

A Lightweight Spatial and Temporal Multi-feature Fusion Network for Defect Detection

Bozhen Hu, Bin Gao, *Senior Member, IEEE*, Wai Lok Woo, *Senior Member, IEEE*, Lingfeng Ruan, Jikun Jin, Yang Yang, Yongjie Yu

Abstract—This paper proposes a hybrid multi-dimensional features fusion structure of spatial and temporal segmentation model for automated thermography defects detection. In addition, the newly designed attention block encourages local interaction among the neighboring pixels to recalibrate the feature maps adaptively. A Sequence-PCA layer is embedded in the network to provide enhanced semantic information. The final model results in a lightweight structure with smaller number of parameters and yet yields uncompromising performance after model compression. The proposed model allows better capture of the semantic information to improve the detection rate in an end-to-end procedure. Compared with current state-of-the-art deep semantic segmentation algorithms, the proposed model presents more accurate and robust results. In addition, the proposed attention module has led to improved performance on two classification tasks compared with other prevalent attention blocks. In order to verify the effectiveness and robustness of the proposed model, experimental studies have been carried out for defects detection on four different datasets.

Index Terms—image segmentation, Sequence-PCA, attention, model compression, defect detection

I. INTRODUCTION

Image segmentation can be defined as a specific image processing technique that is used to divide an image into two or more meaningful regions [1]. Convolution operations are effective in extracting features. Fully Convolutional Networks (FCN) [2] has been introduced to replace the fully connected layers and used the deconvolution (also commonly known as transposed convolution) function to introduce dense predictions for per-pixel tasks. Mask R-CNN [3] extends Faster R-CNN [4] with a parallel branch to perform pixel level object specific binary classification to provide more accurate segments. In order to capture contextual information and maintain the sharpness of the segmented outputs, Chen *et al.* [5] [6] proposed the DeepLab family of algorithms. The usage of diverse methodologies such as atrous convolutions [7], spatial pooling pyramids [8], and conditional random fields (CRF) [9] were demonstrated to perform image segmentation at a level of accuracy which is beyond previous methods. Ronneberger *et al.* [10] proposed UNet based on the encoder-decoder architecture with skip connections, which has shown an excellent segmentation results for a variety of problems. The state-of-the-art models for image segmentation are variants of the encoder-decoder architecture, including Dense-UNet [11], SegNet [12], and UNet++ [13]. By extending the fully connected LSTM (FC-LSTM) [14], Shi *et al.* [15] proposed the convolutional LSTM (ConvLSTM) which have been used to perform instance level segmentation. Bai *et al.* [16] conducted a systematic evaluation of generic convolutional and recurrent architectures for

sequence modeling, evaluated across a broad range of standard tasks and datasets, indicating that a simple convolutional architecture outperforms canonical recurrent networks. As for the image segmentation of multidimensional data, Deep voxel-wise residual networks [17] was proposed by combining the low-level image appearance features for further improving the segmentation performance. V-Net [18] and 3D U-Net [19] were proposed for 3D image segmentation based on volumetric, fully convolutional neural networks. In order to address the issues of high computational cost and GPU memory consumption, a hybrid densely connected UNet (H-DenseUNet) [11] was proposed, which uses a 2D DenseUNet for efficiently extracting intra-slice features and a 3D counterpart for hierarchically aggregating volumetric contexts.

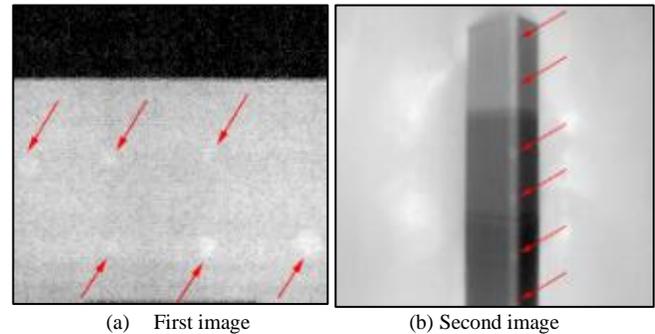


Fig. 1. Infrared thermal images and defects locations

Nondestructive testing (NDT) falls under the research category of computer vision for defects detection without causing damage to the specimens. For quality assurance, the necessity to monitor the health and quality of composite materials becomes ever more important [20]. NDT and evaluation methodologies include safety penetrant testing, eddy current testing, ultrasonic testing, and optical pulsed thermography (OPT) [21]. OPT system has the advantages of large single detection area, fast speed, non-contact, safety, and simple operation compared with traditional detection technology. In this paper, all defects are pre-embedded/produced in the carbon fiber reinforced polymer (CFRP) materials during the manufacturing process for detection validation. These have been provided by the Chengdu Aircraft Design Institute of China Aviation Industry. This type of composite materials are used for aircraft manufacturing. In the experiments, both the OPT and portable optical pulsed thermography (POPT) system are used to get the infrared thermal volumetric data of these materials. When a specimen is heated by external light, the heat energy is transmitted in the

specimen. Due to the existence of defects inside the specimen, the power at the defect area and non-defect area is different. Thus, the temperature of the surface of this specimen is different. Based on these, it is possible to determine the locations of the defects in the infrared thermal images, which are shown in Fig. 1 with the red arrows pointing at the locations of the defects. Our task is to determine all defects' locations in the infrared thermal images accurately and quickly. The methodology on annotating the images and converting segmentation results to defects detection is presented in Section II.E.2.

Various approaches have been proposed to process the infrared thermal volumetric data produced by the OPT system. Principal component analysis (PCA) [22] [23] is a commonly used algorithm, which combines both spatial and temporal information for extracting defect information in time series. Pulsed phase thermography (PPT) [24] [25] is a nondestructive evaluation processing technique based on the discrete Fourier Transform, transforming time domain features into frequency domain, extracting defect information from frequency domain to eliminate noise. Due to the excellent effects on image and video processing, deep learning algorithms have become been applied to Infrared Non Destructive Testing (IRNDT). Since the data produced by the OPT system is multi-dimensional, Luo *et al.* [26] proposed a visual geometry Group-UNet and LSTM cross learning structure which can significantly improve the contrast between the defective and non-defective regions. However, the issue of how to make full use of all dimensional features, especially the time-dimensional information, is still an urgent problem to be solved for infrared thermal volumetric data which suffers from low resolution, high noise as well as uneven heating. Therefore, it is difficult to train an existing network to detect inner defects with a large sufficient capacity of detecting weaker and small defects on a complex and irregular surface end to end.

In order to deal with these limitations, we propose a multi-feature fusion deep network to significantly enhance the detection rate and simultaneously extract both the spatial and temporal information. Different samples with various sizes of defects at different levels are used to validate the accuracy and robustness of the proposed algorithm. The comparable analysis has been undertaken with the state-of-the-art deep semantic segmentation algorithms.

To summarize, the contributions of the paper are as follows:

- (i) Proposal of a hybrid multi-dimensional spatial and temporal segmentation model which fuses features in multiple dimensions. This allows the users to achieve the goal of defects detection with an end-to-end structure.
- (ii) Design of a new attention block to provide spatiotemporal attention to focus on semantically meaningful regions of the volumetric data and recalibrate the feature maps adaptively, based on the weighted channels. In addition, a new layer, termed as Sequence-PCA Layer is proposed in the learning process to provide extra semantic information.
- (iii) Development of Automated Machine Learning (AutoML) for model compression in the volumetric data segmentation task to perform automatic drop connections of

unimportant nest nodes that cause the back propagation invalid as well as simplify the overall network structure.

The rest of the paper is organized as follows: The details of the proposed model and the quantitative detectability assessment indicators are described in Section II. Experiments and results analysis are introduced in the Section III. Finally, Section IV draws the conclusion of the work and highlights the future work.

II. METHODOLOGY

This section interprets the framework of the proposed model by investigating different deep learning networks. Fig. 2 shows the framework of the proposed hybrid multi-dimensional spatial and temporal segmentation model. The framework can be mainly summarized as the three parts: data preprocessing (Part A), whole model and the improved parts of the model (Part B), the prediction and evaluation (Part E).

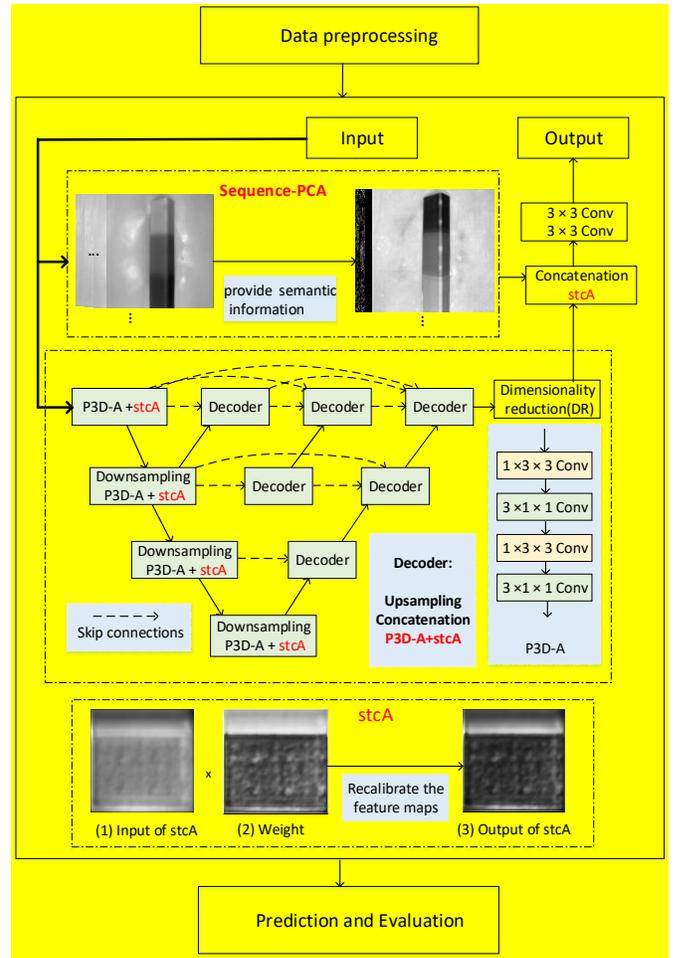


Fig. 2. Proposed model descriptions

A. Data Preprocessing

The infrared thermal training sets are obtained from different types of specimens where each specimen contains the defects of different diameters and depths. The details of data preprocessing are shown in Fig. 3.

For the obtained volumetric data $\mathbf{D}_i \in \mathbb{R}^{m \times n \times f}$, where (m, n) is the size of the frame, f is the number of total frames, and i represents the i^{th} specimen. In order to remove the background

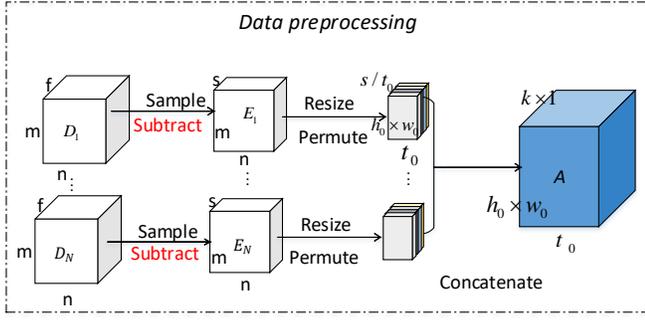


Fig. 3. Data preprocessing

noise and reduce the effects of thermal diffusion, the mean of last b_l frames is subtracted for every frame of the volumetric data:

$$\widehat{Im}_j = Im_j - (\sum_{j=f-b_l+1}^f Im_j) / b_l \quad (1)$$

where Im_j represents the j^{th} slice of data D_i . s frames with the highest contrast are extracted to obtain the sequence data $E_i \in \mathbb{R}^{m \times n \times s}$ with $s < f$. The temperature gap between the defect area and the non-defect area is the largest around the frame with the highest temperature, which is more conducive for the segmentation of the defects as shown in Fig. 4.

Every frame of the data is resized to (h_0, w_0) to maintain a consistent size during the encoding and decoding process, where h_0 and w_0 is set as the power of two. We extract t_0

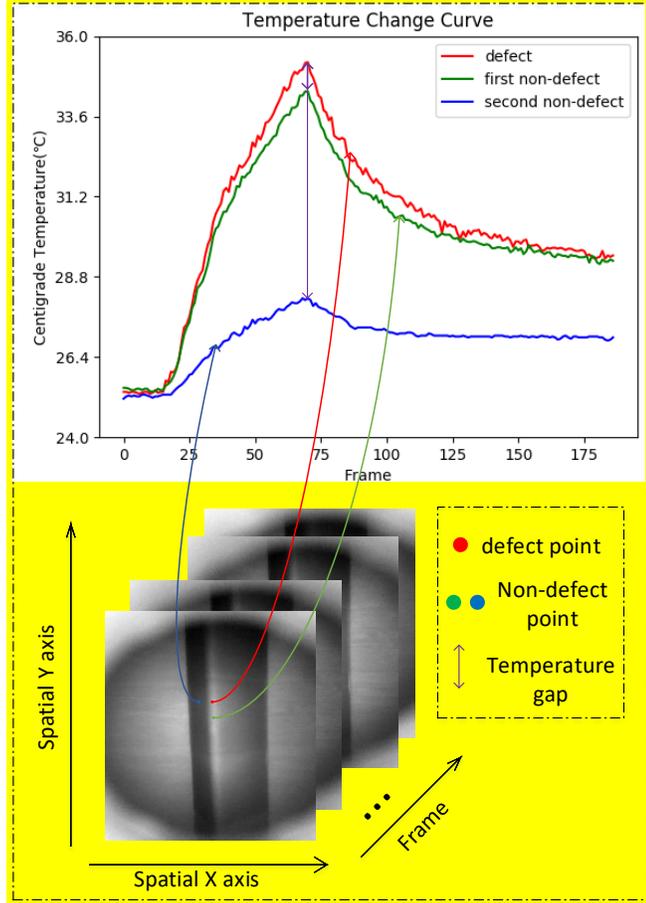


Fig. 4. Temperature change curve at different points corresponding to defective and background areas

frames from s frames at a regular interval for one specimen to construct batches of time sequences, concatenating all the preprocessed data to get the final training data $A \in \mathbb{R}^{k \times t_0 \times h_0 \times w_0 \times c_0}$ where k is the total number of the training data, and c_0 is the number of the image's channels. As for gray-scale images, c_0 is set to one. The specific values of these parameters can be found in Section III.B.

B. Proposed Hybrid 3-Dimensional Spatial and Temporal Segmentation Model

Unlike many models that use convolutional and temporal networks separately, 3D-CNN is applied due to its excellent performance in extracting the multi-dimensional features. In order to save computational cost and meet the required memory, we use $1 \times 3 \times 3$ convolutional filters (equivalent to 2D CNN) plus $3 \times 1 \times 1$ convolutions to construct temporal connections to replace $3 \times 3 \times 3$ convolutions [27]. This significantly reduces the training time and the number of trainable parameters. In [28], the Pseudo-3D (P3D) [27] is shown to be better than 2D and 3D convolutions through a variety of tasks and datasets. P3D-A which considers stacked architecture by making temporal 1D filters follow spatial 2D filters in a cascaded manner is used in the final model as depicted in Fig. 2. UNet++L4 [13] is employed as our baseline mode. A Sequence-PCA layer is designed to provide extra semantic information. A new attention block is proposed to boost semantically meaningful features and suppress the weaker ones, thereby aiding a more fine-grained segmentation, inspired by concurrent spatial and channel squeeze and channel excitation(scSE) block[29][30]. These details are given in the following sections.

1) Sequence-PCA layer

Inspired by the Rectified Local Phase unit [31] and the ReLPV [32] block, we can use a certain algorithm to extract useful information from the volumetric data as a complementary of the outputs from the original baseline deep network. When annotating the slice of the infrared thermal volumetric data, the process suffers from edge blurring, uneven heating and high noise to create labels. To overcome this issue, PCA is used to process the data to aid the annotators to determine the locations of defects. The processed results and annotation methods are tabulated in Table III and Section II.E.2

PCA can extract both spatial and temporal information, and we choose the first n ($n=4$) components that are sufficient to illustrate obvious defects locations among the high noise thermal images. The choice of n is based on the desired amount of the variance proportion retained in the first n eigenvalues. In many cases, more than 95% of variance is contained in the first three to five components [23].

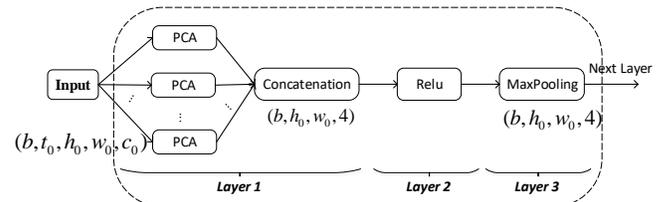


Fig. 5. Architecture of the proposed Sequence-PCA layer

Inspired by the process of artificial annotation on the slice of the thermal data and consideration of some advantages of PCA algorithm including fast speed and simple implementation, we create a Sequence-PCA layer to deal with the batches of data during training to provide more semantic information. Fig. 5 illustrates the architecture of the proposed Sequence-PCA layer and it is a three-layer alternative representation of one standard 3D convolutional layer.

The *Layer 1* takes one batch of the training data as input with size (b, t_0, h_0, w_0, c_0) , converting it into b groups of single tensor with size (t_0, h_0, w_0, c_0) where b is the batch size, (h_0, w_0, c_0) is the initial size of the resized thermal image, and t_0 represents the length of time sequence. This layer conducts the operation for every sequence tensor and obtains the first four components with the irrelevant others abandoned. The obtained feature maps are concatenated for batches. The comparison results are shown in Table III and Table IV. The basic idea behind this approach is to enable the network to learn from its own sequence quickly. The *Layer 2* is the ReLU activation layer [33] which takes the tensor with size $(b, h_0, w_0, 4)$ from *Layer 1* as input and output a tensor with shape unchanged. ReLU layer can add non-linearity to a certain degree. The *Layer 3* is an overlapping maxpooling layer [34], with pooling regions of size 2×2 , stride 1 and padding the same as the latest input tensor size. This can retain more details for the input tensor as well as aggregate the features from pre-defined receptive field, which is observed slightly more difficult to overfit.

As for the forward-backward propagation in the Sequence-PCA layer, let F as the mapping of the internal network parallelized with the Sequence-PCA layer and I as the input, the original mapping with Sequence-PCA layer is then recast into $H(I) := \text{concatenation of } F(I) \text{ and Sequence-PCA}(I)$. Therefore, the end-to-end training of this hybrid model is straightforward. Backward propagation through the *Layer 1*, *2* and *3* of the Sequence-PCA layer is similar to propagating gradients through the layers without any learnable parameters (e.g. Multiply, Add, Pooling etc.).

2) Spatiotemporal and Channel Attention Block (stcA)

The original sSE [29] block squeezes U , a slice of the feature map, along the channels as

$$q = W_{sq} * U \quad (2)$$

where $*$ is the convolution operation and weight $W_{sq} \in \mathbb{R}^{1 \times 1 \times c \times 1}$. q is the projection tensor with size $(1, h, w, 1)$, then which passes through a sigmoid layer $\sigma(\cdot)$ to rescale activations to $[0, 1]$. $\sigma(q)$ is used to recalibrate or excite U spatially, each value in $\sigma(q)$ corresponds to the relative importance of U . The architectural flow is shown in Fig. 6(a).

A 1×1 convolution with one filter is mainly used in the original sSE block to recalibrate the feature maps. Given a spatial location (j, k) with $j \in \{1, 2, \dots, h\}$ and $k \in \{1, 2, \dots, w\}$, each q_{jk} of the projection represents the linearly combined representation for all channels of the feature map. If there is one channel for an input tensor, it is only one trainable parameter in the original sSE block. Thus, it is equivalent to performing a sigmoid function to each scaled value for the input tensor. Moreover, the category of a pixel is often related to its neighbors, usually having the same properties in multi-dimension. Generally speaking, the shallower the layer of the network, the smaller the receptive field of the feature map.

Hence, there is an importance in investigating a more accurate weight selection method for a point, depended on its channels as well as neighbors. To achieve this goal, we define a block with size (k_t, k_s, k_c) which is used to determine the neighbors of one pixel. After obtaining the channel-recalibrated data through the Efficient Channel Attention (ECA) [35] module, our proposed spatiotemporal and channel attention (stcA) block can be efficiently implemented by a fast 3D-convolution with kernel size (k_t, k_s, k_c) , which encourages local interaction among the neighboring pixels. This 3D-convolution expands the receptive field to a larger extent for the input tensor that can get more accurate recalibration results. In addition, the kernel size k_t for temporal dimension is different with k_s for spatial dimension since the infrared thermal volumetric data suffers incompatibility with the anisotropic spatiotemporal dimensions. The choice of k_t , k_s , and k_c has been tabulated in Table II and Section II.D.

In the proposed method, we consider an alternative slicing of the input tensor $U_{3d} = [u^{1,1,1,1}, u^{1,1,1,2}, \dots, u^{1,i,j,k}, \dots, u^{1,t,h,w}]$, where $u^{1,i,j,k} \in \mathbb{R}^{1 \times 1 \times 1 \times c}$, is the tensor of the spatiotemporal location (i, j, k) in the given tensor with c channels, $i \in \{1, 2, \dots, t\}$, $j \in \{1, 2, \dots, h\}$, and $k \in \{1, 2, \dots, w\}$. The projection tensor q with size $(1, t, h, w, 1)$ can be obtained by:

$$q = W_{3d} * U_{3d} \quad (3)$$

where $*$ is the convolution operation and weight $W_{3d} \in \mathbb{R}^{k_t \times k_s \times k_c \times c \times 1}$. This 3D-convolution allows different blocks to share the same learning parameters to reduce model complexity and improve efficiency. The stcA block can be seen in Fig. 6(b). Each value of the projection tensor q represents the linearly combined representation for all its own and neighbors' weighted channels because of the cascade architecture of the stcA block, which is proved more accurate. The effects of the stcA block are analyzed in Section III.D.

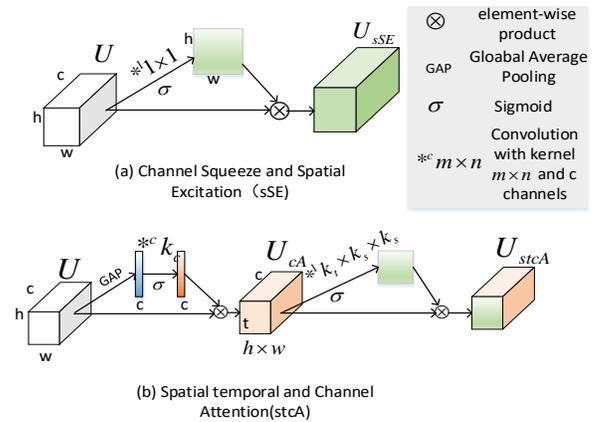


Fig. 6. Architecture design of sSE and stcA blocks

Maxpooling layer with kernel $2 \times 2 \times 2$ and stride 2 are often used for down sampling in such encoder-decoder architectures. Generally, one point in the feature map that has performed three down sampling operations corresponds to about $8 \times 8 \times 8$ pixels in the initial input tensor. Therefore, if we add stcA blocks to the internal feature maps, these recalibrations can be seen to put weights to blocks in the initial input tensor. Defects are usually in such blocks that we can provide more importance and suppress the irrelevant areas relatively.

C. Network Pruning and Compression: Drop Connections

In a typical deep learning model, the shallow layers which extract low level features have less number of parameters whereas the deep layers are opposite. As the layers in deep neural networks are not independent, the design space has exponential complexity with network going deeper while the manual model compression is time-consuming and not optimal. Inspired by the insights of DropConnect [36] and AMC [37] network, we develop AutoML for model compression in the volumetric data segmentation task with differentiable neural architecture search [38][39] to automatically sample the design space and improve the model compression quality.

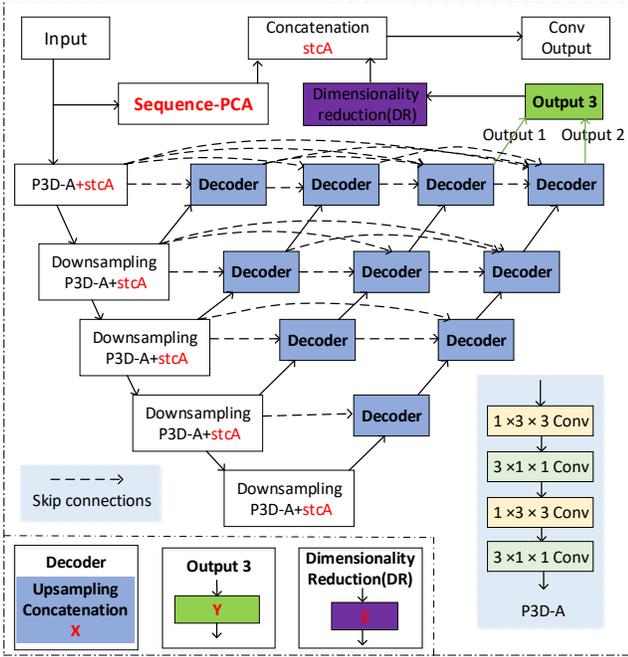


Fig. 7. Hybrid segmentation network architecture. The Decoder cell has a candidate layer \mathbf{X} , which is chosen from $P3D-A+stcA$, $P3D-A$, and $Identity$; Correspondingly, \mathbf{Y} is chosen from $Output\ 1$ or $Output\ 2$, and \mathbf{Z} is chosen from 0 to $t_0 - 1$, in the first layer, t_0 is 16

We define a cell to be a fully convolutional module, typically repeated multiple times to form the entire neural network. The proposed segmentation network follows the encoder-decoder structure with skip pathways while the decoder cells are learned in a differential way [39]. The whole network structure is illustrated in Fig. 7. The blue decoders, green outputs, and purple dimension reduction cell are all in the search space.

The operations of the white background cells are as shown in the Fig. 7. A Decoder cell has candidates about $P3D-A+stcA$, $P3D-A$, and $Identity$ block. These different choices can make the network learn its own best architecture to have a better performance without introducing much additional parameters as analyzed in Section II.D. The cell $Output\ 3$ can choose $Output\ 1$ with only three down sampling operation or $Output\ 2$ which has four down-sampling operations. The *Dimensionality Reduction* cell is designed to make the dimension consistent for the feature maps, which are from Sequence-PCA layer and network $Output\ 3$ for the subsequent concatenation in the next layer.

As for the search space, as can be seen in Fig. 7, several

Decoders, an *Output cell*, and a *Dimension Reduction cell* are designed, which include different choices, the set of possible architectures is denoted as:

$$\varepsilon = \{Decoder_1 (Upsampling + P3D - A + stcA), Decoder_2 (Upsampling + P3D - A), Decoder_3 (Upsampling + Identity), Output\ 3_1 (output1), Output\ 3_2 (output2), DR_i (i | i \in [0,15], i \in N^*)\} \quad (4)$$

where N^* means an integer. As shown in (4), the choices of these candidate cells are listed. To make the search space continuous, we relax the categorical choice of a particular cell operation as a softmax [39] over these candidate cells, this continuous relaxation can make the scalars controlling the connections between different hidden states be a part of the differentiable computation graph [28]. Thus, the gradient descent method can be used to optimize them.

Through sharing parameters among child models, the time of training in this model are reduced. From the consideration about the performance and parameters of the child models, the final hybrid segmentation model can be seen in the Fig. 2. Comparisons about different choices of \mathbf{Y} and the values of \mathbf{X} , \mathbf{Z} are shown in Section III.E.

D. Model Complexity

1) Parameter analysis of the Sequence-PCA layer

The operations in Sequence-PCA layer are Multiply, Add, Concatenation, and Maxpooling, etc. Thus, the designed Sequence-PCA layer can process each batch of data in parallel with the neural network, which achieves the purpose of rendering the defect information to be more differentiated without any learnable parameters.

2) Parameter analysis for stcA blocks

For an encoder or decoder block, outputting a feature map with size (b, t, h, w, c) , the additional parameters introduced by a $stcA$ block are from two convolution layers. The number of trainable parameters is k_c in the 1D-convolution, if $c > 128$, we set k_c to five, otherwise, k_c equals to three, this makes channels of large size have long range of interactions. As for the 3D convolution in the $stcA$ block, the number of additional parameters is $k_t \times k_s \times k_s \times c$. In order to capture larger size of receptive field, we set k_s to three or five and k_t to one or two. So the total number of additional parameters introduced by $stcA$ blocks is

$$\sum_{i=1}^s k_{c_i} + (k_{t_i} \times k_{s_i} \times k_{s_i} \times c_i) \quad (5)$$

where k_{c_i} , k_{t_i} , and k_{s_i} are the kernel size of the convolutions in the i^{th} $stcA$ block, c_i is the number of the channels of the i^{th} encoder or decoder outputs, and s is the total number of the encoder or decoder blocks that contain $stcA$ blocks.

Taking the proposed hybrid network as an example, original UNet++L4 using $3 \times 3 \times 3$ convolutions has approximately 2.6×10^7 parameters. If we use $1 \times 3 \times 3$ convolutional filters plus $3 \times 1 \times 1$ convolutions to replace $3 \times 3 \times 3$ convolutions to construct the 3D UNet++L4 network, the number of the parameters reduces to 1.3×10^7 . After adding Sequence-PCA layer and $stcA$ blocks, which add 3.4×10^4 parameters, only accounting for around 0.2% of the total number of the parameters. We proposed AutoML for this network pruning and

compression, the final model goes to 3.14×10^6 parameters, which reduces by around 75.85%.

E. Loss Function and Quantitative Detectability Assessment

1) Loss Function

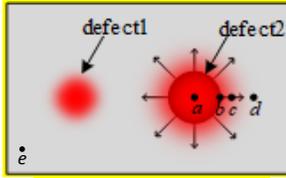
We have added a combination of binary cross-entropy (BCE) and dice coefficient (DC) as the loss function to this semantic task. Mathematically, the hybrid loss is defined as:

$$L(Y, \hat{Y}) = \beta_w \text{BCE}(Y, \hat{Y}) + \text{DC}(Y, \hat{Y}) \quad (6)$$

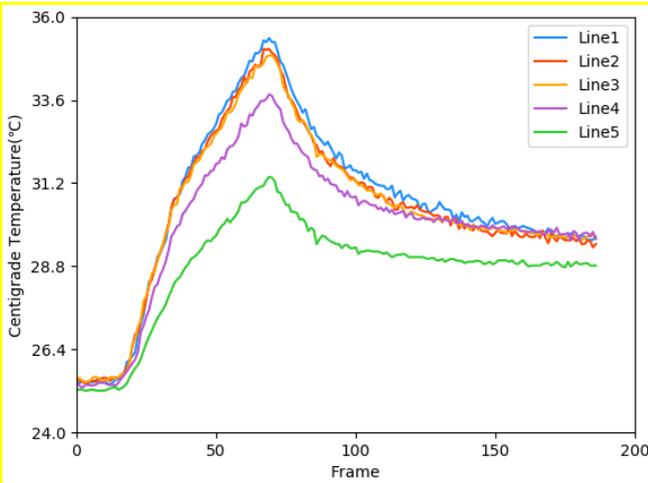
$$\text{BCE}(Y, \hat{Y}) = -\frac{1}{N} \sum_{c=1}^2 \sum_{n=1}^N y_{n,c} \log \hat{y}_{n,c} \quad (7)$$

$$\text{DC}(Y, \hat{Y}) = -\frac{1}{b} \sum_{i=1}^b \frac{2|Y_i \cap \hat{Y}_i|}{|Y_i| + |\hat{Y}_i|} \quad (8)$$

where $\hat{y}_{n,c} \in \hat{Y}_i$ and $y_{n,c} \in Y_i$ denote the i^{th} predicted and ground truth segmentation for class c and n^{th} pixel in the batch, b and N indicate the batch size and the number of pixels within one batch. β_w can adjust the weight of binary cross-entropy and dice coefficient.



(a) Thermal diffusion effect



(b) Corresponding temperature features

Fig. 8. Thermal diffusion and temperature curve

2) Quantitative Detectability Assessment

From the point of view of these infrared thermal images, defects and background are shown with different vision features, however, due to the effects of lateral and longitudinal thermal diffusion, the observed signals are not the expected feature signals as they are superimposed by interference [40]. In Fig. 8(a), we take five points a , b , c , d , and e , where point a represents the center of defect 2. We assume that the thermal feature of point a is the expected feature of defect, point b is the edge of defect 2, point c is the background near defect 2, point d is the background having a distance with defect 2, point e is the background far away from any defect. In Fig. 8(b), Lines 1–5 correspond to the thermal features of points a , b , c , d and e . It shows that Line 1 and Line 3 can be distinguished, while Line 2 and Line 3 are similar. In theory, Line 2 is similar to Line 1,

and Line 3 is similar to Line 4. However, since point b and point c are located at the boundary between the defect and the background regions, they are affected by thermal diffusion, and the features of the two points can be regarded as the superposition of background and defect features. Thus, it is resulting in the features of point b and point c are quite similar.

The locations of the defects can be determined from the infrared thermal images due to the effect of thermal diffusion whereas it is difficult to identify the real size of a defect. Three experienced human annotators have been employed to annotate the original thermal image sequences or the processed images by PCA algorithm independently. Considering the effect of thermal diffusion, they only marked the actual size of the defect as they determined (the circle where point b is located in Fig. 8(a)). Defects are marked in black and the background is marked in white. Finally, we take the intersection of the three annotation images as the ground truth, all frames of one infrared thermal volumetric data have one ground truth image. Fig. 9 shows an example of a thermal image, one component of PCA processed image and its ground truth image.

The method of intersections is adopted to convert segmentation outputs to defects detection. We first use the value of 0.5 as the threshold to binarize the outputs from the network, and then obtain the intersections of the output images and the ground truth images. If the intersection for a defect in a specific position is greater than 50%, the defect at that position in the image is deemed successfully detected.

The Intersection Over Union (IOU) has been used to evaluate the algorithms for most semantic segmentation tasks. However, due to the label reacts to the thermal diffusion of the defect after heating rather than the defect itself. Therefore, according to previous IRNDDT work [40], we use IOU and F-Score simultaneously to estimate the segmentation effect and the detection ability of the proposed algorithms. The IOU is formulated as follows:

$$IOU = \frac{|Y_i \cap \hat{Y}_i|}{(|\hat{Y}_i| + |Y_i| - |Y_i \cap \hat{Y}_i|)} \quad (9)$$

The events based on F-Score is expressed as:

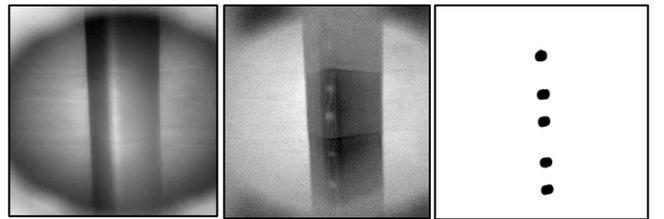
$$F = (\beta_f^2 + 1)(P \times R) / ((\beta_f^2 \times P) + R) \quad (10)$$

where P is precision and R is recall given by:

$$P = TP / (TP + FP) \quad (11)$$

$$R = TP / (TP + FN) \quad (12)$$

where the TP is true positive, which means that the defect is existed and is detected; FP is false positive, meaning no defect exists but is detected by the model; FN is false negative, which denotes a defect exists but is not detected; TN is true negative, which denotes no defect exists and none is detected. β_f is the weight of the Precision and Recall. For the thermal imaging



(a) Original image (b) processed image (c) Ground truth
Fig. 9. A thermal image, its processed image and ground truth

debond diagnostic, the value of β_f is set to 2, which is mean that Recall is more important than Precision, because we want all defects can be detected in order to avoid damage to materials and equipment due to the existence of a certain ignored defect, resulting in heavy losses. In practice, we may sacrifice a certain degree of accuracy to detect all the defects, and then this is followed by resorting to human judgments.

III. EXPERIMENT AND RESULT ANALYSIS

A. Experimental Setup and Samples Preparation

Experiments are carried out in high-precision Optical Pulsed Thermography (OPT) system and in Portable OPT (POPT) system. As shown in Fig. 10(a), the first OPT system has higher precision than portable devices. In the experiments, the infrared thermal camera is IR camera (A655sc) that is used to collect thermal image sequences and the thermal sensitivity is 0.05°C. The maximum resolution of thermal images is (480,640). The halogen lamps with a power of 2kW is applied as an excitation source and the excitation time can be controlled by the excitation source with a maximum power of 3kW.

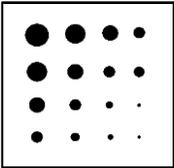
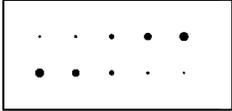
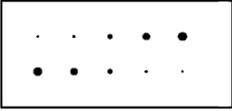
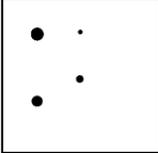
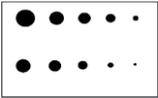
In order to verify the robustness of proposed model under the different experimental conditions and specimens, we use POPT system to do the experiments, which is shown in Fig. 10(b). It

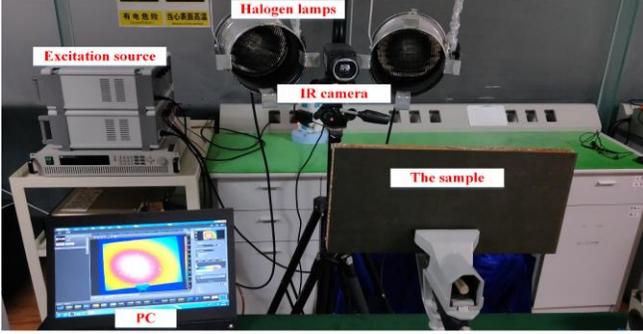
includes an excitation source and a CPU, a display screen, a halogen lamp with a power of 800 W, an IR camera (MAGNITY MAG-62), and a grip. The temperature sensitivity is 0.06°C, and the maximum resolution is (288,384). Thus, the resolution of the thermal images from POPT system are lower than these from OPT system.

Seven different specimens and corresponding descriptions are shown in Table I. Six of them are collective heat type samples while the last one is insulated heat type sample. Among the first six specimens, the sample No.1 is a carbon fiber composite with 16 sub-surface defects of different sizes. The sample No. 3 is a curved surface, which is a little more difficult to detect than the sample No. 2 and No. 4. The samples No. 5 and No. 6 are CFRP with curved shape, which are R-area type material. The defects of the R-type specimen are located in the bend, and it presents considerable difficulty for detection due to its irregular structure. In the experiments, the power of lamps, the positions of the samples, and the excitation time are changed to get different infrared thermal volumetric data to train and test the proposed model.

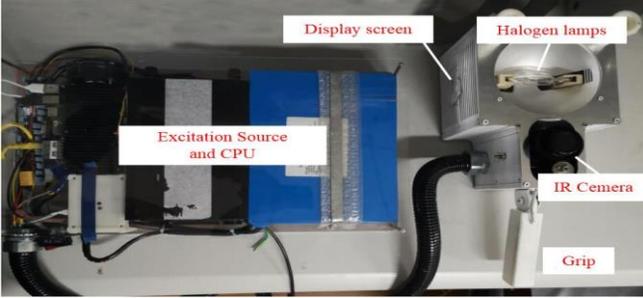
There are four infrared thermal volumetric datasets used in the experiments, which are dataset-one, flat type data from OPT system; dataset two, R-area type data from OPT system; dataset three, R-area type data from POPT system, and dataset four, different R-area type data from POPT system.

TABLE I
DESCRIPTIONS OF SEVEN CFRP SPECIMENS

Types	Samples number	Indication	Size (mm)	Defect information (mm)	Picture
CFRP (collective heat type)	1		250×250×2.2	Depth: 1 or 1.2 Diameter: 2,4,6,8,10,12,16,20	
CFRP (Coating material)	2		150×300×2	Depth: 1 or 1.2 Diameter: 3,5,7,10,12	
	3		150×300×2	Depth: 1 or 1.2 Diameter: 3,5,7,10,12	
	4		150×150×2	Depth: 1.2 Diameter: 3,5,7,1	
CFRP (R-area type)	5		100×100×8	Depth: 0.5, 0.75, 1, 1.25, 1.5 Diameter: 9, 10	
	6		100×100×8	Depth: 0.5, 0.75, 1, 1.25, 1.5 Diameter: 6, 8	
CFRP (radiant heat type)	7		250×300×14.2	Depth: 2 or 2.2 Diameter: 3,6,10,14,18	



(a) OPT system



(b) POPT system

Fig. 10. Experimental systems

TABLE II
SOME PARAMETERS OF THE PROPOSED MODEL

Hyper-Parameter	Value
Batch Size(b)	2
b_t	30
s	160
t_0, h_0, w_0	16, 192, 192
k_c	5 if $c > 128$; else 3
k_t, k_s	2,5 if $t > 4$; else 1,3
β_w	1

B. Implementation

The proposed model is implemented based on the Keras library with TensorFlow backend. The experiments are carried out on two different datasets from OPT and POPT systems, and the datasets have different types of specimens. For the obtained infrared thermal data $\mathbf{D}_i \in \mathbb{R}^{m \times n \times f}$, the maximum (m, n) is (480,640) while f depends on the length of time of the heating and cooling process. In general, $m > 200$, $n > 200$, and $f > 180$.

The proposed model are trained on these different datasets with identical optimization. During the learning process, the Adam optimizer is set up with learning rate of 1×10^{-4} , which would decrease when the metrics stop improving after five epochs, and the lower bound on the learning rate is 1×10^{-6} . The maximum epoch is 60 and training would be early stopped if the monitored validation loss does not exceed the minimum change of 1×10^{-4} after 10 epochs. The values of some

parameters appeared in this paper are shown in Table II. All the experiments are conducted on an NVIDIA GTX 1080 Ti GPU with 11GB RAM.

C. Impact Analysis of PCA and Sequence-PCA Layer

PCA algorithm is often used to process the infrared thermal data to help annotators judge the locations defects. Three slices of different data with six PCA components are visualized in Table III. Red arrows in the original image indicate the locations of defects. In addition, the components with red borders have a clearer appearance of defects.

We visualized the outputs from Sequence-PCA Layer 1 as shown in Table IV with the same six components for the first batch of input data which has been preprocessed. The first column in Table IV is the slices of preprocessed data, the preprocessing operations are introduced in Section II.A. The components with red borders also indicate that have a better appearance of defects. From Table IV, we can clearly see that the desired information almost all in the first four principal components, so we get **the first** four channels in the program. Although the Sequence-PCA layer only processes batches of data with shorter time sequence, however, the results are similar to the images shown in Table III.

When added Sequence-PCA layer to the baseline model, the results have the largest achievements for IOU and F-score, increased by 15.59, 4.71% respectively, which can be seen in Table VI. This shows that the Sequence-PCA layer is meaningful for the final segmentation results.

In addition, a stcA module is added after the concatenation of the feature maps from Sequence-PCA layer and the *Dimension Reduction* cell. Different weights of these channels can be visualized which are shown in Fig. 11. The weights are the mean channels' weights on the test datasets, the last four weights are multiplied with the tensors from Sequence-PCA layer while the first thirty-two weights are for the tensors from that cell. From Fig.11, we can see that the feature maps from Sequence-PCA layer have effects on the outputs, whose weights are not near to zero.

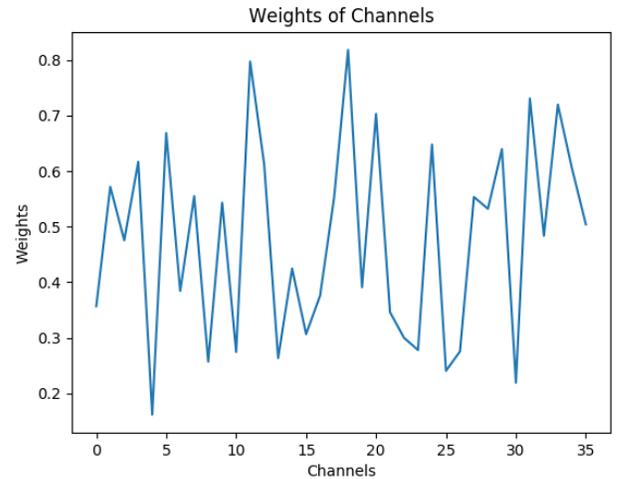


Fig. 11. Curve of different weights of the channels

TABLE III
SOME RESULTS OF PCA

Original image	First component	second component	Third component	Fourth component	Fifth component	Sixth component

TABLE IV
SOME RESULTS OF SEQUENCE-PCA LAYER 1

Preprocessed image	First component	second component	Third component	Fourth component	Fifth component	Sixth component

TABLE V
RESULTS OF DATA FROM OPT SYSTEM

Original image	Ground truth	Simplified structure	Original image	Ground truth	Simplified structure

The model can be simplified through Sequence-PCA layer to achieve the purpose of fast detection for the industrial applications. For example, defects in flat type data is easily to be detected as very deep networks may be not handled in this case. By only keeping the Input, Output, two Convolution and Sequence-PCA layers of the proposed model, we get the simplified structure, which is shown in Fig 12. This simplified structure is directly trained to detect defects for flat type data from OPT system, some results are showed in Table V. We can see that this simplified structure with only about 3K parameters has some effect on the obvious defects segmentation.

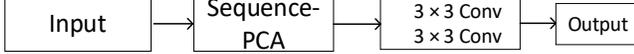


Fig. 12. A simplified structure

D. Effects of stcA Module

1) Defects Segmentation and detection

We compare designed stcA block with several state-of-the-art methods by using the UNet++L4 as backbone model, including SENet [30], scSENet [29], and ECANet [35]. The evaluation metrics concern both efficiency (inference time) and effectiveness (IOU, F-score). The results are the mean of 5-fold cross-validated results on the four different infrared thermal volumetric datasets whose details can be seen in Section III.A. We run them on the same platform and the results are given in Table VI.

TABLE VI
SEGMENTATION AND DETECTION RESULTS OF DIFFERENT METHODS (INFERENCE TIME IS FOR PER IMAGE)

Models	Inference time	Param.	IOU	F-score
3D-UNet++	0.337ms	12.75M	50.091	84.34%
3D-UNet++- SENet	0.346ms	12.80M	51.304	84.88%
3D-UNet++- scSENet	0.348ms	12.81M	51.689	86.42%
3D-UNet++- ECANet	0.340ms	12.75M	52.120	86.01%
3D-UNet++- stcANet($k_t=k_s=1$)	0.341ms	12.75M	52.680	86.70%
3D-UNet++- stcANet (ours)	0.343ms	12.78M	52.892	86.98%
3D-UNet++- Sequence PCA (ours)	0.340ms	12.75M	65.682	89.05%

From Table VI, we can draw conclusions that all the attention blocks can have certain improvements on IOU and F-score, especially for the stcANet with k_t and k_s more than one. The

results are achieving 2.8 and 2.64% gains where it is better than other attention modules in the task of defects detection and segmentation.

2) Classification

We compare our stcA module with the above methods on the public NEU multi-class steel defect dataset [41]. NEU-CLS-64 consists six classes (crazing, inclusion, patches, pitted surface, rolled-in-scale, and scratches) with each image size (64, 64) and 62924 images in total. We used 10-fold cross-validation for training, which accounts for 70%, the rest are for testing. The NEU surface defect database includes two difficult challenges. The first one is the same class of defects existing large differences in appearance, for instance, the scratches may be horizontal, vertical, or slanting scratch, while the different classes of defects have similar aspects, e.g., rolled-in scale, crazing, and pitted surface. The second challenge is the defect images suffer from the influence of illumination and material changes [42].

Performance comparisons are illustrated in Table VII. The stcA module achieves the state-of-the-art Top-1 accuracy in the NEU-CLS-64 dataset with the pretrained DenseNet121 [43] as the backbone model. We can draw a conclusion that stcANet which combines the channel and spatiotemporal information to recalibrate feature maps can gain a better performance. In the experiments, the values of k_t and k_s have been set more than one, in order to combine the information from neighbors and weighted channels with a greater receptive field (so as to recalibrate the feature maps).

In addition, we evaluate the proposed stcANet on CIFAR-10 dataset by using DenseNet121 [43] as the backbone model. Through 10-fold cross-validation, the stcANet has obtained 0.5% higher than scSENet in accuracy.

TABLE VII
PUBLIC NEU DATASETS RESULTS. IN EXPERIMENTS, BATCHSIZE IS SET TO 32 AND IMAGE SIZE IS (64 × 64). (INFERENCE TIME IS FOR PER IMAGE)

Models	Top-1 accuracy	Epochs.	Learning rate	Inference time
SurfNet [44]	0.995	100	0.0007	1.9ms
MultiVis [44]	0.984	50	0.001	3.4ms
FastInf [44]	0.944	100	0.001	0.2ms
DenseNet121- ECANet	0.9974	10	0.0001	0.016s
DenseNet121- scSENet	0.9976	10	0.0001	0.016s
DenseNet121- stcANet($k_t=k_s=1$)	0.9980	10	0.0001	0.016s
DenseNet121- stcANet	0.9984	10	0.0001	0.016s

E. Results of Model Compression

In Fig. 7, there are 16 options for \mathbf{Z} , the results of the different values of \mathbf{Z} are slightly different. The final model chooses \mathbf{Z} as eight. As for the candidate decoders, \mathbf{X} is chosen from $P3D-A+stcA$ block, $P3D-A$, and $Identity$. In the case that the number of additional parameters does not increase considerably, the final model chooses $P3D-A+stcA$ for all decoders. These blocks enable the model to extract more semantic information and gain a better performance for the results. Table VIII shows the comparisons of the different \mathbf{Y} for the loss and accuracy on the validation dataset, and F-score is

obtained as the mean of 5-fold cross-validated results on four different infrared thermal datasets.

TABLE VIII
METRICS FOR DIFFERENT Y

Y	Parameters	Loss	Accuracy	F-score
Output2	12.78M	1.1383	0.9936	89.99
Output1	3.14M	1.3301	0.9914	89.67

From Table VIII we can see that *Output 2* has better performance than *Output 1* in a small range, however, the parameters of *Output 1* is about one quarter of *Output 2*. Due to resolution loss after pooling, it is generally difficult to optimize the number of pooling operations in order to extract high level global features in such encoder-decoder architectures [45]. Therefore, we choose *Output 1* in the final model considering the model simplicity as well as some small size of defects. The framework of the proposed model is shown in Fig. 2.

F. Model Results and Analysis

In this section, in order to evaluate the proposed algorithm, several prevalent semantic segmentation deep learning algorithms have been selected for comparison. These methods consist of TerausNet [46], UNet++ [13], H-DenseUNet [11], and RVOS [47]. The same training sets are used to train each network. It should be noted that TerausNet and UNet++ predict defects in the spatial domain, except that the others are in spatiotemporal domain. The segmentation results of each network are given in Table IX and Table X, respectively.

For better visualization, we combine background information and segmentation results together. The visual result \hat{I} is got by:

$$\hat{I} = I + \sim B(\hat{Y}) \quad (13)$$

where I is a slice of the preprocessed volumetric data, \hat{Y} is a predicted image for I , and $\sim B$ means binary inversion using 0.5 as the threshold.

The defective specimens used as test datasets are considered challenging to detect for many current methods. As shown in Table IX, for the specimens No.1-1 and No.1-2, they are flat type data. The specimen No.1-2 is the CFRP with insulated heat type, having the minimum categories, obtained the worst results compared to other flat samples. Despite this, these deep neural networks can extract features effectively in detecting defects for flat type data from OPT system which have more obvious defect characteristics. The defects in the thermal images are relatively easy for human to discern. In terms of the R-area type data from OPT system, the challenge to detect such defects remains because the background information drowns the defect information and the defects are in the elbow of the samples. For the models of H-DenseUNet and TerausNet, both have mistakenly treated the elbow area as the defects. The RVOS model can extract time information and the UNet++ model adds much low semantic information under the high noise of interference, so they have better results than the H-DenseUNet and TerausNet models. As for the proposed model, besides of the better segmentation performance for the flat type data, it also shows the excellent detection capability for the R-area type data from OPT system.

TABLE IX
VISUAL RESULTS OF DATA FROM OPT SYSTEM

No.	Original image	Ground truth	TerausNet	H-DenseUNet	RVOS	UNet++	Proposed
1-1							
1-2							
2-1							
2-2							

TABLE X
VISUAL RESULTS OF DATA FROM POPT SYSTEM

No.	Original image	Ground truth	TernausNet	H-DenseUNet	RVOS	UNet++	Proposed
3-1							
3-2							
4-1							
4-2							

TABLE XI
F-SCORE OF COMPARISON RESULTS (%)

dataset	TernausNet	H-DenseUNet	RVOS	UNet++	Proposed
1	81.67	83.97	76.10	91.54	93.33
2	37.30	30.81	88.49	83.33	90.42
3	0	0	64.72	89.04	94.25
4	0	0	0	76.02	82.48
average	26.34	25.19	54.16	84.34	89.67

To validate the robustness of the proposed model, we use different R-area type data from POPT system, which is limited by the resolution of the IR camera. The visual results are shown in Table X. It is shown that H-DenseUNet and TernausNet models have no effect on detecting defects for such hard type data from POPT system. In addition, these defects are considered challenging even for annotators to recognize. The RVOS model can distinguish some defects for the specimens No.3-1 to No.3-2, but fails on the specimens No.4-1 to No.4-2 which suffer heavily from high noise and low resolution. UNet++ can detect defects well on dataset three, whereas it has difficulty in detecting defects on dataset four. Among these methods, the proposed model has yielded good performance on these difficult-to-detect specimens as shown by the detected defects which are more comprehensive and clearer than other methods.

Table XI shows the F-score on the test datasets, which are obtained as the mean of the 5-fold cross-validation results. For the R-area type data from POPT system, the F-score of TernausNet and H-DenseUNet is 0%, which means that these

algorithms failed to detect defects, so their averaged F-score is less than 30%. RVOS model have a better situation with averaged F-score 54.16%, and UNet++ method ranks the second. The proposed model gives the highest capability on detecting defects with the averaged F-score 89.67%. Therefore, the proposed method is better than other methods in terms of detection ability, which is more robust and adaptable.

IV. CONCLUSION AND FEATURE WORK

In this paper, a hybrid multi-dimensional spatial and temporal segmentation model is proposed to detect defects for the infrared thermal volumetric data. In addition, a new Sequence-PCA layer and attention modules have been developed. The proposed model has been tested on four different datasets from POPT and OPT systems with different depths and diameters of defects. Compared with other segmentation models, the proposed model has yielded considerable better performance as shown by the F-score and IOU. Moreover, the proposed model has only 3.14 million

parameters, which is lightweight compared with other deep networks. The stcA block has also been evaluated with other attention modules on two classification tasks and the obtained results have shown promising prospects. Future research will focus on the target reconstruction for the large-scaled specimens and the separation of the background and defects.

V. ACKNOWLEDGEMENT

The work was supported by Defense Industrial Technology Development Program (Grant No. JSZL2019205C003), National Natural Science Foundation of China (No. 61971093, No. 61527803, No. 61960206010). The work was supported by Science and Technology Department of Sichuan, China (Grant No.2019YJ0208, Grant No.2018JY0655, Grant No. 2018GZ0047) and Fundamental Research Funds for the Central Universities (Grant No. ZYGX2019J067).

REFERENCES

- [1] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding Deep Learning Techniques for Image Segmentation," *ACM Comput. Surv.*, vol. 52, no. 40, pp. 1–58, 2019.
- [2] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.
- [3] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 2980–2988, 2017.
- [4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, pp. 834–848, 2018.
- [6] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *arXiv:1706.05587*, p. 14, 2017.
- [7] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," in *International Conference on Learning Representations*, 2016, p. 13.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, pp. 1904–1916, 2015.
- [9] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with Gaussian edge potentials," in *25th Annual Conference on Neural Information Processing Systems*, 2011, pp. 1–9.
- [10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.
- [11] X. Li, H. Chen, X. Qi, Q. Dou, C. W. Fu, and P. A. Heng, "H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes," *IEEE Trans. Med. Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.
- [12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [13] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11045 LNCS, pp. 3–11, 2018.
- [14] N. Srivastava, "Unsupervised Learning of Video Representations using LSTMs," *arXiv:1502.04681*, p. 12, 2014.
- [15] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in *Neural Information Processing Systems Conference*, 2015, pp. 1–11.
- [16] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv:1803.01271v1*, 2018. .
- [17] D. Maturana and S. Scherer, "VoxNet: A 3D Convolutional Neural Network for real-time object recognition," *IEEE Int. Conf. Intell. Robot. Syst.*, vol. 2015-December, pp. 922–928, 2015.
- [18] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *4th International Conference on 3D Vision*, 2016, pp. 565–571.
- [19] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9901 LNCS, pp. 424–432, 2016.
- [20] H. Fernandes, H. Zhang, A. Figueiredo, C. Ibarra-Castanedo, G. Guimaraes, and X. Maldague, "Carbon fiber composite inspection and defect characterization using active infrared thermography: numerical simulations and experimental results," *Appl. Opt.*, vol. 55, no. 34, p. D46, 2016.
- [21] P. G. Bison, S. Marinetti, E. G. Grinzato, V. P. Vavilov, F. Cernuschi, and D. Robba, "Inspecting thermal barrier coatings by IR thermography," *Thermosense XXV*, vol. 5073, no. April, p. 318, 2003.
- [22] S. Marinetti *et al.*, "Statistical analysis of IR thermographic sequences by PCA," *Infrared Phys. Technol.*, vol. 46, no. 1-2 SPEC. ISS., pp. 85–91, 2004.
- [23] N. Rajic, "Principal component thermography for flaw contrast enhancement and flaw depth characterisation in composite structures," *Compos. Struct.*, vol. 58, no. 4, pp. 521–528, 2002.
- [24] X. Maldague and S. Marinetti, "Pulse phase infrared thermography," *J. Appl. Phys.*, vol. 79, no. 5, pp. 2694–2698, 1996.
- [25] C. Ibarra-Castanedo and X. P. V. Maldague, "Interactive methodology for optimized defect characterization by quantitative pulsed phase thermography," *Res. Nondestruct. Eval.*, vol. 16, no. 4, pp. 175–193, 2005.
- [26] Q. Luo, B. Gao, W. L. Woo, and Y. Yang, "Temporal and spatial deep learning network for infrared thermal defect detection," *NDT E Int.*, vol. 108, no. August, p. 102164, 2019.
- [27] Z. Qiu, T. Yao, and T. Mei, "Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 5534–5542, 2017.
- [28] Z. Zhu, C. Liu, D. Yang, A. Yuille, and D. Xu, "V-NAS: Neural Architecture Search for Volumetric Medical Image Segmentation," *arXiv:1906.02817v1*, pp. 1–9, 2019.
- [29] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11070 LNCS, pp. 421–429, 2018.
- [30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 7132–7141, 2018.
- [31] S. Kumawat and S. Raman, "LOCAL PHASE U-NET FOR FUNDUS IMAGE SEGMENTATION," in *International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 1209–1213.
- [32] S. Kumawat and S. Raman, "LP-3DCNN: Unveiling Local Phase in 3D Convolutional Neural Networks," *arXiv:1904.03498v1*, 2019.
- [33] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *J. Mach. Learn. Res.*, vol. 15, no. January, pp. 315–323, 2011.
- [34] K. Jansen and H. Zhang, "ImageNet Classification with Deep Convolutional Neural Networks," *NIPS Proc.*, vol. 1, pp. 1097–1105, 2012.
- [35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," *arXiv:1910.03151v1*, 2019.
- [36] M. Zeiler and R. Fergus, "Regularization of Neural Networks using DropConnect," in *International Conference on Machine Learning*, 2013, no. 28, pp. 1058–1066.
- [37] Y. He, J. Lin, Z. Liu, H. Wang, L. J. Li, and S. Han, "AMC: AutoML for model compression and acceleration on mobile devices," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11211 LNCS, pp. 815–832, 2018.

- [38] C. Liu *et al.*, “Auto-DeepLab: Hierarchical Neural Architecture Search for Semantic Image Segmentation,” *arXiv:1901.02985v2*, 2019.
- [39] H. Liu, K. Simonyan, and Y. Yang, “DARTS: Differentiable Architecture Search,” *arXiv:1806.09055v2*, pp. 1–13, 2018.
- [40] Y. Wang *et al.*, “Thermal pattern contrast diagnostic of microcracks with induction thermography for aircraft braking components,” *IEEE Trans. Ind. Informatics*, vol. 14, no. 12, pp. 5563–5574, 2018.
- [41] K. Song and Y. Yan, “A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects,” *Appl. Surf. Sci.*, vol. 285, no. PARTB, pp. 858–864, 2013.
- [42] K. Song, S. Hu, and Y. Yan, “Automatic recognition of surface defects on hot-rolled steel strip using scattering convolution network,” *J. Comput. Inf. Syst.*, vol. 10, no. 7, pp. 3049–3055, 2014.
- [43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017, vol. 2017-Janua, pp. 2261–2269.
- [44] S. Arikan, K. Varanasi, and D. Stricker, “Surface Defect Classification in Real-Time Using Convolutional Neural Networks,” *arXiv:1904.04671v1*, 2019.
- [45] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing, “Modified U-Net (mU-Net) with Incorporation of Object-Dependent High Level Features for Improved Liver and Liver-Tumor Segmentation in CT Images,” *IEEE Trans. Med. Imaging*, vol. 00, no. 00, pp. 1–1, 2019.
- [46] V. Iglovikov, S. Seferbekov, A. Buslaev, and A. Shvets, “TernausNetV2: Fully convolutional network for instance segmentation,” *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2018-June, pp. 228–232, 2018.
- [47] C. Ventura, M. Bellver, A. Girbau, A. Salvador, F. Marques, and X. Giro-i-Nieto, “RVOS: End-to-End Recurrent Network for Video Object Segmentation,” *arXiv:1903.05612v2*, 2019.