Journal of Cultural Heritage 58 (2022) 186–198

Contents lists available at ScienceDirect

Journal of Cultural Heritage

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

Original article

Convolutional neural network exploiting pixel surroundings to reveal hidden features in artwork NIR reflectograms

Tomáš Karella[a,1,*], Jan Blažek[a,1], Jana Striová[b,1]

[a] Department of Image Processing, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věž4, Praha, 18000, Czech Republic
[b] National Research Council (CNR), National Institute of Optics (INO),Largo Enrico Fermi 6, Florence, 50125, Italy

ARTICLE INFO

Article history:
Received 21 March 2022
Accepted 19 September 2022

Keywords:
Signal separation
Concealed features visualization
Artwork analysis
Infrared reflectography
Convolutional neural networks

ABSTRACT

Near-infrared reflectography (NIR) is a well-established non-invasive and non-contact imaging technique. The NIR methods are able to reveal concealed layers of artwork, such as a painter's sketch or repainted canvas. The information obtained may be helpful to historians for studying artist technique, attributing an artwork reconstructing faded details. Our research presents the improved method previously developed that reveals the hidden features by removing the information content of the visible spectrum from NIR. Based on convolutional neural networks (CNN), our model estimates the transfer function from visible spectra to NIR, which is nonlinear and specific for painting materials. Its parameters are learnt for particular paintings on the subsamples randomly selected across the canvas, and the model is further utilised to enhance the whole artwork. In addition to the previously developed model, our algorithm exploits each pixel's surroundings to estimate its NIR response. This leads to more precise results and increased robustness to various noises. We demonstrate higher accuracy than the previous method on the historical paintings mock-ups and higher performance on well-known artworks such as Madonna dei Fusi attributed to Leonardo da Vinci.

1. Introduction

The investigation of hidden features in artworks by non-invasive analytical methods is a hot topic in Heritage Science domain. Traditional and innovative tools are being exploited and developed to provide insights into the composition, structure, and other properties of concealed layers. Broadband Infrared Reflectography represents a traditional way of investigating underdrawings and retouches of paintings. With the technological advances, state-of-art devices were developed to acquire the reflected radiation from the painting within narrow spectral bands either in single-point or line-scan imaging modality providing us with 2D multi-spectral/hyperspectral information [1]. The potential of many other imaging techniques - i.e. TeraHertz [2–4], Optical Coherence Tomography [5,6], photoacoustic [7–9] and non-linear [10] - is being currently explored for the inspection of artwork subsurface features.

Our research capitalizes on the results achieved with deep learning in multimodal data processing in other research fields, mainly in medical applications [11] or in consolidating video, audio or text [12]. Artificial intelligence, in general, can provide more comprehensible and enhanced information. [13] The use of black-box models in Heritage Science could bring positive results, but this area requires further research. Several studies [14–18] point to the applicability of the deep learning and especially the Convolutional neural networks (CNN), which have recently been applied for underdrawing recovery, ghost-painting reconstruction [19], x-ray separation [20], or for registration of numerous modalities such as infrared reflectography, visual light photography or x-radiography [21].

This study builds on and improves our previously developed algorithm [22], which exploits neural networks to separate the visible cover contribution from the hidden layers in the near infrared (NIR) reflectograms. In the original method, previously, the authors defined the relationship estimating NIR response based on the visible spectra (VIS) using a shallow neural network as an approximation due to its non-linearity. They kept their model size as tiny as possible to prevent overfitting. This is even more significant concern in our case as we are using higher-dimensional input and larger architectures. The limitation of the previous model is that it works independently with individual pixels. The new model takes into consideration the fact that the pixel's neighbourhood

* Corresponding author.
E-mail address: karella@utia.cas.cz (T. Karella).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.culher.2022.09.022
1296-2074/

areas are strongly dependent on each other. It follows that inclusion of neighbourhood pixels into the input should lead to a more accurate estimate of the NIR-VIS relation. To utilize enlarged input dimension, our models are based on the well-known convolutional neural networks [23] that are specialized for grid data processing and proved to be usefull in various image processing tasks such as object detection [24] or image segmentation [25].

This work's scope is to explore how the inclusion of the pixel's closest surroundings affects the prediction of NIR response. For the first time, we present a novel way of creating the mock-up dataset reflecting pixel neighbours. As in the previous model [22], our research addresses two crucial tasks: the performance measure and the demonstration of the developed method on historical artworks. Phantoms - examples, realized according to historical painting recipes with known pigments and underdrawing coverage, are used first to estimate model accuracy and architecture search. As a further step of our research, we choose only the best performing models on phantoms to reveal hidden layers in artworks to demonstrate the practical value of our model. Our experiments reveal the limitations of our models' hyperparameters, the potential network architectures and the enhanced outputs when applied to historical artworks.

2. Research aim

NIR spectrography is among the most widely used technique for noninvasive examining historical paintings and proved to be usefull in various tasks such as conservation [26] authentication [27] or composition analysis of an art piece [28,29].

However, NIR scanning delivers a mixture of beneath layers with the visible painted layers. A challenging area in the field of NIR spectrography is to separate these two layers and show beneath layers as they look like without those visible contribution. The transition from visible to NIR could still be poorly described, mainly because no exactly defined relation represents that [30].

This paper outlines a new approach for unmixing these two signals utilising the latest methods of Artificial Intelligence dealing with the nonlinearity nature of the problem. We have designed a model based on Convolutional Networks for enriching the hidden layer clarity.

3. Materials and methods

3.1. Mock-ups

M3art database [30] provided us with a set of mock-up samples (physical samples) as shown in Fig. 1. Authors prepared many canvases with color samples from 2012 to 2014. Their database contains 25 materials combined in up to 3 layers (ground layer, underdrawing/underpainting, and top covering layer). The materials were used for creating 634 samples of four kinds - ground layer only, ground layer with drawing, ground layer with color, and finally, groundlayer with drawing covered by color layer.[2]

We use the canvas shown in Fig. 1 for our experiments. Each square 4×4 cm corresponds to twelve different pigments, and its right half contains underdrawing (one set with natural willow charcoal and the second set with red clay (sinopia)). The composition of the pigmented layer was following: 2 g of pigment for 10 drops of 5% solution of animal glue, 5 drops of turpentine, 3 drops of egg yolk and 1 drop of ethanol. The canvas was prepared by mixing 3 volume parts of Bologna chalk (calcium sulfate), 2 volume parts of 7 % aqueous solution of gelatin, 1 egg yolk and 1/4 volume parts of polymerized linseed oil. These samples were used to create our virtual phantoms.

[2] https://m3art.utia.cas.cz/db/overview

Fig. 1. Mock-up from M3art database [30] used for virtual phantoms. First two samples in the first and fourth row are underdrawings only without any pigments (graphite and red clay). Pigments follow (lead white, cinnabar, red lead, madder, indigo, lead-tin yellow, green earth, verdigris, azurite, ultramarine, bone black and umber raw). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Visible and near-infrared imaging

The visible and near-infrared (VIS-NIR) scanning of these samples was performed with a multispectral VIS-NIR scanner constructed by CNR-INO [31]. The scanner has been described in detail here [32]. In brief, it acquires in a single point modality, simultaneously a set of 32 self-registered and aberration-free reflectance images (16 in visible and 16 in near-infrared range). It follows that the available scan contains 32 dimensions representing a wavelength window separating spectral space between 380 nm to 2500 nm, with the resolution of 20–30 nm in VIS and 60–120 nm in NIR.

3.3. Artworks

The two artworks were selected so that the result could be compared our research with our previous work [22]. The first

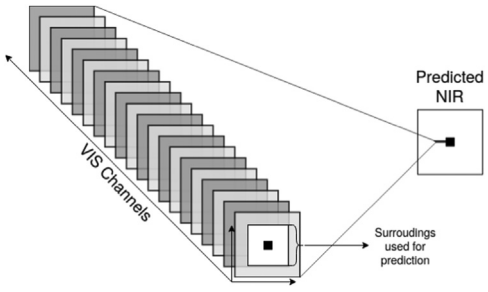

Fig. 2. On the left, there is a 3D cube expressing VIS channels of the image. The right part represents the desired output, which is the expected response of the input in NIR wavelength. The prediction of one pixel (its coordinates represented by a black square) is not only based on the reflectivity of this pixel in the VIS, but also takes into account the reflectivity of neighboring pixels.

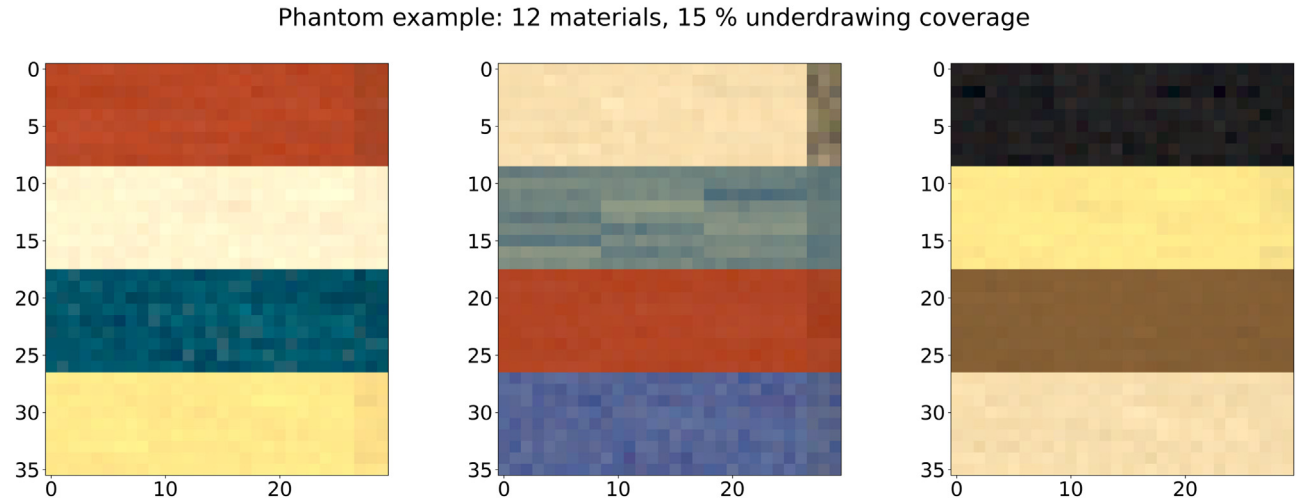

Fig. 3. Phantom example of spatial dimensions (144 × 108) consisting of 12 distinct painting materials with 15% underdrawing coverage. Materials form a grid of (4 × 3), each cell contains single distinct material and underdrawing area is on the right. Shown colors are in RGB derived from 16 VIS bands. The example parameters slightly differs from the experiment dataset for better clarity.

is a 'Still life' oil on canvas by an anonymous painter from the Fine Arts Museum of Asturias. It shows a series of pottery using oil on canvas painting (23.4 x 28.4 cm) and originated in the early 20th century. Infrared reflectography discovered underdrawing and underpaintings scenes with buildings not visible by the naked eye.

The second example is 'Madonna dei Fusi' (Madonna of the Yarnwinder) attributed to Leonardo da Vinci. The painting is owned by a private collector. The canvas is also oil painted and dates back to 1501–1507; the analysis revealed various *pentimenti*.

VIS-NIR multispectral scanner detailed in [32] and [33] yields 32 channel scans (380, 2800 nm) of these artworks. The scanners spatially registered the images, and we then used the resulting hypercube.

3.4. Model

Our goal is to predict NIR response (in. specific spectral window) given the vector of VIS reflectancies, and this is motivated by the fact that some NIR content is not involved in VIS spectra. The estimation is further intended to distinguish between concealed features noticeable only in the infrared domain by removing the expected VIS contribution in NIR. Therefore, the result should display just invisible parts of NIR, remove residuals of VIS contribution, and thus present a more accurate estimation of the painting hidden features.

We estimate the transition function from VIS to NIR by neural nets to address its non-linearity. Particular painting has its own and different estimated function because the material property is strongly dependent on many factors. The input consists of 16 VIS bands, and the output is a single NIR band; including more NIR bands as output does not lead to better results, that is inspired by our predecessor [22].

To extend the capabilities of preceding architecture, we consider the pixel spatial surroundings for this estimation illustrated in Fig. 2. To make use of the newly included information, our model incorporates CNN, which are traditionally selected for dealing with spatial space [34].

Instead of matrix multiplication employed by classical neural nets, CNN processes the input using a convolutional operation with the trainable kernel. Let us define the convolution operation $\circledast$ as:

Definition 1 (Discrete Convolution)**.**

$$S(i,j) = (K \circledast I)(i,j) = \sum_{m}^{M}\sum_{n}^{N} I(i-m, j-n)K(m,n), \tag{1}$$

where I stands for pixels intensities (dimension of I is M - width and N - height). Convolution filter K (kernel) is also 2D matrix with dimensions typically lower than I. Convolution is moving the filter K to every possible spatial coordinate (i, j) of image I, and for each (i, j) multiplying image intensities by corresponding weights of kernel K and summing weighted intensities.[3]

The convolutional layers also include adding bias and activation function after convolution operation like typical neurons.

The Information Gain (IG) metric given in [22] seems to be a reliable model performance metric in terms of separation in the NIR spectra.

Input data are represented as functions VIS and NIR_λ non-zero only on spatial grid of dimension $(m \times n)$. The function $\hat{f}$ (our CNN) predicts the NIR response given only VIS bands illustrated as "Predicted NIR" in Fig. 4.

Definition 2 (VIS)**.** Function $\mathrm{VIS}(i,j) : \mathbb{N}^2 \to \mathbb{R}^{l\times l\times b}$; gives reflectancy in b consequent bands of visible spectra for pixel at (i, j) including $(l \times l)$ neighbourhood.

For our datasets $b = 16$, $l \in \{3, 5, 7, 9\}$.

Definition 3 (NIR)**.** Function $NIR_\lambda(i,j) : \mathbb{N}^2 \to \mathbb{R}$; returns pixel reflectence in single NIR band at spatial coordinates (i, j). The NIR band is centered at wavelength λ.

In our case λ is different with respect to the painting: 1730 nm for *Still life* and 950 nm for *Madonna dei Fusi*.

Performance of $\hat{f}$ is measured as a difference between real NIR response and predicted. "The Information Gain for a certain spectral window and a given pixel is the change of captured radiation caused by painting layers not included in VIS." [22]

[3] Note that most CNNs use the flipped kernel $(K \circledast I)(i,j) = \sum_m^M \sum_n^N I(i+m, j+n)K(m,n)$, therefore from the mathematical point of view the $\circledast$ operator should be termed as the cross-correlation.

Definition 4 (IG)**.** The IG metric is defined as:

$$\Delta = \sum_{i}^{M}\sum_{j}^{N} \hat{f}(VIS(i,j)) - NIR_{\lambda}(i,j), \tag{2}$$

where VIS function and NIR_{λ} are defined beforehand, $\hat{f}$ is model $\mathbb{R}^{l\times l\times b} \to \mathbb{R}$. Formula sums over spatial dimension of input data $(M \times N)$.

However, IG is not suitable for training neural networks because it requires a smooth and differentiable loss function; we used the function from our previous article, where the network weights were adjusted based on a loss function defined as the Mean Square Error (MSE), which reflects IG and is easily differentiable. Thus, IG is used to measure validation error and MSE for the training.

Definition 5 (MSE)**.** The MSE metric is defined as:

$$\mathrm{MSE}(\hat{f}) = \frac{1}{M\times N}\sum_{i}^{M}\sum_{j}^{N}(\hat{f}(VIS(i,j)) - NIR_{\lambda}(i,j))^2, \tag{3}$$

where VIS function and NIR_{λ} are defined beforehand, $\hat{f}$ is model $\mathbb{R}^{l\times l\times b} \to \mathbb{R}$. Formula sums over spatial dimension of input data $(M \times N)$.

The output of all models was the same single neuron corresponding to the pixel intensity distribution of greyscaled image. The input size limits the kernel size and the depth because only a "valid" padding of the CNN layer was allowed, which means that not all the parameter options are available for each input size.

Table 1
CNN architecture parameters.

Parameter name	Values
Filters per CNN layer	10, 15, 25, 50
Kernel size	3, 5, 9
Number of CNN layers (Depth)	1, 2, 3, 4
Activation functions	sigmoid, relu

We choose the number of learnable paremeters close to the foregoing model capacity to overcome the overfitting phenomenon. Hyperparameter behaviour is described in detail in Appendix C. Even for our smaller convolutional layers made of consecutive 3×3 kernels with 30 filters, the model tends to memorize the patterns, and the validation error drops after approximately 20 epochs. We employ the early stopping regularization technique to avoid this undesired behaviour and we work with higher capacity models. Still, we do not enlarge the filters above 50 because such models' training was notably non-stable and requires sensitive hyperparameter selection. The following Table 1 summarizes the tested parameters related to these restrictions.

3.5. Virtual phantoms generators

In order to evaluate the model performance, the hidden features of the NIR have to be known; therefore, we employ the

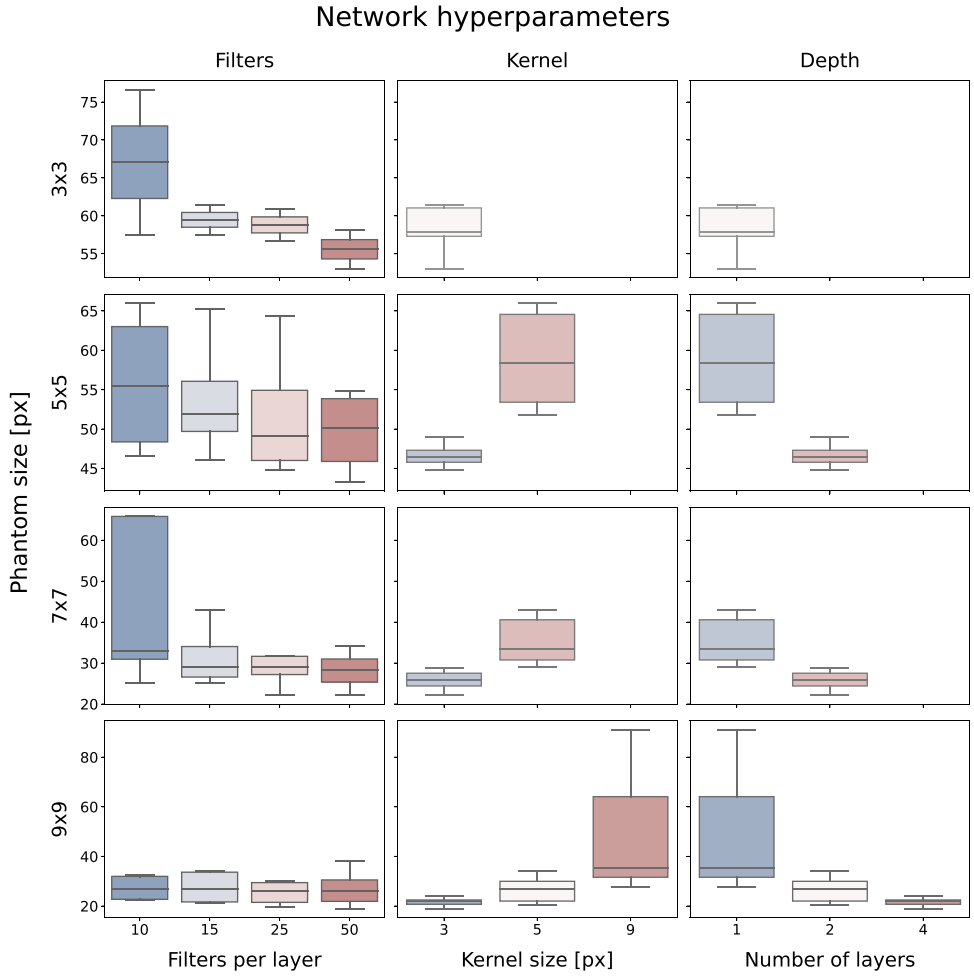

Fig. 4. The figure displays the MSE performance according to the network parameters and the phantom input size. Rows represent a different size of the phantom and columns the architecture parameter. Boxplots reflect the influence on the MSE expressed in the y-axis of the particular parameter value on the x-axis. (the MSE is multiplied by 10^5 to improve readability.)

Table 2
The table enlists the results of our experiments compared to the baseline model (the predecessor model [22]). The MSE metric refers to an average performance on the 20 distinct virtual phantoms. The input column stands for training sample dimensions, the depth for a number of hidden layers, the filters for a number of convolutional filters or neurons in the baseline case, the kernel for the size of convolutional filters, the activation for a chosen activation function.

Input [px]	Depth	Filters	Kernel [px]	Activation	Validation MSE
baseline	2	25	-	sigmoid	0.000636 ± 0.000139
3x3	1	50	3	relu	0.000530 ± 0.000104
3x3	1	25	3	sigmoid	0.000567 ± 0.000106
3x3	1	10	3	sigmoid	0.000575 ± 0.000127
3x3	1	15	3	sigmoid	0.000575 ± 0.000085
5x5	2	50	3	sigmoid	0.000433 ± 0.000070
5x5	2	25	3	sigmoid	0.000447 ± 0.000064
5x5	2	15	3	sigmoid	0.000461 ± 0.000073
5x5	2	25	3	relu	0.000464 ± 0.000056
7x7	3	25	3	sigmoid	0.000221 ± 0.000053
7x7	3	50	3	sigmoid	0.000222 ± 0.000064
7x7	3	15	3	sigmoid	0.000251 ± 0.000080
7x7	3	10	3	relu	0.000252 ± 0.000055
9x9	4	50	3	sigmoid	0.000189 ± 0.000045
9x9	4	25	3	sigmoid	0.000196 ± 0.000046
9x9	2	25	5	sigmoid	0.000206 ± 0.000043
9x9	4	15	3	sigmoid	0.000211 ± 0.000063

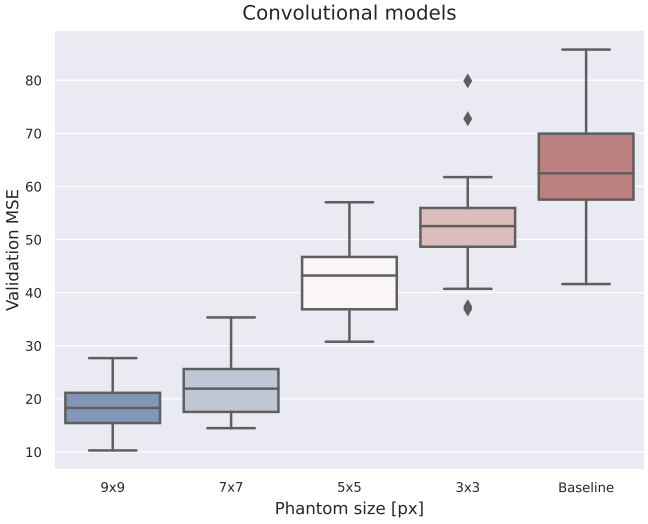

Fig. 5. The plot shows the performance of our best models using pixel surroundings for predicting the NIR response. The x-axis represents the size of the employed neighbourhood and the y-axis the MSE measured on the validation set. The predecessor (baseline) model [22] not using any neighbourhood pixels is plotted on the very right. It was trained and validated on the same virtual phantoms only with omitted surroundings of the pixel. (The MSE is multiplied by 10^5 to improve the readability.)

physical samples (mock-ups) with the precise position of underdrawings. Based on these, we can generate artificially datasets of a changeable size, a material count or an underdrawing ratio, which we call virtual phantoms.

Mock-ups as a source of truth consist of pigments without overdrawing, pigments over graphite underdrawings, and pigments over clay underdrawings. This gives us 12×3 groups with different reflectance characteristics. From each group we take patches (400) which have required dimensions and contains all 16 VIS reflectivity values per pixel and single NIR band that is same for all patches each group. The dimensions of a sample are then $l \times l \times 17$, where l vary according to the CNN architecture. Let us denote sample $\boldsymbol{x_i}$.

At this point we needed to be able to generate artificial samples to increase their number so we derive a sample generator for each group represented by group mean vector and variance for each dimension. According to our predecessor [22] we assume that material reflectance matches the normal distribution. Hence, generator is represented by a multivariate Gaussian distribution of the k-dimensional random vector.

Let us assume $D_{m,l}$ as the m material distribution with spatial size $l \times l$. We approximate $D_{m,l}$ by a multivariate Gaussian distribution $I_{m,l}(\mu, \Sigma)$. The $I_{m,l}$ parameters are estimated by using all the $\boldsymbol{x_i}$ samples of material m and l size. We calculate μ as the sample mean vector and Σ as the sample covariance matrix $\mathbf{Q}$ given by the following formulas:

$$\boldsymbol{\mu}(m) \sim \frac{1}{S}\sum_{i=1}^{S}\mathbf{x_i} = \begin{bmatrix}\bar{x}_1\\ \vdots \\ \bar{x}_K\end{bmatrix} \; with \; \bar{x}_J = \frac{1}{S}\sum_{i=1}^{S} x_{ij} \tag{4}$$

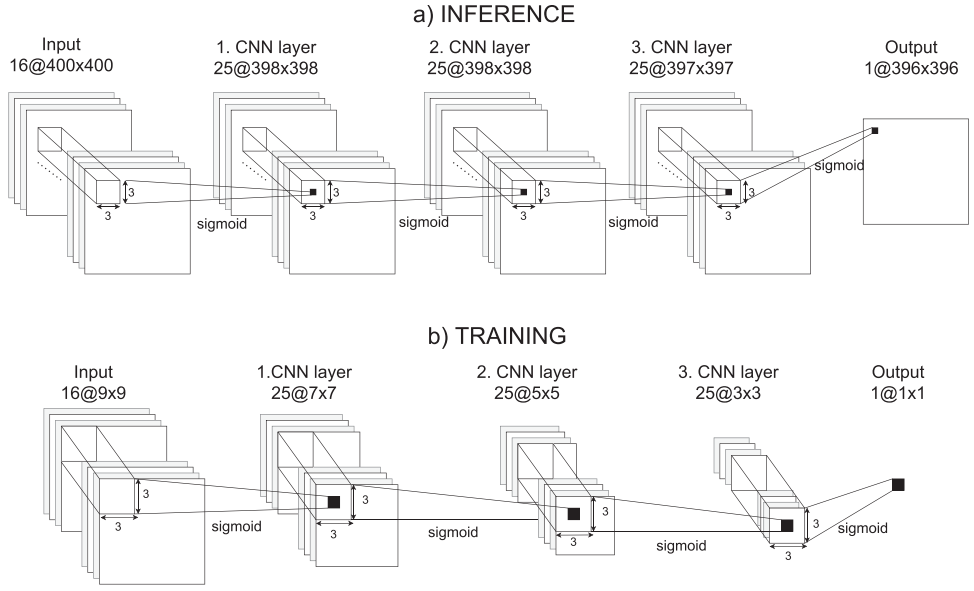

Fig. 6. a) Inference; the trained network is used for estimation of the NIR response of VIS. **b)** Training of network; parameters are trained on a subset of samples of size 9×9 of the original painting. (The filter notation is as follows depth@width x height.)

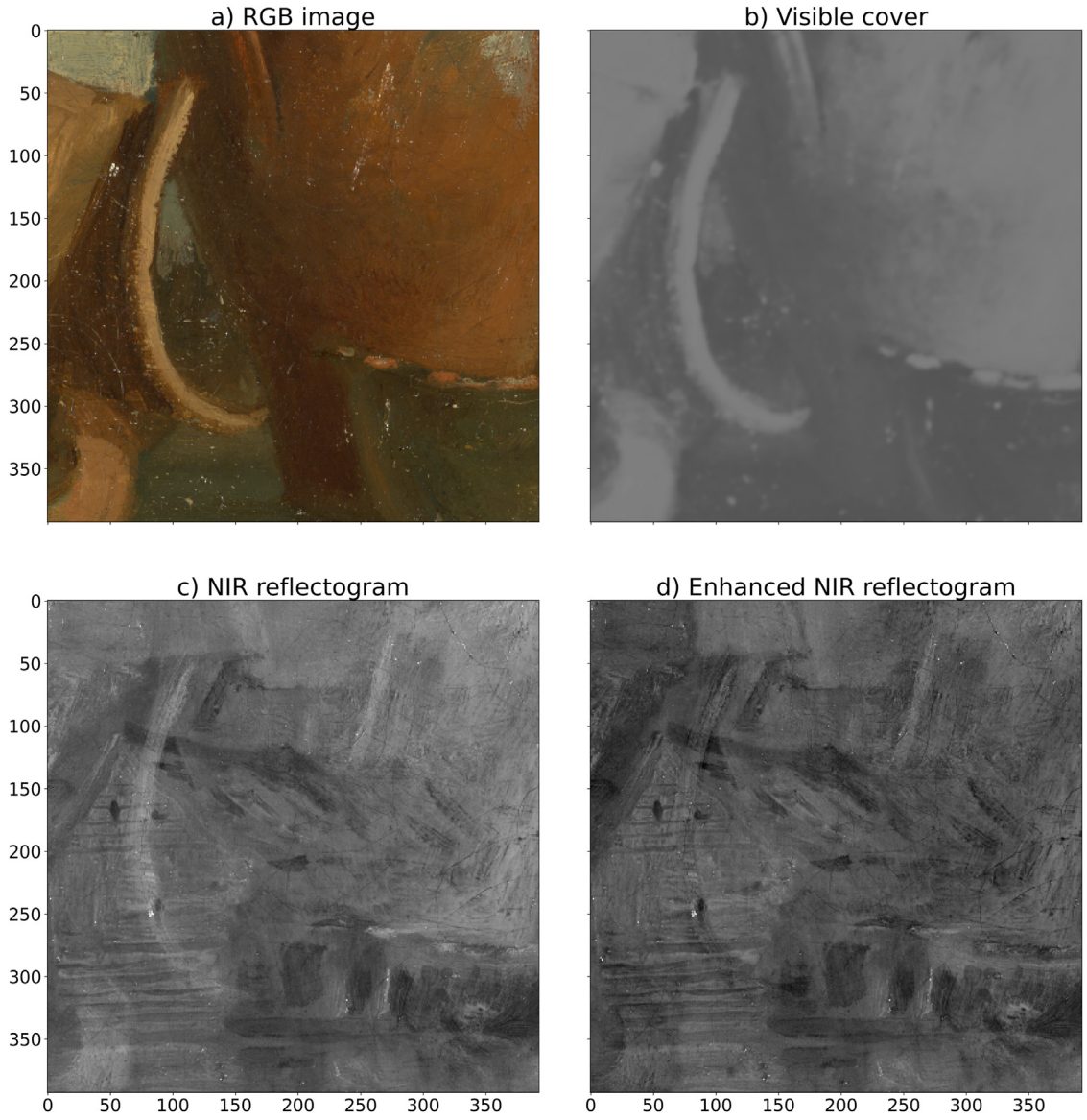

Fig. 7. The detail (10 x 10 cm) of the *Still life* painting: **a)** the RGB image, **b)** the visible cover (output of the neural network), **c)** the NIR reflectogram centered at 1730 nm, **d)** the enhanced reflectogram (the subtraction image c-b).

$$\boldsymbol{\Sigma}(m) \sim \mathbf{Q} = [q_{jk}] \; with \; q_{jk} = \frac{1}{S-1}\sum_{i=1}^{S}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \tag{5}$$

where $\mathbf{x_i}$ is a sample of $\mathcal{D}_{m,l}$ with dimension $K = (l \times l \times s)$ and S stands for the total samples count.

For drawing vector $\boldsymbol{v}$ from $I_{m,l}$, we need decompose covariance matrix (for example by Cholesky decomposition),

$$\boldsymbol{\Sigma}(m) = \boldsymbol{L}\boldsymbol{L}^T, \tag{6}$$

where Σ is covariance matrix and $\boldsymbol{L}$ is a real lower triangular matrix with positive diagonal entries. Vector $\boldsymbol{u} = (u_1, \ldots, u_n)$, where u_i is independent sample from a standard normal distribution $N(0, 1)$. Then the vector $\boldsymbol{v}$ is calculated by following equation:

$$\boldsymbol{v} = \boldsymbol{\mu}(m) + \boldsymbol{L} \cdot u \tag{7}$$

4. Results

The result section includes two sections. Firstly, we showcase the experiments with virtual phantoms primarily to examine various CNN architectures and to demonstrate that our improved model can achieve better accuracy than the antecedent model [22]. The experiments are evaluated by the MSE metric to give an objective and straightforward score. Secondly, we apply the best models from the first phase on real artworks to illustrate its contribution to better visualization of hidden features in NIR reflectograms. We present figures confirming the reduction of VIS residuals and improved visibility of thin drawings.

4.1. CNN architecture design

To study the model design, we prepared four sets of virtual phantoms with a variable size of neighbourhood surroundings (from 1 to 4) generated from distributions $I_{m,l}$. Each set contains 20 different phantoms; each phantom contains samples of spatial size 400×400 equally distributed among 12 distinct materials m (sampled from $I_{m,l}$); each of them includes 2% samples with underdrawing and 17 different bands (16 VIS bands and single NIR band). Phantom example with 12 materials and 15% underdrawing coverage is shown in Fig. 3. These settings reflect the predeces-

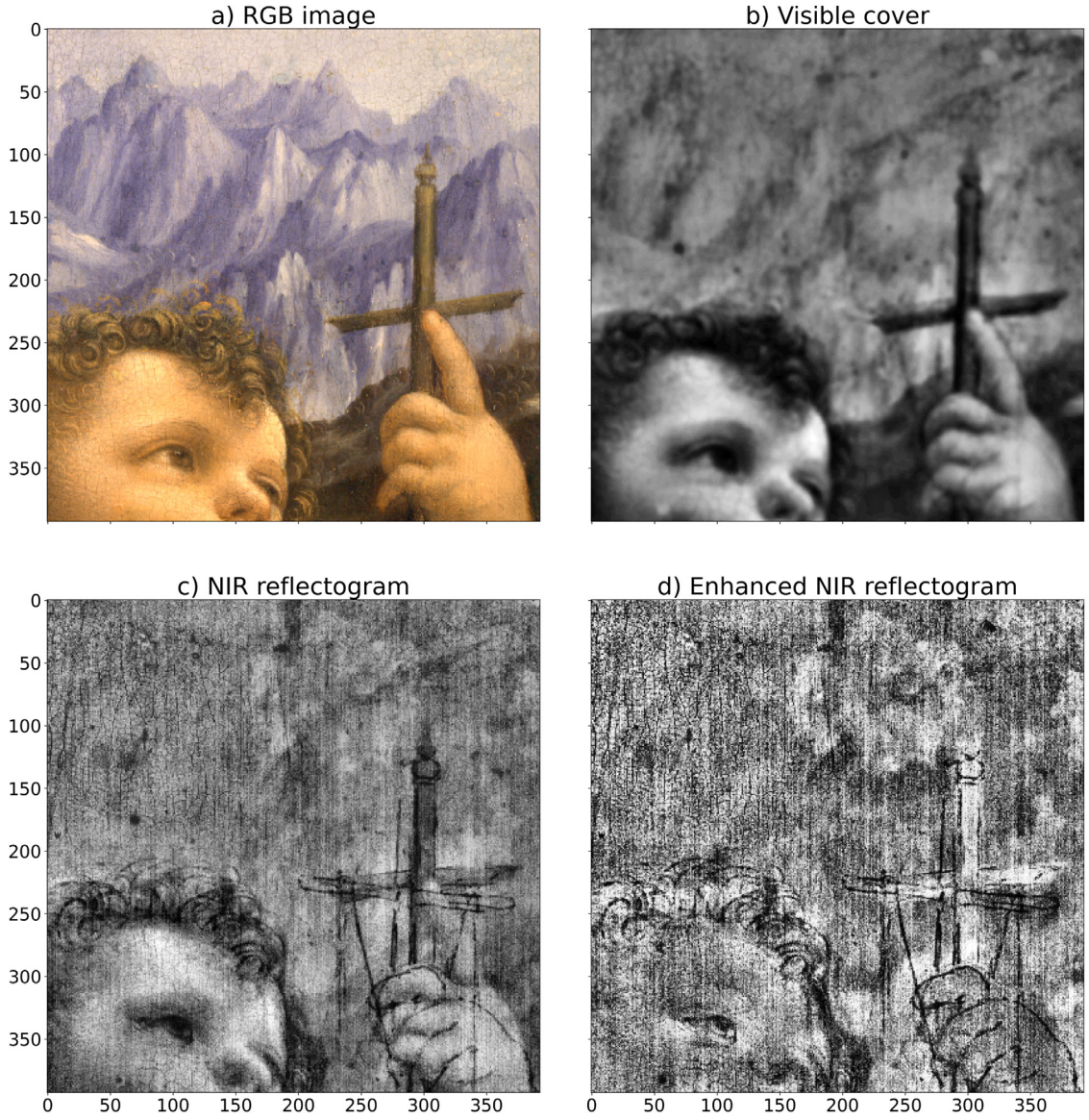

Fig. 8. The detail (10 x 10 cm) of *Madonna dei Fusi* painting attributed to Leonardo Da Vinci: **a)** the RGB image, **b)** the visible cover (the output of the neural network), **c)** the NIR reflectogram centered at 950 nm, **d)** the enhanced reflectogram (the subtraction image c-b). Histograms of images b, c, d were equalized by CLAHE algorithm [35] (the original NIR is included in Appendix B.1).

sor performance measuring environment [22]. These virtual phantoms were split into validation and training sets and served as a dataset to train various CNN architectures and to measure their performance.

We have tested several hyperparameters such as the number of filters per CNN layer, the kernel size, and the number of convolutional layers. We report the MSE performance as a function of these hyperparameters for a given phantom input size in the Fig. 4. It can be seen that 25 filters are sufficient (a higher number has little impact on MSE), the kernel size 3×3 is adequate, and deeper networks achieve lower standard deviation and lower error rates, especially for the 7×7 and 9×9 input size.

The results, as shown in Table 2 with the detailed description of the most successful models, confirms that our CNN architectures outperform the baseline model [22] in terms of the MSE metric. In addition to the findings about hyperparameters from Fig. 4, Table 2 points to the sigmoid activation function as more appropriate. The best networks of these settings are further applied to artworks to demonstrate their practical usability.

As Fig. 5 shows, there is a clear trend of the positive contribution of a pixel surroundings. All of our best models achieved lower MSE than the baseline model, and their error decreased with the larger surroundings along with the standard deviation until the 9×9 dimension of the input.

4.2. CNN to enhance information about NIR reflectograms in historical paintings

As a further step of our research, we have tested the developed method on artworks. We selected regions of interest on the paintings and for each of them, we trained a model on 2500 randomly chosen pixels with 9×9 surroundings. We have selected architecture with four CNN layers, each containing 25 filters, with 3×3 kernel size and sigmoid activation summarized in Fig. 6.

The trained model estimates the VIS contribution (the visible cover) in NIR reflectograms. For the improved reflectogram, the visible cover is subtracted from the acquired NIR reflectogram. We describe here two representative examples of enhanced visibility

of hidden features in the NIR reflectograms of an anonymous oil painting from 20th century shown in Fig. 7 and a panel painting attributed to Leonardo da Vinci in Fig. 8.

5. Conclusion

The performed tests validate our assumption that including pixel spatial surroundings leads to more accurate results in the NIR estimation for the virtual phantoms and the real paintings. As shown in Fig. 5, our models overcame the MSE metric of the previous model, and the actual experiments with artworks confirm the achievement of the practical result as shown in Figs. 7 and 8.

The architecture search confirms the common knowledge that the kernels 3×3 are sufficient and enlarging the kernel decreases the model stability and the performance as seen in the middle column in Fig. 4. The last column of Fig. 4 points out to the fact that the bigger network depth (more CNN layers) the better performing model. Increasing the number of filters per CNN layer improves the model accuracy, the plateau is reached with 50 filters. However, the model capacity should not be larger, the stack of too many CNN layers suffers from a vanishing gradient, and models with too many filters tend to overfit.

When it comes to historical artworks, our models cope better with colour transitions such as a wide brush stroke in the middle of the artwork shown in Fig. 7 and reducing the impact of high-brightness points from VIS. Although the output corrected several misregistration errors, there are still some imperfections, such as the brush stroke in the lower corner in Fig. 7. As shown in Fig. 8 our networks reduced noises and enhanced numerous cracks of the paintings, allowing us to better distinguish each angel's hair better and change the background to a firmer one.

This study is a first step towards a better understanding of how the spatial environment of pixels affects the neural network modelling of the VIS to NIR transition. In the following research, we will focus on a more complex modelling of virtual phantoms, employing the multilayered paint stratigraphy (overpaints). Our future work will also investigate the benefits of more advanced CNN models such as the residual connection introduced in [36] along with the extensive regularization techniques to handle the vanishing gradient problem and prevent overfitting.

Statements and Declarations

- Funding: This work was supported by GA21-03921S and Strategy AV21 of the Czech Academy of Sciences “Hopes and Risks of the Digital Era”.
- Ethics approval: Not applicable
- Consent to participate: Not applicable
- Consent for publication: Not applicable
- Availability of data and materials: The source codes, trained deep learning model, virtual and physical phantoms data of this study are available from the corresponding author upon reasonable request. Data containing artworks such as Madonna dei Fusi and Still life are not publicly available; for further information, contact Jana Striov.
- Authors' contributions: All authors contributed to the study conception and design. Material preparation and analysis were performed by Tom Karella. The first draft of the manuscript was written by Tom Karella and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We are grateful to the National Institute of Optics (INO-CNR) for sharing this rare dataset for the presented analysis. This research was supported by Strategy AV21 of the Czech Academy of Sciences ”Hopes and Risks of the Digital Era” and GACR GA21-03921S.

Appendix A. Abbreviations

NIR near infrared
VIS visible spectra
IG Information Gain
CNN Convolutional neural networks
MSE Mean Squared Error

Appendix B. Artworks

The original non improved version of the Fig. 8 is displayed in Fig. B.1, we used the histogram equalization algorithms to improve readability for the reader, but no other changes have been made.

To manifest the distinctions between our and the baseline [22] model, we additionally incorporate Figs. B.3, B.2 with the highlighted difference between model outputs.

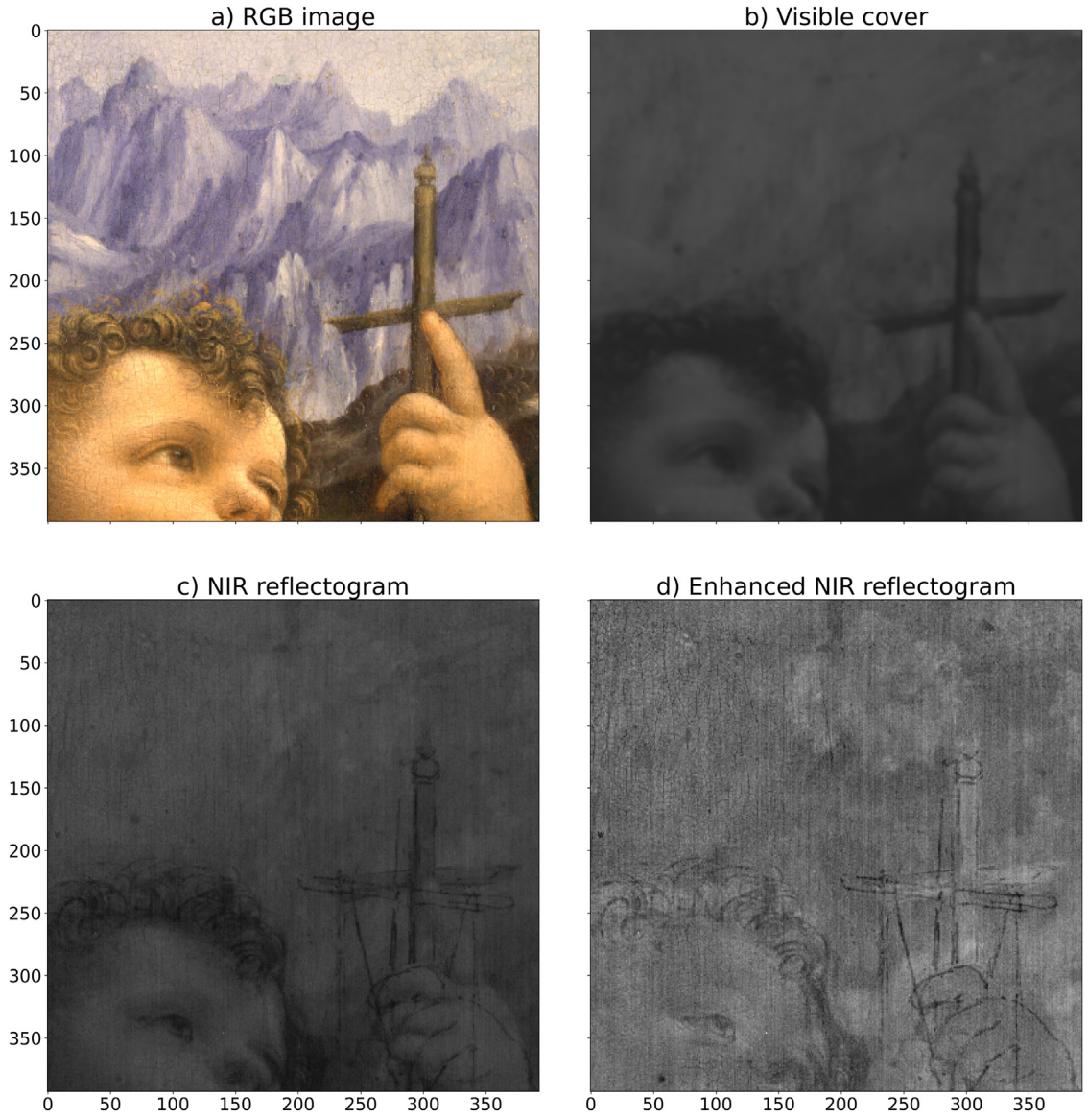

Fig. B.1. The detail (10 x 10 cm) of *Madonna dei Fusi* painting attributed to Leonardo Da Vinci: **a)** the RGB image, **b)** the visible cover (the output of the neural network), **c)** the original NIR reflectogram centered at 950 nm, **d)** the enhanced reflectogram (the subtraction image c-b).

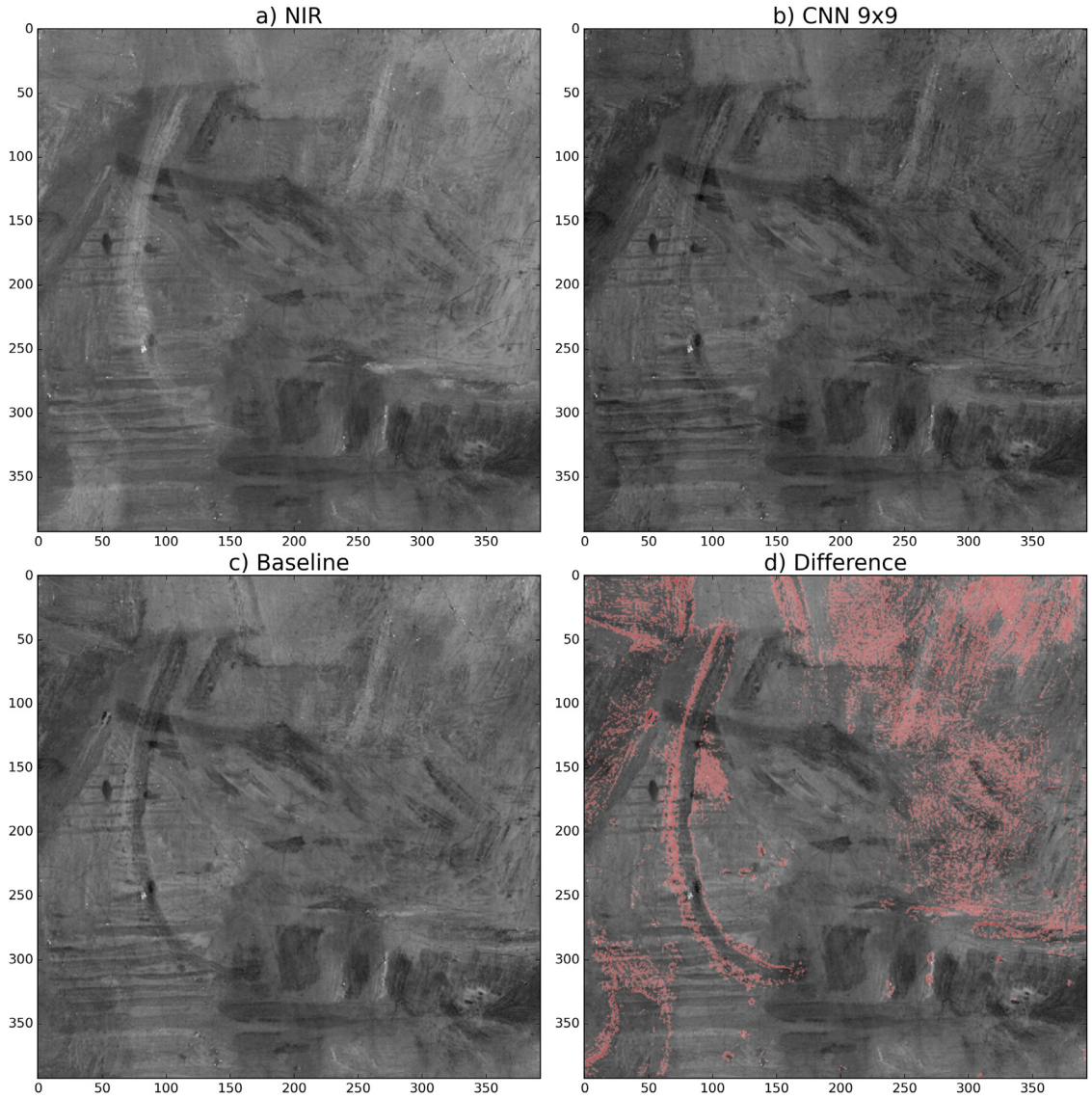

Fig. B.2. the detail (10 x 10 cm) of *Still life* painting: **a)** the NIR reflectogram centered at 1730 nm, **b)** the enhanced reflectogram by our baseline model (the substraction image a - output of the neural network), **c)** the enhanced reflectogram by our model using 9×9 surroundings (the substraction image a - the output of the neural network), **d)** the diference between b and c, the red color highlights diferences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

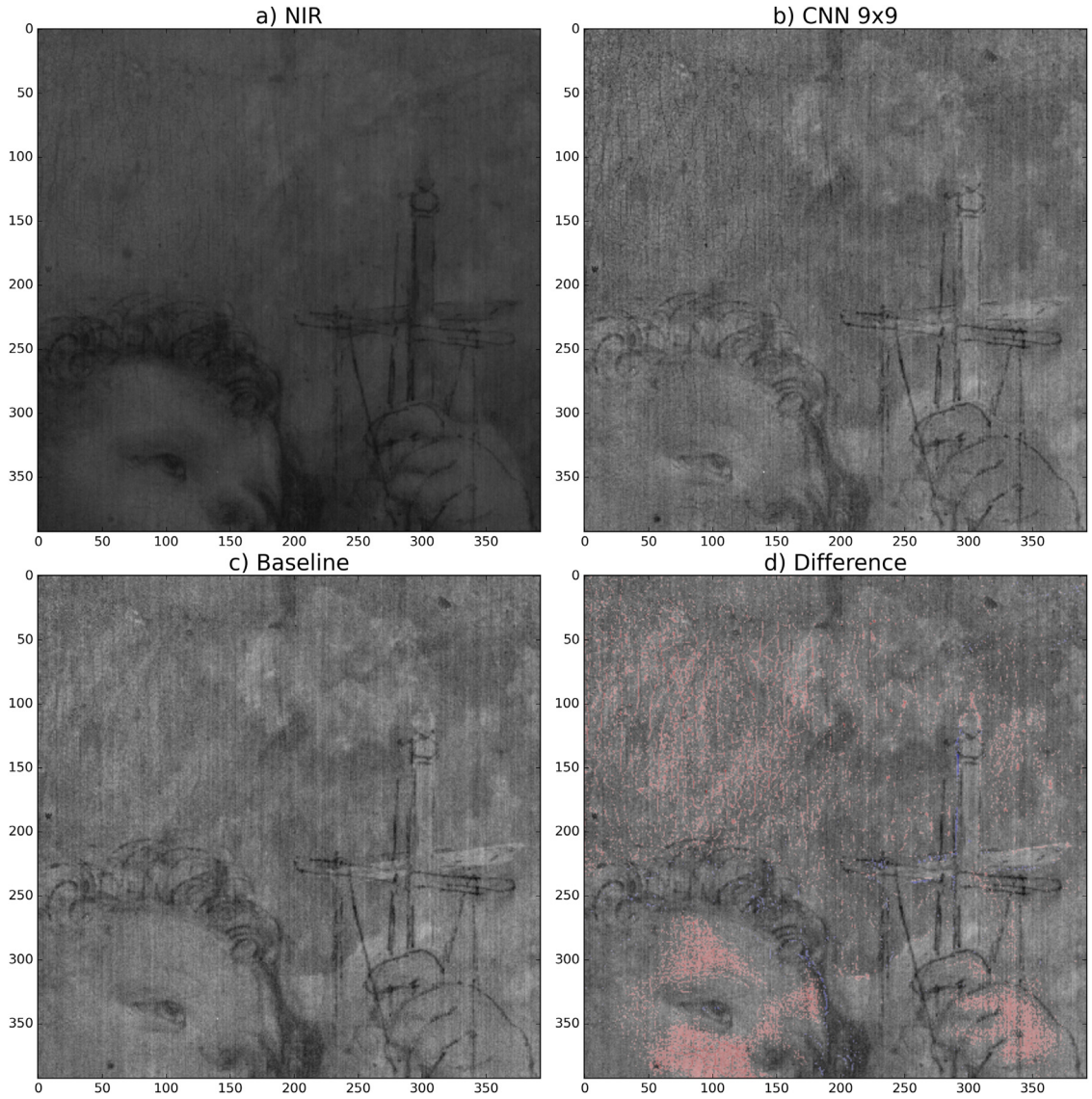

Fig. B.3. The detail (10 x 10 cm) of *Madonna dei Fusi* painting attributed to Leonardo Da Vinci: **a)** the NIR reflectogram centered at 950 nm, **b)** the enhanced reflectogram by our baseline model (the substraction image a - output of the neural network), **c)**the enhanced reflectogram by our model using $9 \times 9$the surroundings (the substraction image a - the output of the neural network), **d)** the difference between b and c, the red and blue color highlights differences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

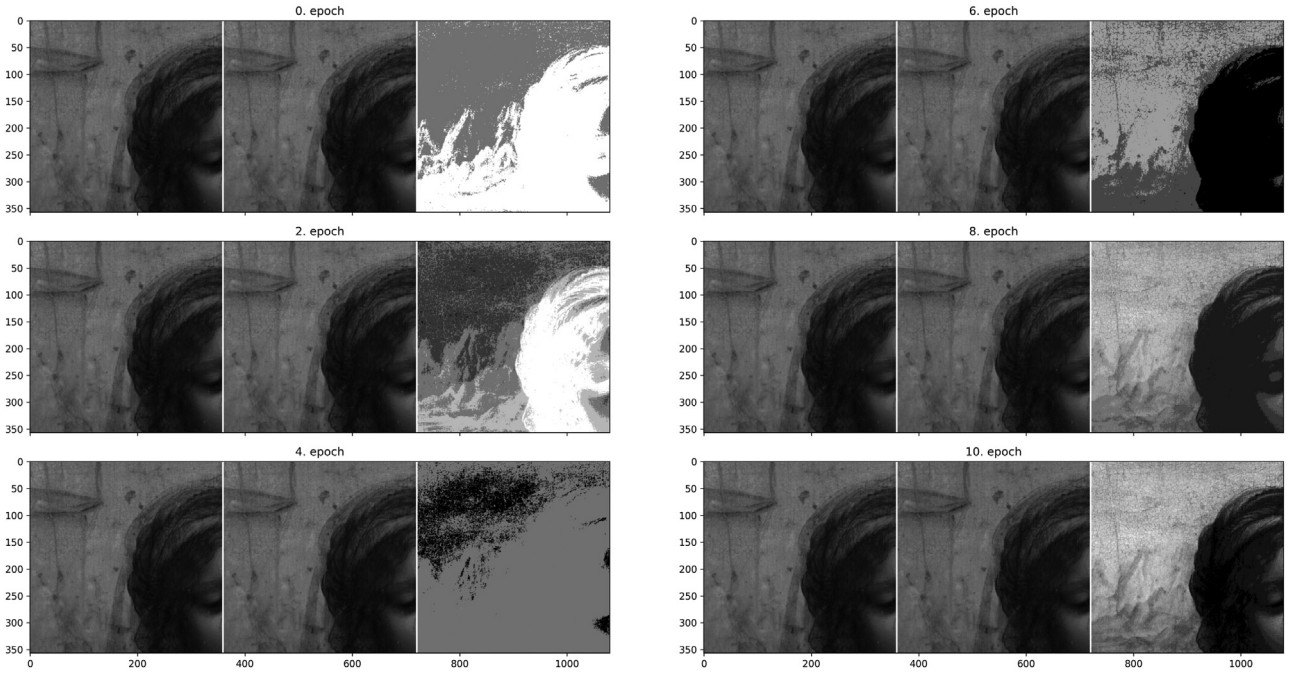

Fig. C.1. The progress of CNN output over the epochs: each cell contains grayscaled VIS reflectogram, NIR reflectogram and the visible cover (the output of the neural network).

Appendix C. CNN filters and epochs

In this section, we sum up some of CNN hyperparameter's behavior we were adjusting during our experiments.

CNN contains blocks of convolutional layers, that is, convolution operation with n learnable kernels of given size $(l \times l)$; we usually refer n as a number of filters and l as a kernel size. Using a convolutional layer with n filters on an image gives n output feature maps (all kernels applied on the image). The more filters, the more complex function is approximated by CNN (similar to modeling by higher order polynomial), but it could lead to overfitting phenomena.[4] The goal is to choose a suitable number of filters and CNN layers based on the validation error. For further details, see chapter "Convolutional Networks" in book [34].

Learning of neural networks refers to using stochastic gradient descent optimization. We usually use our training data in small patches (called batches), calculate the gradient for each batch, and adjust the learnable parameters of the neural network. As long as the gradient descent is an iterative method, we have to repeat that several times; in single epoch means that we adjusted for every batch. We illustrate the change of our CNN during the first ten epochs in Fig. C.1. For further details, see chapter "Optimization for Training Deep Models" in book [34].

References

[1] J. Striova, A. Dal Fovo, R. Fontana, Reflectance imaging spectroscopy in heritage science, La Rivista del Nuovo Cimento 43 (10) (2020) 515–566, doi:10.1007/S40766-020-00011-6.
[2] C. Ludovica, K. Dandolo, P.U. Jepsen, Wall painting investigation by means of non-invasive terahertz time-domain imaging (THz-TDI): inspection of subsurface structures buried in historical plasters, J. Infrared Millimeter Terahertz Waves 37 (2) (2016) 198–208, doi:10.1007/s10762-015-0218-9.
[3] G. Filippidis, M. Massaouti, A. Selimis, E.J. Gualda, J.-M. Manceau, S. Tzortzakis, Nonlinear imaging and THz diagnostic tools in the service of cultural heritage, Appl. Phys. A 106 (2) (2011) 257–263, doi:10.1007/S00339-011-6691-7.
[4] A. Redo-Sanchez, B. Heshmat, A. Aghasi, S. Naqvi, M. Zhang, J. Romberg, R. Raskar, Terahertz time-gated spectral imaging for content extraction through layered structures, Nat. Commun. 7 (1) (2016) 1–7, doi:10.1038/ncomms12665.
[5] P. Targowski, M. Iwanicka, Optical coherence tomography: its role in the non-invasive structural examination and conservation of cultural heritage objects a review, Appl. Phys. 106 (2) (2011) 265–277, doi:10.1007/S00339-011-6687-3.
[6] K. Kim, P. Kim, J. Lee, S. Kim, S. Park, S.H. Choi, J. Hwang, J.H. Lee, H. Lee, R.E. Wijesinghe, M. Jeon, J. Kim, Non-destructive identification of weld-boundary and porosity formation during laser transmission welding by using optical coherence tomography, IEEE Access 6 (2018) 76768–76775, doi:10.1109/ACCESS.2018.2882527.
[7] G.J. Tserevelakis, A. Chaban, E. Klironomou, K. Melessanaki, J. Striova, G. Zacharakis, Revealing hidden features in multilayered artworks by means of an epi-illumination photoacoustic imaging system, J. Imaging 7 (9) (2021) 183, doi:10.3390/JIMAGING7090183.
[8] G.J. Tserevelakis, P. Siozos, A. Papanikolaou, K. Melessanaki, G. Zacharakis, Non-invasive photoacoustic detection of hidden underdrawings in paintings using air-coupled transducers, Ultrasonics 98 (2019) 94–98, doi:10.1016/J.ULTRAS.2019.06.008.
[9] A.D. Fovo, A. Papanikolaou, G.J. Tserevelakis, G. Zacharakis, R. Fontana, Combined photoacoustic imaging to delineate the internal structure of paintings, Opt. Lett. 44 (4) (2019) 919–922, doi:10.1364/OL.44.000919.
[10] A. Dal Fovo, M. Castillejo, R. Fontana, Nonlinear optical microscopy for artworks physics, La Rivista del Nuovo Cimento 2021 44:9 44 (9) (2021) 453–498, doi:10.1007/S40766-021-00023-W.
[11] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, doi:10.1016/j.media.2017.07.005.
[12] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: ICML, 2011, pp. 689–696. https://icml.cc/2011/papers/399_icmlpaper.pdf. Bellevue
[13] M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue, S. James, Machine learning for cultural heritage: a survey, Pattern Recognit. Lett. 133 (2020) 102–108, doi:10.1016/J.PATREC.2020.02.017.
[14] M. Sabatelli, M. Kestemont, W. Daelemans, P. Geurts, Deep transfer learning for art classification problems, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018.
[15] B. Saleh, A. Elgammal, Large-scale classification of fine-art paintings: learning the right metric on the right feature, 2015, arXiv preprint arXiv:1505.00855.
[16] A. Elgammal, Y. Kang, M.D. Leeuw, Picasso, matisse, or a fake? Automated analysis of drawings at the stroke level for attribution and authentication, 2018, 42–5032nd AAAI Conference on Artificial Intelligence, AAAI 2018
[17] W.R. Tan, C.S. Chan, H.E. Aguirre, K. Tanaka, Ceci n'est pas une pipe: a deep convolutional network for fine-art paintings classification, Proceedings - International Conference on Image Processing, ICIP vol. 2016-August (2016) 3703–3707, doi:10.1109/ICIP.2016.7533051.
[18] M. Ghosh, S.M. Obaidullah, F. Gherardini, M. Zdimalova, Classification of geometric forms in mosaics using deep neural network, 2021, J. Imaging, 7(8), 10.3390/jimaging7080149

[4] https://en.wikipedia.org/wiki/Overfitting

[19] A. Bourached, G. Cann, R.-R. Griffiths, D.G. Stork, Recovery of underdrawings and ghost-paintings via style transfer by deep convolutional neural networks: a digital tool for art scholars, IS&T Int. Symp. Electron.Imaging Sci. Technol. 2021 (14) (2021).
[20] Z. Sabetsarvestani, B. Sober, C. Higgitt, I. Daubechies, M.R.D. Rodrigues, Artificial intelligence for art investigation: meeting the challenge of separating x-ray images of the ghent altarpiece, Sci. Adv. 5 (8) (2019). 10.1126/SCIADV.AAW7416/ASSET/3CA4CA13-8855-45ED-BD07-C656DDF1857A/ASSETS/GRAPHIC/AAW7416-F6.JPEG
[21] A. Sindel, A. Maier, V. Christlein, Craquelurenet: Matching the crack structure in historical paintings for multi-modal image registration, in: 2021 IEEE International Conference on Image Processing (ICIP), 2021, pp. 994–998, doi:10.1109/ICIP42928.2021.9506071.
[22] J. Blažek, J. Striová, R. Fontana, B. Zitová, Improvement of the visibility of concealed features in artwork NIR reflectograms by information separation, Digit. Signal Process. Rev. J. 60 (2017) 140–151, doi:10.1016/j.dsp.2016.09.007.
[23] Y. LeCun, Generalization and network design strategies, Connectionism Perspect. 19 (1989) 143–155.
[24] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen, Deep learning for generic object detection: asurvey, Int. J. Comput. Vis. 128 (2) (2020) 261–318, doi:10.1007/s11263-019-01247-4.
[25] S. Minaee, Y.Y. Boykov, F. Porikli, A.J. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: a survey, IEEE Trans. Pattern Anal. Mach. Intell. (2021), doi:10.1109/TPAMI.2021.3059968.
[26] M. Attas, E. Cloutis, C. Collins, D. Goltz, C. Majzels, J.R. Mansfield, H.H. Mantsch, Near-infrared spectroscopic imaging in art conservation: investigation of drawing constituents, J. Cult. Herit. 4 (2) (2003) 127–136, doi:10.1016/S1296-2074(03)00024-4.
[27] M.R. Derrick, D. Stulik, J.M. Landry, Infrared Spectroscopy in Conservation Science, The Getty Conservation Institute, Los Angeles, 2000.
[28] A. Casini, F. Lotti, M. Picollo, L. Stefani, E. Buzzegoli, Image spectroscopy mapping technique for noninvasive analysis of paintings, 1999, 10.1179/sic.1999.44.1.39, 44(1), 39–48,
[29] A. Orlando, M. Picollo, B. Radicati, S. Baronti, A. Casini, Principal component analysis of near-infrared and visible spectra: an application to a XIIth century Italian work of art, Appl. Spectrosc. 49 (4) (1995) 459–465.
[30] J. Blažek, J. Soukup, B. Zitová, J. Flusser, J. Hradilová, D. Hradil, T. Tichý, M3art: a database of models of canvas paintings, in: Euro-Mediterranean Conference, Springer, Limassol, Cyprus, 2014, pp. 176–185.
[31] C. Bonifazzi, P. Carcagnì, R. Fontana, M. Greco, M. Mastroianni, M. Materazzi, E. Pampaloni, L. Pezzati, D. Bencini, A scanning device for VIS-NIR multispectral imaging of paintings, J. Opt. A Pure Appl. Opt. 10 (6) (2008) 064011.
[32] J. Striova, C. Ruberto, M. Barucci, J. Blažek, D. Kunzelman, A. Dal Fovo, E. Pampaloni, R. Fontana, Spectral imaging and archival data in analysing madonna of the rabbit paintings by Manet and Titian, Angew. Chem. Int. Ed. 57 (25) (2018), doi:10.1002/anie.201800624.
[33] R. Fontana, M. Barucci, E. Pampaloni, J. Striova, L. Pezzati, From leonardo to raffaello: insights by vis-IR reflectography, acta artis academica, Interpretation of Fine Art's Analysis in Diverse Contexts, 2014.
[34] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge, Massachusetts, 2016. http://www.deeplearningbook.org
[35] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J.B. Zimmerman, K. Zuiderveld, Adaptive histogram equalization and its variations, Comput. Vis. Graph. Image Process. 39 (3) (1987) 355–368.
[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.