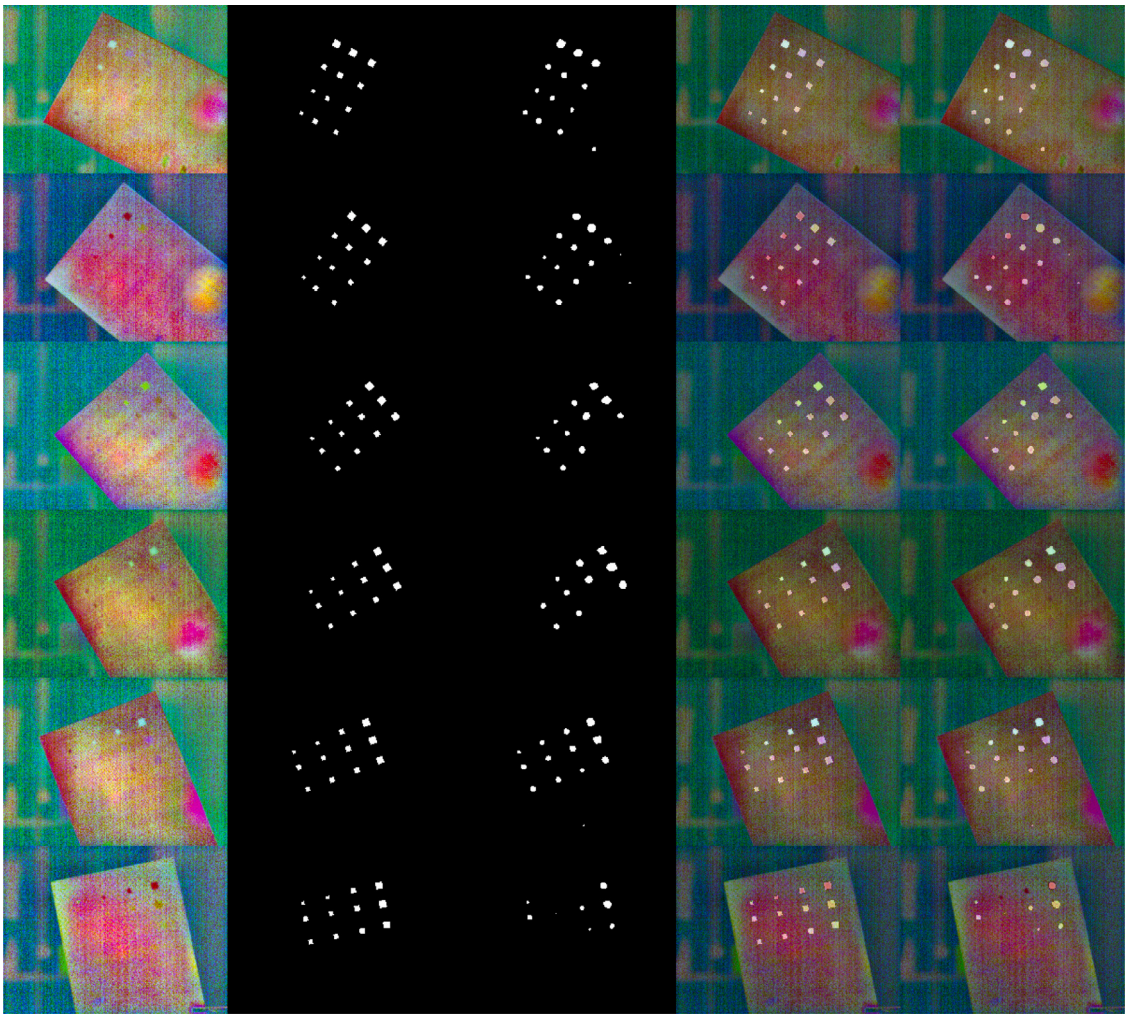


Fig. 13. Visualization of the predicted results for UNet evaluated with 30 channels. Only the first three channels are shown in the image, which consist of the first, third and fourth components, using exactly the same channels as the 3-channel images. (**1st col.**) Original images, (**2nd col.**) ground truth masks, (**3rd col.**) predictions, (**4th col.**) original images and ground truth masks, (**5th col.**) original images with predictions.

Table 10
Metrics for specimen 3.

Experiment	Precision	Recall
UNet (3-channel images)	.56	.50
DeepLabV3+ (3-channel images)	.84	.32

reflectivity (See Fig. 18). This ply is 20 mm thick and is called "Rohacell core" and is a registered trademark of structural foams that have high mechanical performance (https://www.rohacell.com/en). These foams have been used in the aeronautical sector for a long time to lighten composite materials and are currently used for the same purpose in other industries, such as the automotive and wind sectors.

There are no defects deeper than that of the Rohacell core because with thermography it is not possible to detect defects due to the fact that it is a great thermal insulator. The depth of the defects are 0.38 mm, 0.75 mm, 1.125 mm from left to right. The bottom three defects are steel chips defects and the rest are PTFE defects.

As can be seen in Fig. 19, most of the defects are not detected successfully. It seems that the new layer aggressively affects the reflectivity and therefore the behavior of the model for defect detection. In Table 10 the metrics for this evaluation are obtained. These metrics obviously present very poor results.

As a result of these evaluations, it can be observed that as long as the tested specimen has a similar structure to that of the training specimen, high quality detections can be achieved even if the depth of the specimen is not exactly the same as in the training specimen. However, if the specimen structure is severely altered, by adding an inner layer with different reflectivity, or very drastic depth changes, the semantic segmentation models are not able to find all the defects in the specimen. In particular, the Rohacell core changes the boundary conditions of the heat transfer problem, which affects the results obtained in the inspections.

5. Conclusion

This paper studies different solutions from the computer vision branch for pixel-based defect detection in CFRP specimens. It evaluates older and more common models such as Random Forest and Support Vector Machines against state-of-the-art approaches such as convolutional neural networks for semantic segmentation.

Semantic segmentation networks are capable of detecting subsurface defects far outperforming older methods such as Random Forest or Support Vector Machines. In addition, semantic segmentation has a great advantage over object detection thanks to its ability to detect defects of any shape, not only square defects.

More complex and modern networks like DeepLabV3+ tend to perform better, but increasing the amount of data per sample given to the model is almost as effective, as seen with UNet. Using 30-channel images instead of 3-channel images significantly improves

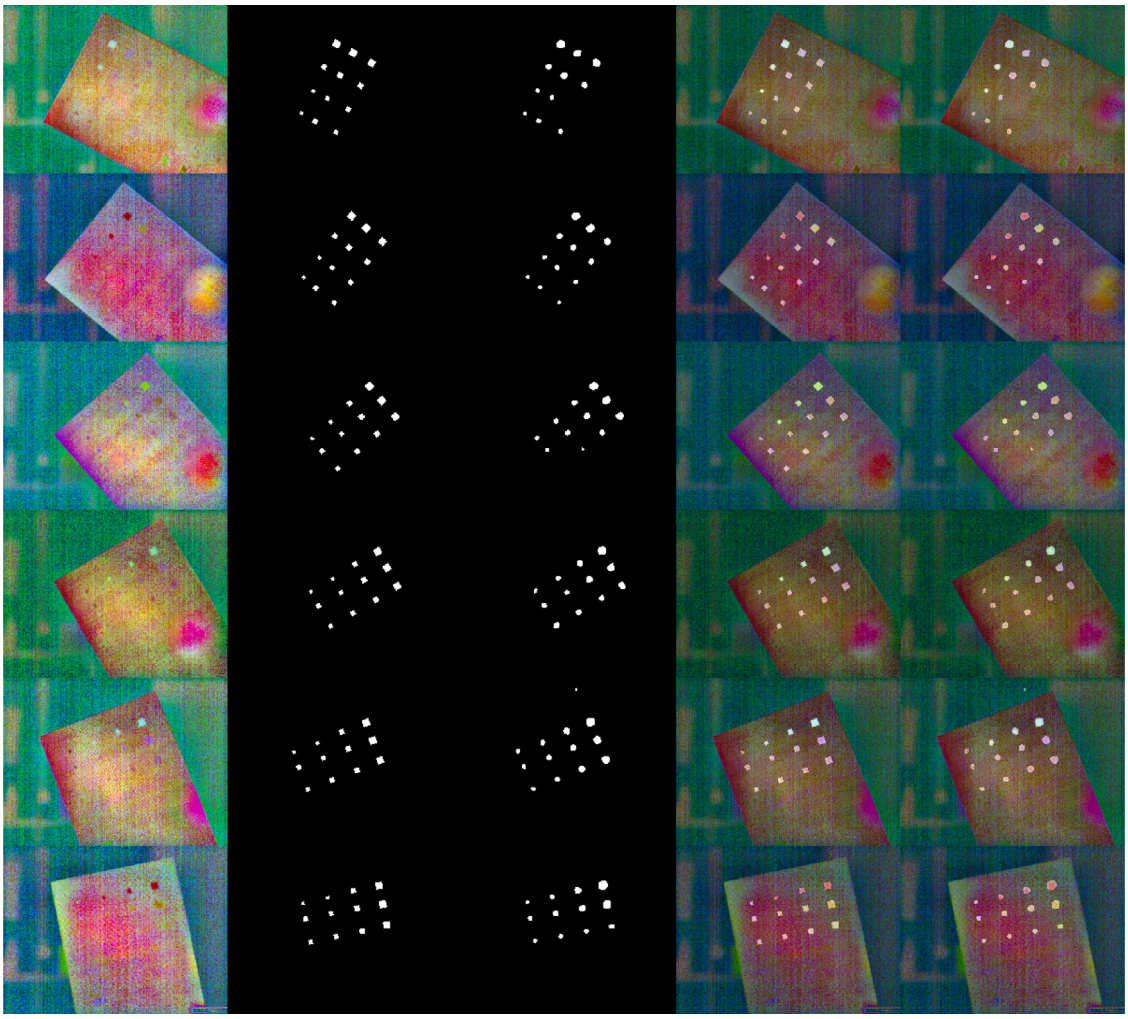


Fig. 14. Visualization of the predicted results for DeepLabV3+. (**1st col.**) Original images, (**2nd col.**) ground truth masks, (**3rd col.**) predictions, (**4th col.**) original images with ground truth masks, (**5th col.**) original images with predictions.

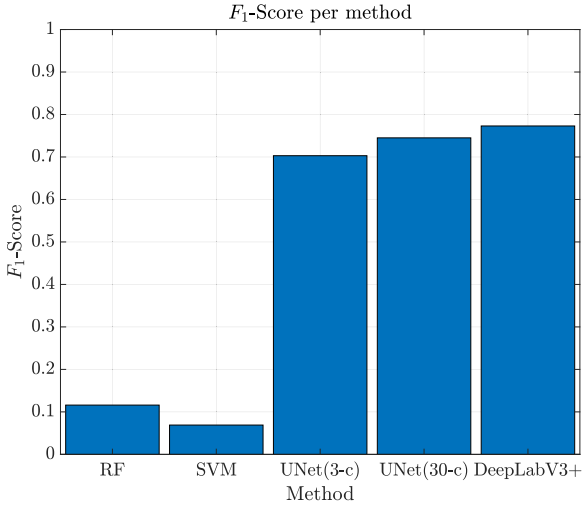


Fig. 15. Bar graph for all the methods.

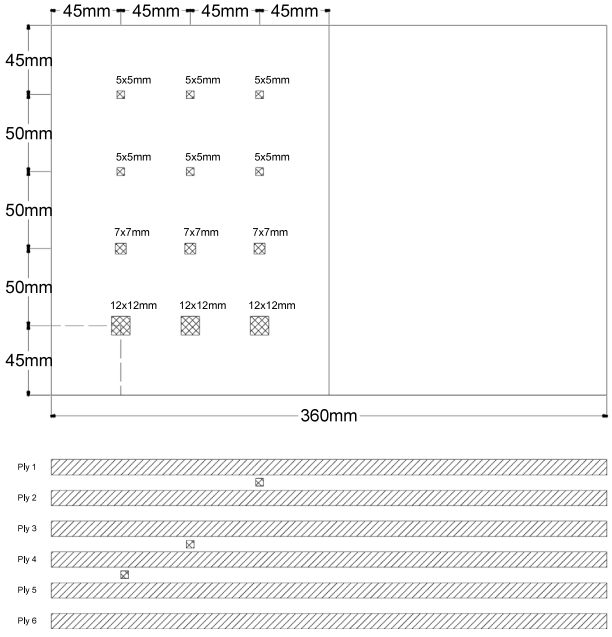


Fig. 16. Diagram of specimen 2.

predictability. To ensure reproducibility and further investigations, the dataset generated for this article is publicly available in the following DOI: https://doi.org/10.5281/zenodo.5426792.

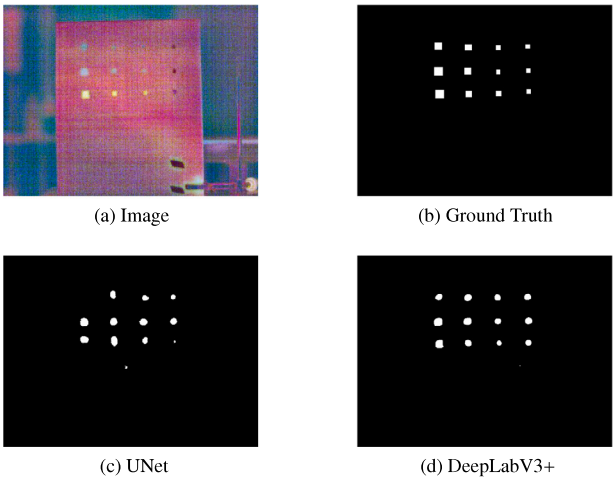


Fig. 17. Specimen 2.

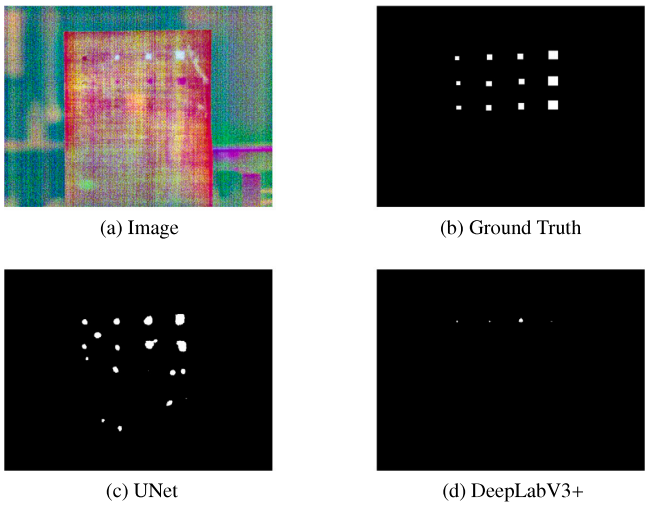


Fig. 19. Specimen 3.

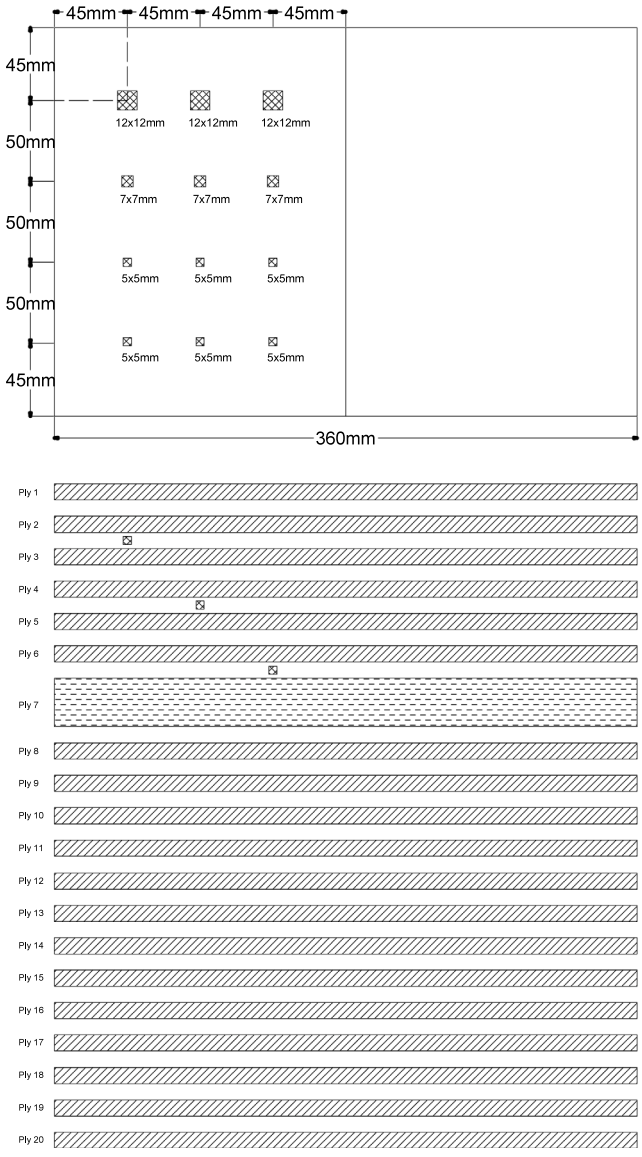


Fig. 18. Diagram of specimen 3.

To increase the validity of this study, evaluations of the DeepLabV3+ and UNet trained models for the 3-channel dataset are performed on new specimens with different internal structures. This evaluation proves that as long as the specimen has a similar internal structure, defect detection with strong results is possible.

Without performing testing on composite specimens with naturally occurring flaws, it is not possible to validate this technique. Almost all defects are easily detected and without false positives in this test for artificially induced defects with DeepLabV3+. The structure of the specimens needs to be similar to the training samples. A larger and more varied dataset would produce improved results.

It is apparent that these technologies could provide a solid support to help experts who have to check each specimen manually. Considering how fast the field of computer vision is evolving, it would be no surprise if deep learning algorithms become the norm for subsurface defect detection.

This study shows that there is still room for improvement in this field. For example, a GPU with more than eleven gigabytes of VRAM could slightly improve the results offered in this evaluation work by increasing the batch size. In addition, if the architecture of DeepLabV3+ were modified to accept 30-channel images, it could improve its results, although this would further limit the VRAM required.

CRediT authorship contribution statement

Oscar D. Pedrayes: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Darío G. Lema:** Investigation, Resources, Supervision. **Rubén Usamentiaga:** Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Pablo Venegas:** Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision. **Daniel F. García:** Investigation, Resources, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is available in https://doi.org/10.5281/zenodo.5426792.

Acknowledgments

This work has been partially funded by the project RTI2018-094849-B-I00 of the Spanish National Plan for Research, Development and Innovation, Spain.

References

[1] B. Kamsu-Foguem, Knowledge-based support in Non-Destructive Testing for health monitoring of aircraft structures, Adv. Eng. Inform. 26 (4) (2012) 859–869.
[2] S. Gholizadeh, A review of non-destructive testing methods of composite materials, Proc. Struct. Integr. 1 (2016) 50–57.
[3] B. Yousefi, D. Kalhor, R. Usamentiaga Fernández, L. Lei, C.I. Castanedo, X.P. Maldague, et al., Application of deep learning in infrared non-destructive testing, in: QIRT 2018 Proceedings, 2018.
[4] Q. Luo, B. Gao, W.L. Woo, Y. Yang, Temporal and spatial deep learning network for infrared thermal defect detection, NDT & E Int. 108 (2019) 102164.
[5] Q. Fang, X. Maldague, A method of defect depth estimation for simulated infrared thermography data with deep learning, Appl. Sci. 10 (19) (2020) 6819.
[6] R. Marani, D. Palumbo, U. Galietti, T. D'Orazio, Deep learning for defect characterization in composite laminates inspected by step-heating thermography, Opt. Lasers Eng. 145 (2021) 106679.
[7] Y. He, B. Deng, H. Wang, L. Cheng, K. Zhou, S. Cai, F. Ciampa, Infrared machine vision and infrared thermography with deep learning: a review, Infrared Phys. Technol. (2021) 103754.
[8] P. Theodorakeas, E. Cheilakou, E. Ftikou, M. Koui, Passive and active infrared thermography: An overview of applications for the inspection of mosaic structures, in: J. Phys. Conf. Ser., vol. 655, IOP Publishing, 2015, 012061.
[9] R. Usamentiaga, P. Venegas, J. Guerediaga, L. Vega, I. López, A quantitative comparison of stimulation and post-processing thermographic inspection methods applied to aeronautical carbon fibre reinforced polymer, Quant. InfraRed Thermogr. J. 10 (1) (2013) 55–73.
[10] C. Ibarra-Castanedo, X. Maldague, Pulsed phase thermography reviewed, Quant. Infrared Thermogr. J. 1 (1) (2004) 47–70.
[11] Y. Liu, G. Tian, B. Gao, X. Lu, H. Li, X. Chen, Y. Zhang, L. Xiong, Depth quantification of rolling contact fatigue crack using skewness of eddy current pulsed thermography in stationary and scanning modes, NDT & E Int. 128 (2022) 102630.
[12] I. Sgibnev, A. Sorokin, B. Vishnyakov, Y. Vizilter, Deep semantic segmentation for the off-road autonomous driving, Int. Archiv. Photogram. Remote Sens. Spatial Inform. Sci. 43 (2020) 617–622.
[13] O.D. Pedrayes, D.G. Lema, D.F. García, R. Usamentiaga, Á. Alonso, Evaluation of semantic segmentation methods for land use with spectral imaging using sentinel-2 and PNOA imagery, Remote Sens. 13 (12) (2021) 2292.
[14] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: ECCV, 2018, pp. 801–818.
[16] C. Zhang, Y. Ma, Ensemble Machine Learning: Methods and Applications, Springer, 2012.
[17] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, 2000, http://dx.doi.org/10.1017/CBO9780511801389.
[18] Y. Cao, Y. Dong, Y. Cao, J. Yang, M.Y. Yang, Two-stream convolutional neural network for non-destructive subsurface defect detection via similarity comparison of lock-in thermography signals, NDT & E Int. 112 (2020) 102246.
[19] C. Schmidt, T. Hocke, B. Denkena, Artificial intelligence for non-destructive testing of CFRP prepreg materials, Product. Eng. 13 (5) (2019) 617–626.
[20] H.-T. Bang, S. Park, H. Jeon, Defect identification in composite materials via thermography and deep learning techniques, Compos. Struct. 246 (2020) 112405.
[21] Y. Dong, C. Xia, J. Yang, Y. Cao, Y. Cao, X. Li, Spatio-temporal 3D residual networks for simultaneous detection and depth estimation of CFRP subsurface defects in lock-in thermography, IEEE Trans. Ind. Inf. (2021).
[22] Z. Nie, J. Xu, S. Zhang, Analysis on DeepLabV3+ performance for automatic steel defects detection, 2020, ArXiv preprint arXiv:2004.04822.
[23] R. Usamentiaga, C. Ibarra-Castanedo, M. Klein, X. Maldague, J. Peeters, A. Sanchez-Beato, Nondestructive evaluation of carbon fiber bicycle frames using infrared thermography, Sensors 17 (11) (2017) 2679.
[24] K. Zheng, Y.-S. Chang, K.-H. Wang, Y. Yao, Improved non-destructive testing of carbon fiber reinforced polymer (CFRP) composites using pulsed thermograph, Polym. Test. 46 (2015) 26–32.
[25] D. Schumacher, N. Meyendorf, I. Hakim, U. Ewert, Defect recognition in CFRP components using various NDT methods within a smart manufacturing process, in: AIP Conference Proceedings, vol. 1949, AIP Publishing LLC, 2018, 020024.
[26] R. Marani, D. Palumbo, V. Renò, U. Galietti, E. Stella, T. D'Orazio, Modeling and classification of defects in CFRP laminates by thermal non-destructive testing, Composites B 135 (2018) 129–141.
[27] N. Rajic, Principal Component Thermography, Tech. Rep., Defence Science and Technology Organisation Victoria, Australia ..., 2002.
[28] B. Milovanović, M. Gaši, S. Gumbarević, Principal component thermography for defect detection in concrete, Sensors 20 (14) (2020) 3891.
[29] H.J. Nussbaumer, The fast Fourier transform, in: Fast Fourier Transform and Convolution Algorithms, Springer, 1981, pp. 80–111.
[30] J. Bodzenta, A. Kaźmierczak, T. Kruczek, Analysis of thermograms based on FFT algorithm, J. Physique IV 129 (2005) 201–205.
[31] F.J. Madruga, C. Ibarra-Castanedo, O.M. Conde, X.P. Maldague, J.M. López-Higuera, Enhanced contrast detection of subsurface defects by pulsed infrared thermography based on the fourth order statistic moment, kurtosis, in: Thermosense XXXI, vol. 7299, International Society for Optics and Photonics, 2009, p. 72990U.
[32] F. Madruga, C. Ibarra-Castanedo, O. Conde, J. Lopez-Higuera, X. Maldague, Automatic data processing based on the skewness statistic parameter for subsurface defect detection by active infrared thermography, in: Proc. QIRT, vol. 9, Citeseer, 2008, p. 6.
[33] S. Shepard, J. Lhota, B. Rubadeux, D. Wang, T. Ahmed, Reconstruction and enhancement of active thermographic image sequences, Opt. Eng. 42 (2003) 1337–1342, http://dx.doi.org/10.1117/1.1566969.
[34] D. Balageas, B. Chapuis, G. Deban, F. Passilly, Improvement of the detection of defects by pulse thermography thanks to the TSR approach in the case of a smart composite repair patch, Quant. InfraRed Thermogr. J. 7 (2) (2010) 167–187.
[35] E.O. Brigham, The Fast Fourier Transform and Its Applications, Prentice-Hall, Inc., 1988.
[36] D.L. Balageas, J.-M. Roche, F.-H. Leroy, W.-M. Liu, A.M. Gorbach, The thermographic signal reconstruction method: A powerful tool for the enhancement of transient thermographic images, Biocybern. Biomed. Eng. 35 (1) (2015) 1–9.
[37] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, 2014, ArXiv preprint arXiv:1412.7062.
[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[39] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, 2017, ArXiv preprint arXiv:1706.05587.
[40] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A.L. Yuille, L. Fei-Fei, Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 82–92.
[41] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
[42] E. Fernandez-Moral, R. Martins, D. Wolf, P. Rives, A new metric for evaluating semantic segmentation: leveraging global and contour accuracy, in: 2018 Ieee Intelligent Vehicles Symposium (Iv), IEEE, 2018, pp. 1051–1056.
[43] S.H. Khaleefah, S.A. Mostafa, A. Mustapha, M.F. Nasrudin, Review of local binary pattern operators in image feature extraction, Indonesian J. Electric. Eng. Comput. Sci. 19 (1) (2020) 23–31.
[44] E. Miyamoto, T. Merryman, Fast Calculation of Haralick Texture Features, Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, USA. Japanese Restaurant Office, 2005.
[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.