Infrared thermal defect identification and reconstruction of artworks using a spatiotemporal deep neural network

M Moradi[1,2,*]**, R Ghorbani**[3]**, S Sfarra**[4]**, D M J Tax**[3] **and D Zarouchas**[1,2]

[1] Structural Integrity & Composites Group, Aerospace Engineering Faculty, Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands
[2] Center of Excellence in Artificial Intelligence for structures, prognostics & health management, Aerospace Engineering Faculty, Delft University of Technology, Kluyverweg 1, Delft, 2629 HS, The Netherlands
[3] Pattern Recognition Laboratory, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands
[4] Department of Industrial and Information Engineering and Economics (DIIIE), University of L'Aquila, Piazzale E. Pontieri 1, Monteluco di Roio, 67100 L'Aquila, AQ, Italy

[*] M.Moradi-1@tudelft.nl

Abstract. Assessment of cultural heritage assets is now extremely important all around the world. Non-destructive inspection is essential for preserving the integrity of the artworks while avoiding the loss of any precious materials that make it up. The use of Infrared Thermography (IRT) is an interesting concept since surface and subsurface faults can be discovered by utilizing the 3D diffusion inside the object caused by external heat. The primary goal of this research is to detect defects in artworks, which is one of the most important tasks in the restoration of mural paintings. To this end, a spatiotemporal deep neural network (STDNN) is utilized for defect identification in a mock-up reproducing an artwork, taking into account both the temporal and spatial perspectives of step-heating (SH) thermography. Finally, the outcomes are compared to those of other conventional algorithms.

1. Introduction

The preservation of cultural heritage assets has been nowadays attended because they carry valuable information. Therefore, the use of Non-Destructive Testing (NDT) procedures in conservation is highly valued by restorers and art historians [1]. Thermal non-destructive testing is a smart option to inspect cultural heritage objects since surface and subsurface defects can be detected by exploiting the 3D diffusion inside the object induced by external radiation [2, 3]. However, a painting surface is one of the challenging items for Infrared Thermography (IRT) [4] as an NDT method because pigments composing the colors cause the emissivity variations of the surface, which is the most important coefficient in the emitted radiation energy and Stefan–Boltzmann law [5].

External surface treatments, typically given by spray, are frequently employed as a technique to improve the emissivity value in order to increase the thermal contrast of subsurface faults projected on the surface. This strategy, however, is not always appropriate, especially when sensitive layers of items,

such as cultural heritage and artworks, must be investigated [6-8]. Therefore, other techniques without any manipulation of the objects should be considered to identify the defects of an artwork.

Data analysis by taking the heat equations into account can be helpful in extracting valuable knowledge from thermograms recorded by an IR camera [9, 10]. Yet, because an artwork may have complex damage configurations (various sizes, depths, materials, and types), analyzing IR data using only physical models is extremely difficult, especially when access to thermal data is limited to the external surface. Data-driven models and artificial intelligence (AI) are key mathematical approaches to overcome this challenge.

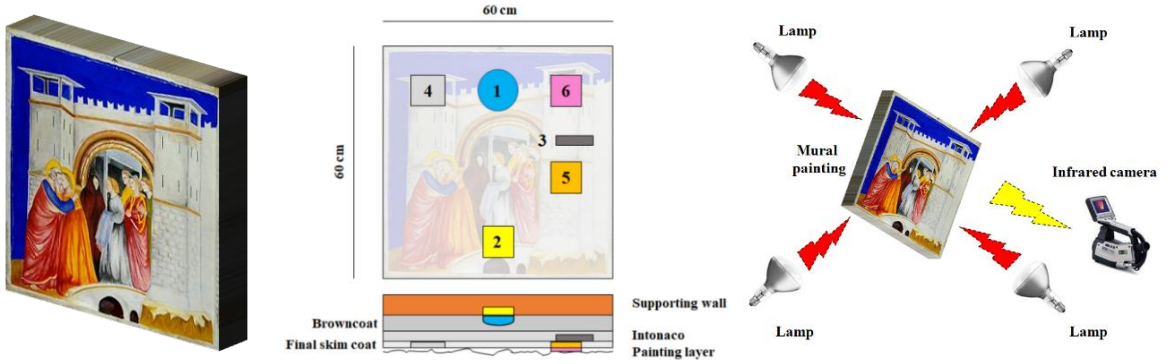
In recent years, infrared machine vision has gained increasing interest in the various domains due to the increasing growth of Machine Learning, notably with Deep Learning (DL) algorithms that use multiple layer networks to extract higher-level features from raw IR input sequentially [11]. The complexity of identifying damage in an artwork can be addressed by employing AI, particularly DL. Despite significant research advances in IRT processing using unsupervised learning, generally employed detection algorithms still have difficulties in defect identification due to weak signal-to-noise ratio (SNR), complicated interference, and so on. The development of supervised learning to research IRT is a prospective trend based on the spatial-temporal physic properties of the IRT sequences [12]. To this end, thermal video can be analyzed from two perspectives:

1. The aspect of temporal information that includes the temperature variation of each pixel over time, and could be regarded as a time-series input.
2. The spatial information aspect, which includes the temperature variation of each frame (at one moment) over all pixels and could be regarded as an image input.

In the present work, a Machine Learning Framework based on the Deep Neural Networks (DNN) is designed to classify pixels into healthy and defective regions, presenting the pertinent intact and damaged areas of the object under inspection. The proposed framework consists of two sub-models: a multilayer perceptron (MLP) to classify each time-series (1D signal) into healthy or defective pixels; and a convolutional neural network (U-Net) to segment images into healthy or defective areas. These two networks are fused sequentially together in order to enhance the performance in such a way that after training the former one, the latter one is trained. The developed framework's performance is compared to the results of popular algorithms such as Pulsed Phase Thermography (PPT), Principal Component Thermography (PCT), and Thermographic Signal Reconstruction (TSR).

2. Experimental setup

The IRT-inspected artwork is a replica of Giotto's "Meeting at the Golden Gate" (a mural painting) that is preserved in Padua's Scrovegni Chapel (Italy). The size of the replica is 60 × 60 cm. The sample contains several faults at various layers, indicating typical degrading mural painting faults. A photograph of the replica, a map of defects, and a sketch of the experimental setup for the laboratory IRT inspection can be seen in figure 1. Details about the different fabricated defects and more information were provided in [13].



(a) (b) (c)

Figure 1. (a) Photograph of the mock-up, (b) defects map, and (c) sketch of the experimental IRT setup.

The sample was stimulated by four halogen lamps (OSRAM SICCATHERM, 250 W), and the thermal response of the surface was recorded by an infrared camera (FLIR S65 HS, 7.5–13 μm, 320×240, 50 Hz). The heating and cooling phases lasted 52.5 and 164 seconds, respectively (for a total of around 216.5 seconds). The final thermography dataset contains 10826 thermal images, of which the frames with a 25 Hz rate will be collected for further analysis to reduce processing costs, and the total number of final frames is 434. In addition, each image is cropped to remove the additional marginal pixels, reducing the size of each frame to 230×230. As a result, the network's input data set size is 230×230×434.

3. Methodology

This section outlines the proposed framework, a spatiotemporal deep neural network (STDNN), which is depicted in figure 2. The framework is divided into two parts: temporal and spatial sub-models, which will be discussed in detail in the following subsections.

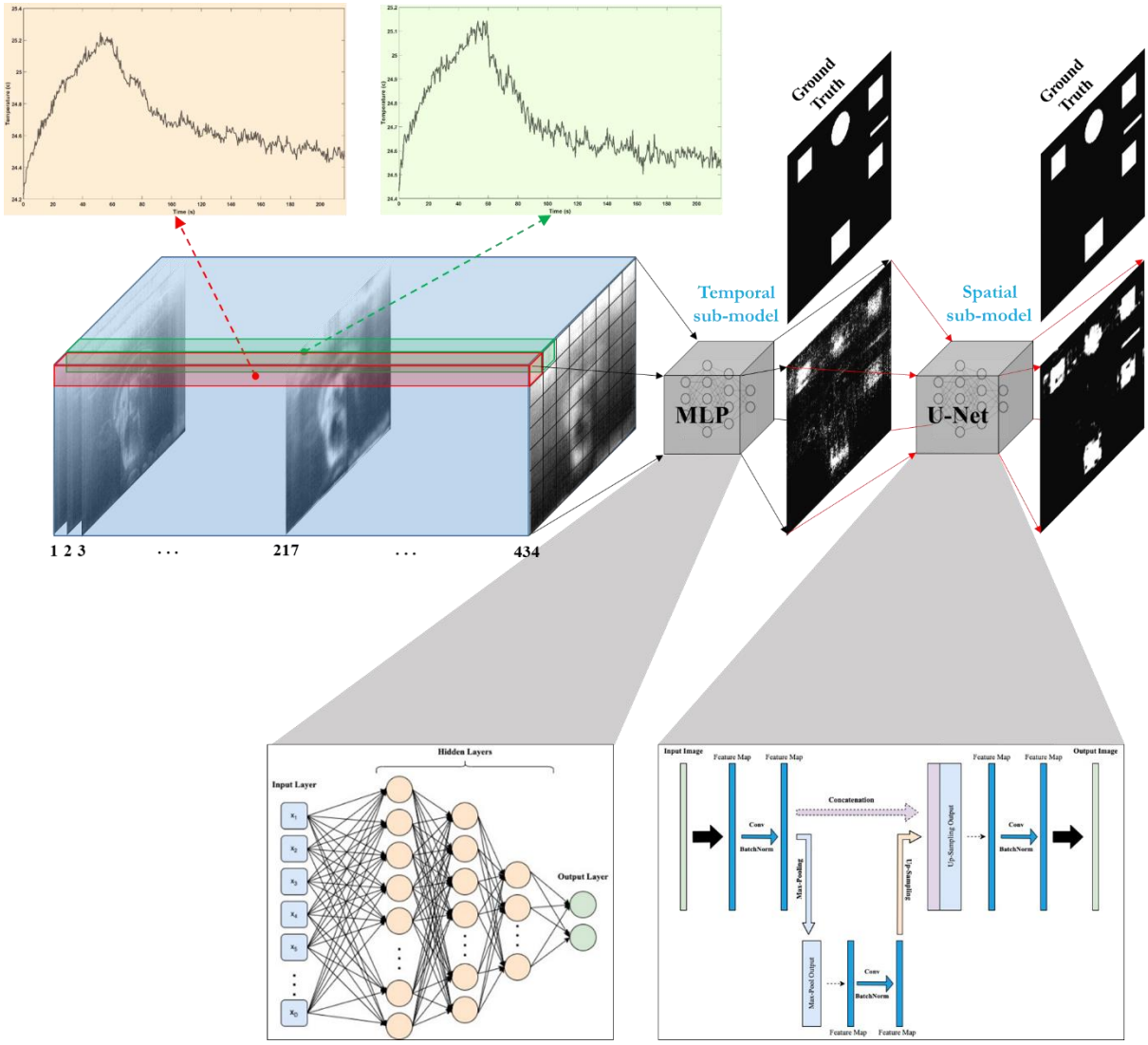


Figure 2. The proposed spatiotemporal deep neural network (STDNN).

3.1. Temporal network

For the temporal information, a multilayer perceptron (MLP) is used to classify the temporal signals related to the pixels into healthy or defective. First, the thermal signals relevant to each pixel are labeled

using the ground truth image, which was built artificially using the actual locations and dimensions of the defects. The dataset is then windowed across the spatial point of view by a mask with a size of 10-by-10 pixels, which is a patch of neighboring pixels, and the moving stride is 10 pixels in both directions of the image's height and width (i.e., zero overlap). As a result, the dataset is composed of 23×23 patches of 3-dimensional data, each of which is 10×10×434 pixels in size. This type of partitioning was carried out because the dataset needed to be divided into training and test datasets from only one specimen, and because pixels near to each other in the spatial network needed to be imported into the sub-model for the segmentation task and morphology based on topology. Each thermal signal related to a pixel is imported into the MLP to be classified between zero (healthy) and one (defective).

The temporal sub-model comprises three hidden layers, respectively containing 20, 10, and 5 neurons. The hyperbolic tangent (tan-sigmoid) function is used as the activation function in all hidden layers, whereas the logistic (log-sigmoid) function is used in the output layer. The scaled conjugate gradient backpropagation [14] is adopted as the optimizer, and a binary cross-entropy function with a regularization parameter of 0.1 is used as the loss function.

3.2. Spatial network

A U-Net as a Convolutional Neural Network (CNN) model [15] is employed to segment images into healthy and faulty regions based on the relationships between surrounding pixels in spatial information. The U-Net architecture is composed of an encoder network followed by a decoder network. U-Net, as opposed to a simple autoencoder architecture, has additional interconnections between the encoder and decoder sections. In the present work, this U-Net model is trained on Keras API which is an open-source software library that provides a Python interface for artificial neural networks. It should be noted that the Adam, which is a stochastic gradient descent algorithm according to the adaptive estimation of first- and second-order moments, is used as the optimizer with a learning rate of 0.001. Moreover, binary cross-entropy is the loss function. The U-Net model's architecture is presented in detail in table 1. The default settings are used for the parameter that are not mentioned in the table 1. Patches of 10×10×434 are inputted to the U-Net, and then all resultant patches of 10×10×1, including training and test ones, are concatenated to reconstruct the full image of 230×230.

Table 1. The U-Net (spatial sub-model) model's architecture.

Layer Number	Layer Type	Input of Layer
1	Conv2D (Filters: 128, (3 × 3), Activation: Exponential Linear Unit (ELU)), Kernel_Initializer: he_normal	MLP's output
2	Batch Normalization	Layer 1
3	Conv2D (Filters: 128, (3 × 3), Activation Function: ELU), Kernel_Initializer: he_normal	Layer 2
4	Max-Pooling (Size:(2 × 2))	Layer 3
5	Conv2D (Filters: 128, (3 × 3), Activation Function: ELU), Kernel_Initializer: he_normal	Layer 4
6	Batch Normalization	Layer 5
7	Conv2D (Filters: 128, (3 × 3), Activation Function: ELU), Kernel_Initializer: he_normal	Layer 6
8	Conv2D-Transpose (Filters: 128, (3 × 3), Strides = (2 × 2), Activation Function: ELU)	Layer 7
9	Conv2D (Filters: 128, (3 × 3), Activation Function: ELU), Kernel_Initializer: he_normal	Concatenation of Layers 3 and 8
10	Batch Normalization	Layer 9
11	Conv2D (Filters: 128, (3 × 3), Activation Function: ELU), Kernel_Initializer: he_normal	Layer 10
12	Dropout (0.3)	Layer 11
Final Layer	Conv2D (Filters: 1, (1 × 1), Activation: Sigmoid)	Layer 12

4. Results and discussion

The dataset, including 529 (23×23) patches, is divided into training and test datasets with a ratio of 7:3. As a result, the training and test datasets are composed of 370 and 159 patches, respectively, with each

patch size of 10×10×434. Having only one mock-up reproduced artwork was one of the biggest challenges in this research. The dataset is randomly split with a 7:3 ratio and repeated ten times to investigate the model's stability. The temporal sub-model has been evaluated on the test dataset after training the MLP network on the training dataset for 1000 epochs. Since this is a Class-Imbalanced problem, the AUC value, which is the area under the ROC (receiver operating characteristic) curve, is provided to analyze the performance of the sub-models. The AUC averaged over ten repeats across ten datasets of the MLP is presented in table 2. The results on the test set demonstrates the stability and high performance of MLP sub-model in classifying the pixels. The images obtained by the temporal sub-model (MLP) from the first dataset and the first repeat can be seen in figure 3 (the 2nd line).

Table 2. AUC calculated from the (temporal sub-model) MLP's results for the ten generated datasets.

Dataset	1	2	3	4	5	6	7	8	9	10	Mean ± Pop std
Training	0.96±0.01	0.96±0.01	0.96±0.01	0.96±0.01	0.97±0.01	0.97±0.00	0.97±0.00	0.96±0.00	0.96±0.01	0.96±0.01	0.96±0.00
Test	0.86±0.01	0.89±0.01	0.87±0.01	0.88±0.01	0.88±0.00	0.87±0.01	0.89±0.01	0.87±0.01	0.89±0.01	0.86±0.01	0.88±0.01

The first repeat of the MLP's outputs from each dataset has been selected as the input of the U-Net model. To reduce the uncertainty in performance of the U-Net model, this model is implemented ten different times on each of these datasets, and the mean performances with their deviations are presented. This random repeat technique will help to find out how stable the spatial sub-model is with its performance.

Table 3 shows the U-Net model's performance across all ten datasets. Due to the Class-Imbalanced condition in this experiment, AUC and F-1 score evaluation metrics are calculated to investigate the performance of the predictive model. Furthermore, the evaluation metrics of Recall and Precision are computed for more detailed information regarding the model's performance. It should be noted that Macro-Averaged is used to calculate the Recall, Precision, and F1-Score metrics.

Table 3. AUC, Precision, Recall, and F1-Score calculated from the (spatial sub-model) U-Net's results for the ten generated datasets.

Dataset	AUC		Precision (Specificity)		Recall (Sensitivity)		F1-Score	
	Train	Test	Train	Test	Train	Test	Train	Test
1	1.00 ± 0.00	0.94 ± 0.00	0.97 ± 0.01	0.82 ± 0.03	0.96 ± 0.02	0.75 ± 0.03	0.96 ± 0.01	0.78 ± 0.01
2	1.00 ± 0.00	0.95 ± 0.01	0.91 ± 0.02	0.88 ± 0.02	0.97 ± 0.01	0.85 ± 0.02	0.93 ± 0.01	0.86 ± 0.01
3	1.00 ± 0.00	0.94 ± 0.00	0.95 ± 0.02	0.83 ± 0.04	0.94 ± 0.01	0.78 ± 0.03	0.95 ± 0.01	0.80 ± 0.01
4	0.99 ± 0.00	0.93 ± 0.01	0.92 ± 0.02	0.83 ± 0.01	0.95 ± 0.01	0.79 ± 0.03	0.93 ± 0.01	0.80 ± 0.02
5	0.99 ± 0.00	0.95 ± 0.00	0.95 ± 0.01	0.85 ± 0.02	0.94 ± 0.02	0.81 ± 0.02	0.94 ± 0.01	0.82 ± 0.01
6	1.00 ± 0.00	0.94 ± 0.00	0.89 ± 0.03	0.83 ± 0.02	0.97 ± 0.01	0.81 ± 0.02	0.93 ± 0.02	0.82 ± 0.01
7	1.00 ± 0.00	0.95 ± 0.00	0.96 ± 0.02	0.89 ± 0.03	0.96 ± 0.01	0.83 ± 0.03	0.96 ± 0.01	0.85 ± 0.01
8	0.99 ± 0.00	0.93 ± 0.00	0.92 ± 0.01	0.83 ± 0.02	0.90 ± 0.04	0.74 ± 0.05	0.91 ± 0.02	0.77 ± 0.03
9	1.00 ± 0.00	0.95 ± 0.00	0.96 ± 0.01	0.87 ± 0.02	0.94 ± 0.02	0.81 ± 0.01	0.95 ± 0.01	0.84 ± 0.01
10	1.00 ± 0.00	0.93 ± 0.00	0.92 ± 0.02	0.81 ± 0.02	0.95 ± 0.02	0.78 ± 0.03	0.94 ± 0.01	0.79 ± 0.01
Mean ± Pop std	1.00 ± 0.00	0.94 ± 0.01	0.94 ± 0.02	0.84 ± 0.03	0.95 ± 0.02	0.80 ± 0.03	0.94 ± 0.01	0.81 ± 0.03

Precision estimates the accuracy for the minority class in the case of a Class-Imbalanced situation since it is a measure that quantifies the amount of correct positive (Minority class which is the defective pixels) predictions. The results indicate that the mean precision on the test set is 0.84 with a low standard deviation, which can be considered an outstanding performance. Although precision is valuable and the findings appear to be excellent, it does not reflect on how many true positive class samples are predicted as belonging to the negative class (Majority Class which is the healthy pixels).

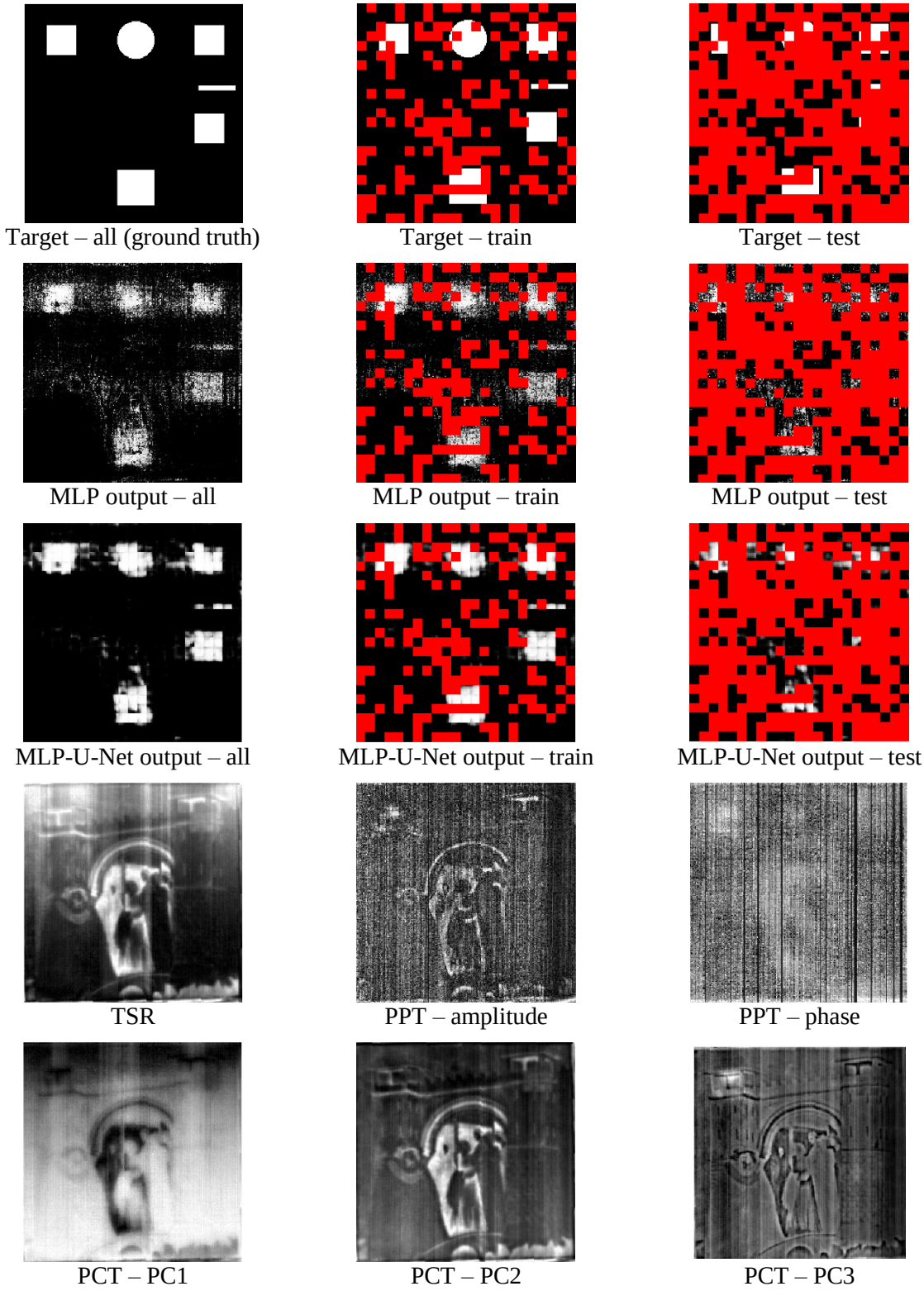


Figure 3. Results of MLP based on only temporal information (the 2nd row), MLP-U-Net based on both temporal and spatial information (the 3rd row), and the well-known algorithms of TSR, PPT, and PCT (the 4th and 5th rows) compared to the ground truth image (the 1st row) for the first dataset. The red color for the first three rows represents the unused pixels for training or test.

Precision only indicates the correct positive prediction out of all positive prediction, but Recall metric, as opposed to precision, shows missed positive predictions. The Recall is a measure that reflects the number of correct positive predictions made out of all possible positive forecasts. The purpose in imbalanced datasets is to optimize Recall without diminishing Precision. Nevertheless, both criteria are often in contradiction, as increases in Recall often lead to the expense of declines in Precision. The F1-score integrates Precision and Recall into a single metric that encompasses both. The overall F1-score of the U-net model on the test set demonstrates that the model performs acceptably in segmenting the image into healthy or faulty regions based on spatial information relationships between surrounding pixels. Above all, the AUC is sensitive to the class imbalance in the view that the minority class has a significant impact on its value. Compared to Accuracy, this is very desirable behavior from the AUC metric. The AUC scores verify that the model's performance is excellent and, more interestingly, stable. Figure 3 (the 3rd line) depicts the images generated by the spatial sub-model (U-Net) once performed on the MLP outputs for the first dataset and first repeat.

To compare, the TSR results, as well as the amplitude and phase of PPT at the frame with the high contrast (maximum kurtosis), are displayed in figure 3. (the 4th line). Figure 3 (the 5th line) depicts the first three principal components (PCs) obtained by PCT. Despite the fact that these popular, conventional algorithms do not require training, their results are not comparable to the MLP-U-Net's, and the effect and reflection of the drawing on the surface are evident. As a result, the proposed framework was able to rectify the emissivity problem induced by pigment effects. However, its shortcoming is that it is supervised and requires training data from similar types of samples, which is not the case in this work.

5. Conclusions

In this work, a spatiotemporal deep neural network (STDNN) was utilized for defect identification in a mock-up reproducing an artwork, taking into account both the temporal and spatial perspectives of SH thermography. Initial results indicated that the mean F1-score evaluation metric is acceptable with a low standard deviation, which can be considered an outstanding performance despite the fact that there is a class-imbalance problem in the data. These results were supported by the AUC scores verifying that the model's performance was excellent and, more interestingly, stable. Finally, the outcomes of the STDNN were compared to those of other conventional algorithms (i.e., PCT, PPT, TSR). It was found that their results cannot be considered comparable to the MLP-U-Net's; for example, the effect and reflection of the drawing on the surface are still evident.

It is possible to say that the proposed framework was able to rectify the emissivity problem induced by pigment effects. For the future, training data from similar types of samples (e.g., panel paintings) will be collected in order to reduce the shortcoming of the proposed STDNN mainly linked to the supervised learning approach.

Acknowledgements

S. Sfarra wish to thank Mr. N. Zaccagnini and Eng. R. Di Biase for the fabrication of the mock-up.

References

[1] R. Shrestha, S. Sfarra, S. Ridolfi, G. Gargiulo, and W. Kim, "A numerical–thermal–thermographic NDT evaluation of an ancient marquetry integrated with X-ray and XRF surveys," Journal of Thermal Analysis Calorimetry, pp. 1-15, 2021.

[2] N. Tao, C. Wang, C. Zhang, and J. Sun, "Quantitative measurement of cast metal relics by pulsed thermal imaging," Quantitative InfraRed Thermography Journal, pp. 1-14, 2020.

[3] K. Liu, K.-L. Huang, S. Sfarra, J. Yang, Y. Liu, and Y. Yao, "Factor analysis thermography for defect detection of panel paintings," Quantitative InfraRed Thermography Journal, pp. 1-13, 2021.

[4] M. Moradi and S. Sfarra, "Rectifying the emissivity variations problem caused by pigments in artworks inspected by infrared thermography: A simple, useful, effective, and optimized approach for the cultural heritage field," Infrared Physics Technology, vol. 115, p. 103718,

2021.
[5] D. Watmough, P. W. Fowler, and R. Oliver, "The thermal scanning of a curved isothermal surface: implications for clinical thermography," Physics in Medicine Biology, vol. 15, no. 1, p. 1, 1970.
[6] A. Chulkov et al., "Evaluating quality of marquetries by applying active IR thermography and advanced signal processing," Journal of Thermal Analysis Calorimetry, vol. 143, no. 5, pp. 3835-3848, 2021.
[7] B. Yousefi, S. Sfarra, C. Ibarra-Castanedo, N. P. Avdelidis, and X. P. Maldague, "Thermography data fusion and nonnegative matrix factorization for the evaluation of cultural heritage objects and buildings," Journal of Thermal Analysis Calorimetry, vol. 136, no. 2, pp. 943-955, 2019.
[8] Y. Yao, S. Sfarra, C. Ibarra-Castanedo, R. You, and X.P.V. Maldague, "The multi-dimensional ensemble empirical mode decomposition (MEEMD): an advanced tool for thermographic diagnosis of mosaics," Journal of Thermal Analysis Calorimetry, vol. 128, no. 3, pp. 1841-1858, 2017.
[9] N. Avdelidis and A. Moropoulou, "Applications of infrared thermography for the investigation of historic structures," Journal of Cultural Heritage, vol. 5, no. 1, pp. 119-127, 2004.
[10] X. P. Maldague, Nondestructive evaluation of materials by infrared thermography. Springer Science & Business Media, 2012.
[11] Y. He et al., "Infrared machine vision and infrared thermography with deep learning: a review," Infrared Physics Technology, vol. 116, p. 103754, 2021.
[12] Q. Luo, B. Gao, W. L. Woo, and Y. Yang, "Temporal and spatial deep learning network for infrared thermal defect detection," NDT and E International, vol. 108, p. 102164, 2019.
[13] S. Sfarra et al., "How to retrieve information inherent to old restorations made on frescoes of particular artistic value using infrared vision?," International Journal of Thermophysics, vol. 36, no. 10, pp. 3051-3070, 2015.
[14] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," Neural networks, vol. 6, no. 4, pp. 525-533, 1993.
[15] 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: Springer.