npj | heritage science

Article

https://doi.org/10.1038/s40494-025-01724-9

A deep learning approach for anomaly detection in X-ray images of paintings

Anzhelika Mezina[1] ✉, Radim Burget[1] & Marek Kotrly[2]

The intersection of technological advancements and cultural heritage studies has intensified the exploration of historical treasures, captivating historians and enthusiasts alike. Artificial intelligence now plays a key role in forensic art investigations by uncovering hidden patterns to detect forgeries. This study focuses on anomaly detection in X-ray images of paintings using the Ghent Altarpiece for training and testing purposes. We propose a novel model combining a Discriminatively Trained Reconstruction Anomaly Embedding Model (DRAEM), a Nested U-Net, and a new dataset derived from the Altarpiece. The proposed architecture was benchmarked against several state-of-the-art deep learning techniques in anomaly detection. Our model achieved an accuracy of 0.8399 and an F1 score of 0.7869, outperforming other methods in both accuracy and computational efficiency. Results, validated by a domain expert, show strong precision and computational efficiency through semi-supervised learning.

Cultural Heritage (CH) is critical in epitomizing our collective past and shaping our identity. Digitizing these artifacts enables a broader access to museums and art galleries through virtual exhibitions. For example, supporting the tourist journey[1], augmented reality in museums[2], classification of large 3D CH[3]. However, not only this, various imaging techniques provide additional information about the artwork that cannot be observed by the human eye and can help reveal additional important information. Among others, this information can be used by experts to confirm originality or forgery[4].

Currently, multiple methods are available. For example, chemical analysis stands out as one of the most trustworthy, but it is regrettably invasive[5] and often damages valuable artwork. An automated method for the detection of cracks[6], which is mainly for restoration purposes, was presented. Many cutting-edge technologies like Artificial Intelligence (AI), especially Deep Learning (DL) techniques, have recently demonstrated significant success. These AI applications present many opportunities for researchers and enthusiasts alike, with the added benefit of being non-invasive to the artwork itself. The challenge of pinpointing an area of interest lies in discerning unexpected patterns, especially those that have never been observed. This point presents a significant hurdle for Machine Learning (ML) techniques, as they face the challenge of detecting entirely novel patterns.

Anomaly detection in paintings is the process of identifying irregularities or deviations from the expected norm within an artwork. These anomalies may indicate underlying issues such as deterioration, unauthorized alterations, or the use of materials that do not match the artwork's historical period.

In art conservation and analysis, anomaly detection is frequently performed using several imaging techniques, including X-ray radiography, infrared reflectography, and ultraviolet fluorescence. These methods enable conservators and researchers to observe the underlying layers of a painting and identify hidden layers, restorations, or materials that differ from the original composition.

Anomaly detection can involve statistical models such as statistical outliers, pattern recognition, spectral analysis, dimensionality reduction, ML algorithms, and image processing techniques to analyze data. The objective is to identify patterns that do not align with the painting's known characteristics or the artist's typical style.

Here, unsupervised or semi-supervised anomaly detection algorithms are pivotal in emphasizing suspicious areas that should be the subject of further investigation. They can reveal hidden or altered features in historical art, such as under-drawings, pigments, damages, and restorations, and can be a basis to confirm or disprove the forgery. Unfortunately, anomaly detection algorithms are primarily designed for industrial applications, such as visual quality inspection, and many fail when applied to painting analysis. Another significant obstacle is the difficulty and expense of data collection, the need for expert evaluation and consultation, and the lack of publicly available data, which is vital for DL methods. These aforementioned challenges render the study of paintings nearly inaccessible to most researchers. Only well-equipped teams with access to specialized hardware can gather the necessary data and conduct experiments.

In recent years, rapidly developing information technologies have elevated the analysis and investigation of paintings to a new level. It is

[1]Dept. of Telecommunications, Brno University of Technology, Brno, Czech Republic. [2]Institute of Criminalistics, Prague, Czech Republic.
✉e-mail: anzhelika.mezina@vut.cz

Table 1 | Summary of related works

Ref.	Solved problem	Method	Results
7	Crack detection	CNN	recall 0.6570, precision 0.5624, F1 0.6060
8	Paint loss detection	Translation invariant UNet	average IoU 0.213, average accuracy 0.838
6	Multimodal registration	Fully Convolutional Neural Network (FCN)	success rate of MSE 84.6, detector repeatability 43.6, matching inlier score 68.5
9	Image separation	Learned coupled iterative shrinkage thresholding algorithms, several linear convolutional mapping	MSE 0.0032 and 0.0052
10	Image separation	'Connected' auto-encoders	MSE 0.00078
12	Defect detection	Anisotropic diffusion method	result of opinions experts are presented
13	Artist identification	An ensemble of CNNs	60-96% accurate
14	Pure pigments identification	CNNs	training accuracy 99.93%

possible to extract hidden features, automate labeling, detect objects, and search for similarities. Crack detection is one of the most popular tasks in this field of research, especially when AI algorithms are rapidly developed.

One of those works is introduced in the paper[7]. The authors adapted a Convolutional Neural Network (CNN) to detect cracks combining different modalities. They created the data set by cutting images from the multimodal Ghent Altarpiece dataset into patches of size 15 × 20 centimeters. The work shows improvement in this task by extending the training dataset. The achieved results are precision – 0.7964, and F1 – 0.8185.

Similar work is introduced in paper[8]. The authors utilize a pre-trained U-Net model with dilated convolution to detect the painting loss in multimodal data. Contrary to previous work, this approach solved the segmentation task, and the main metric is Intersection over Union (IoU). The best-achieved result is 0.213. The results were compared with manual expert annotations and also compared with actual physical restorations.

Another approach[6] used CNN to register several modalities of the image using cracks, considering that they are visible in all modalities. Additionally, the data set consists of large German panel paintings from the 15 to 16th century and 16th century portraits by Lucas Cranach the Elder.

The authors of research[9] discuss the problem of concealed earlier designs that were painted over. The main challenge is that mixed X-Ray (XR) image is the 2D representation of 3D dimensional work, which means the XR image contains the features of concealed painting and the visible one. Consequently, it is not very easy to separate these two different natures. With this motivation, the authors proposed a separation network that utilizes the XR image and the RGB image of the surface painting. It consists of two parts: an analysis component and a synthesis component.

Another work of the same authors is in ref. 10. The authors used the convolutional autoencoder to decompose the mixed XR images into two individual images from double-sided paintings. The main advantage of it is the self-supervised principle of work. It can reconstruct the original RGB images, reproduce the respective XR images, and regenerate the mixed XR images. The Ghent Altarpiece dataset was also used for this experiment. Original images were divided into patches 64 × 64, and 9724 patch triplets were generated with an overlap of 8 px. The result achieved by algorithms is Mean Squared Error (MSE) – 0.00078, which is higher than the method[11] used for comparison (0.0011).

There is an approach that focuses on defect detection[12]. The aim is to reveal hidden details and damage in radiography images of paintings. The proposed approach is based on an anisotropic diffusion method that reconstructs images with sharper edges. In this way, details, such as the effect of brushstrokes, the different types of construction wood, etc., are more visible than in the original image.

Approach[13] uses the so-called height data to identify the artist. The small patches are fed into pre-trained CNN backbones, and averaged predicted probabilities give the final decision regarding the author of the painting. One of the conclusions of this work is that the division of images into relatively small patches leads to the loss of the subject's information and the artist's intended style.

Hyperspectral images, combined with hyperspectral metric data, can also be utilized for pigment identification. This research is provided in the paper[14]. The authors proposed a three-branch CNN, which is aimed at visualizing and identifying pure pigments. The first branch focuses on identifying feature maps with reflectance at different wavelengths. The second branch analyses the derivative of the smoothed reflectance. The third branch aims to compute the error between the reflectance of the analyzed pixels and spectral signatures from the reference pigment database. The model output shows the probability that a given sample belongs to someone in the database.

According to the literature studied, the summary of which is presented in Table 1, no works focus on advanced DL anomaly detection methods in XR images to detect damage or suspicious areas, which, consequently, should be studied in more detail by experts. This work aims to detect and visualize such areas using advanced approaches, such as Neural Network (NN)s.

This study aims to bridge this gap by introducing a novel DL-based framework specifically designed for anomaly detection in XR images of paintings. We propose a novel NN architecture that integrates Discriminatively trained Reconstruction Anomaly Embedding Model (DRAEM) and Nested U-Net, tailored for detecting anomalies in XR images. This approach enables the model to detect and segment anomalous areas with higher precision, making it more applicable to the nuanced requirements of cultural heritage research. We developed a dataset of XR images of the Ghent Altarpiece, a well-studied historical artwork, to train and evaluate the model for anomaly detection and validation. The results were compared with several other state-of-the-art anomaly detection methods, such as CS-Flow[15], FastFlow[16], CFA[17], DRAEM[18], Reverse Distillation[19], and STFPM[20].

Our work is unique and can be helpful for further study of paintings. The models were evaluated on the testing set, which was not included in the training phase. In addition, visual and quantitative comparisons are provided. Our findings may be useful for researchers, restorers, and forgery investigations.

The main contributions of this paper are:

- We developed a NN architecture that combines DRAEM with a nested U-Net structure. This integration leverages DRAEM strengths in anomaly detection while enhancing feature extraction capabilities. As a result, it enables better segmentation and visualization of anomalous regions in XR images of paintings.
- Leveraging high-resolution XR images of the Ghent Altarpiece, we curated a specialized dataset of painting patches pre-processed for the anomaly detection task. Each patch was labeled according to the presence of abnormal parts.
- Our architecture was benchmarked against leading models in anomaly detection, including CS-Flow, FastFlow, CFA, and other architectures, which are developed for industrial defect detection. Comparative analysis demonstrated that our model exceeds others in accuracy and computational efficiency, achieving an accuracy of 83.99% and an

F1 score of 0.7869, indicating its robustness in identifying anomalies specific to XR paintings.
- Through both quantitative and qualitative assessments, our model has shown practical potential for real-world applications in art conservation and forensic analysis. Its effective performance in detecting anomalies, validated by expert visual inspections, suggests that it is a suitable assistive tool for professionals in art restoration and historical research.

The rest of this paper is structured as follows. Section "Methods", describes NN models used for the experiment. Section "Results", shows the results of the experiments and provides visualization of detected anomalies. Section "Discussion", discusses the results and provides conclusions.

Methods

This work is conducted in three steps: first, the dataset is prepared. Secondly, the selected methods and the proposed model are trained on the created dataset, and finally, evaluation is performed. The whole experiment is introduced in Fig. 1. Section "Dataset description", describes the dataset and its preparation. Methods and evaluation metrics are introduced in the "Baseline" section. The newly designed and developed model is presented in the "Proposed architecture" section.

Dataset description

The original data used for the experiments is collected from ref.[21]. One of the advantages of this virtual gallery is that the paintings are presented in different modalities: macrophotography, XR, and infrared macrophotography. These modalities were collected before, during, and after restoration. This collection provides a unique opportunity to analyze the progression of restoration and examine areas with potential anomalies, such as material inconsistencies or hidden layers, which are of particular interest to conservators and heritage scientists. Another advantage is that the modalities are co-registered. It is possible to get a painting of some size using macrophotography and XR, and they do not need any co-registration or any other processing.

For the experiment, 19 XR images before restoration and 12 XR images after restoration were collected. All images were of high resolution.

This selection serves multiple purposes. Firstly, processing high-resolution images is necessary due to the Institute of Criminalistics's real-world application requirements.

Secondly, data is needed for reparations to enhance the dataset by including this specific anomaly, which is visible in X-rays but not in macrophotography (commonly referred to as RGB) images).

Processing full-size images with the algorithm is computationally intensive, as a single PNG file can reach around 100 MB and have dimensions in thousands of pixels. Given the relatively limited GPU memory, conducting NN training and testing with full-sized images is nearly impossible in a given environment. Therefore, it was decided to split the images into smaller patches of size 512 × 512 pixels. This approach not only increases the number of samples available for training and testing, but also alleviates the hardware demands, allowing the experiment to run on standard equipment hardware.

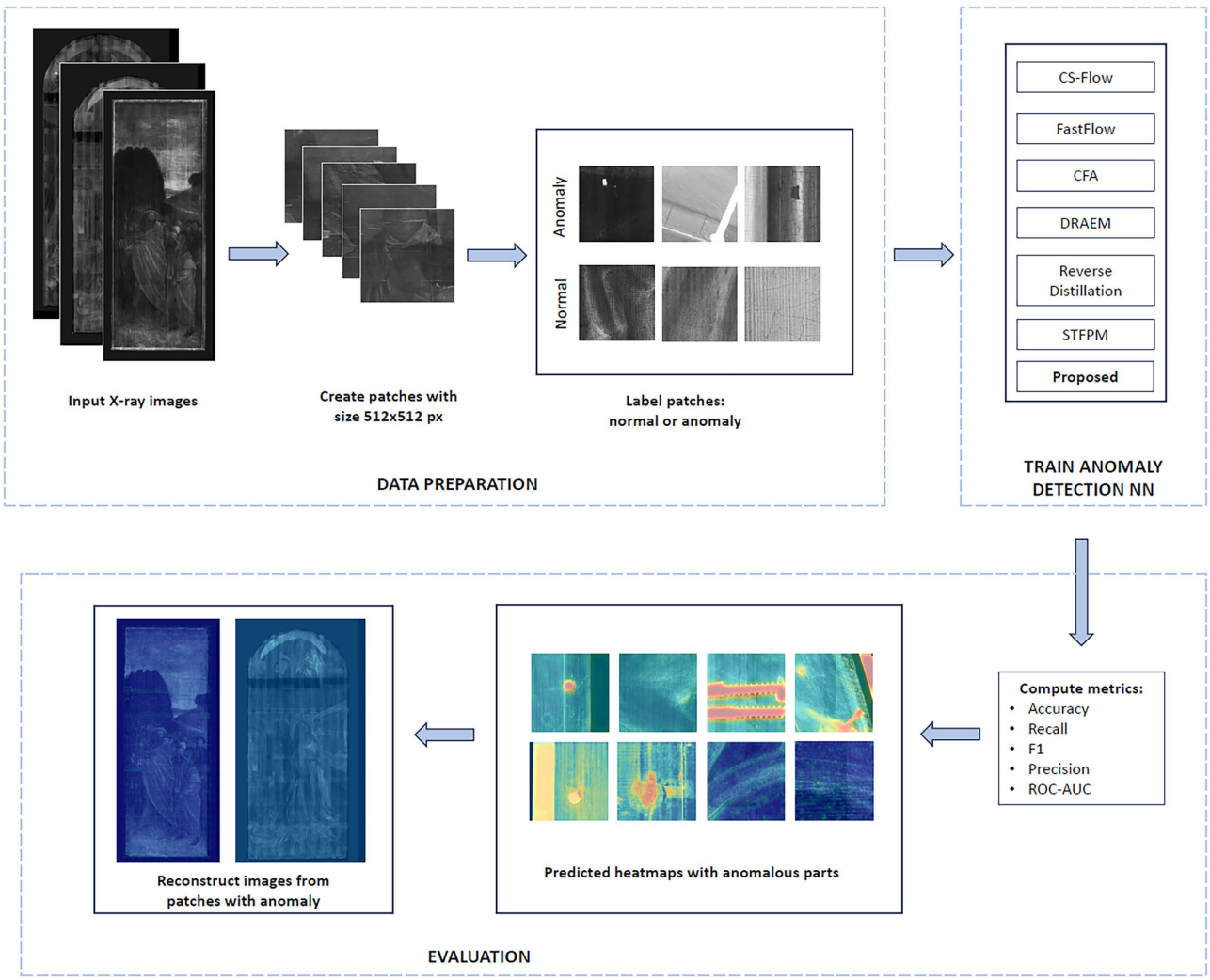
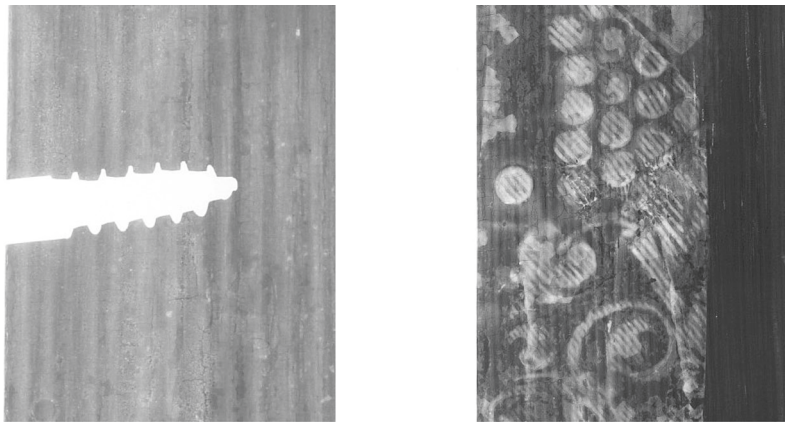


Fig. 1 | The overall scheme of experiment.



(a) Anomaly (b) Normal

Fig. 2 | Examples of patches in the used dataset. **a** Patch with the anomalous part (labelled as "Anomaly"). **b** Patch without anomalous part (labelled as "Normal").

Furthermore, this segmentation enables more precise identification of anomalies and allows deep learning models to process data more efficiently. Each patch is carefully labeled on the basis of expert evaluations, categorizing regions as either "normal" or "anomalous". Anomalies may include irregularities such as paint loss, material degradation, foreign objects, and underdrawings, elements that are vital in the analysis of cultural heritage, but often go unnoticed.

As a result, there are 1372 anomalous samples and 11, 988 normal ones. The example of patches is shown in Fig. 2.

Baseline

This section describes the methods that were used for comparison. Those methods are initially aimed at detecting visual defects in the industrial environment. Since anomalous samples contain damaged or foreign objects, painting loss, which is seen as a very dark place, or, on the contrary, some white spots appeared in XR in unexpected places, the closest research field is anomaly detection in the industry.

For this experiment, the following models were selected: CS-Flow[15], FastFlow[16], CFA[17], DRAEM[18], Reverse Distillation[19], and STFPM[20]. The reason for this choice is the possibility of executing the code without a memory capacity problem and adequate time for the training process.

The main limitation during the attempt to evaluate the other models in this field of research, such as PatchCore[22] or PaDiM[23], is the lack of memory. This complication comes from the principle of work of these models: they use and keep information from normal samples and compute over them. Consequently, the memory complexity depends on the number of normal samples. Their number is much larger than the usually used MVTec AD Dataset[24] size. With this limitation, the possible models that can be applied to this task should be carefully selected. A brief description of selected models is given below.

CS-Flow. This approach belongs to the family of models based on the normalization flow. Generally, this model transforms the data into a tractable distribution. The main idea of CS-Flow is to process the features of images at different scales. It allows the model to utilize information and correlations in local and global contexts and precisely learn distribution to identify defective samples.

FastFlow. FastFlow has a similar principle to CS-Flow. This model also considers local and global features, which can be extracted with the ResNet or Vision Transformer (ViT) model. During the training phase, only normal images are used. FastFlow learns to map the feature distribution of normal images to a standard normal distribution. Probabilities are used as an anomaly score in the testing phase. The main advantage of this architecture is the end-to-end inference phase, which is performed faster than other models such as PatchCore or CFlow. The used backbone is ResNet-18.

CFA. Another way to detect abnormal parts in the image is to extract features using some model and apply the patch descriptor with a memory bank to distinguish the anomalous samples. One of the advantages of this model is the proposed Coupled-hyper-sphere-based feature Adaptation, which adjusts to a target dataset. Another advantage is the memory bank, which is compressed independently of the target dataset size. Consequently, the complexity is decreased and the problems mentioned for PaDiM and PatchCore are solved.

DRAEM. DRAEM[18] is the approach that is based on the principle of Generative Adversarial Network (GAN). The architecture consists of the reconstructive sub-network, which learns to detect and reconstruct anomaly-free content, preserving the non-anomalous part, and the discriminative sub-network, which aims to learn the original and reconstructed images, producing an anomaly detection map. Since the approach simulates the anomaly on anomaly-free samples and it is not necessary to have the anomalous one during training, it is one of the suitable methods for the problem where there are no anomaly samples.

Reverse distillation. This approach is based on the teacher-student model, where the teacher is represented with an encoder – ResNet, which is pre-trained on ImageNet, and the student is with the decoder. The student part processes the one-class embedding produced with the teacher's network to restore the multiscale representation. In this approach, the one-class bottleneck embedding was proposed, effectively preserving the important information from normal samples and ignoring the anomaly perturbations.

STFPM. STFPM has the same principle as the reverse distillation approach, it is a teacher-student model. Both parts, teacher and student, have identical architecture. The teacher is a pre-trained network, distilling knowledge into the student network. The last one is aimed at learning the distribution of non-anomalous images. One of the specific points of this model is the multiscale feature matching, which is supposed to enhance the robustness, allowing the student network to receive multi-level knowledge, detect anomalies of various sizes, and match the features with counterpart features in the teacher network.

Proposed architecture

In addition to the architectures used for comparison, this paper also introduces a novel model of a NN based on the DRAEM architecture. The DRAEM architecture has been described in Section 2, detailed information can be found in the original paper[18], and also, in brief, it is described in this section. The main difference between the proposed model and the

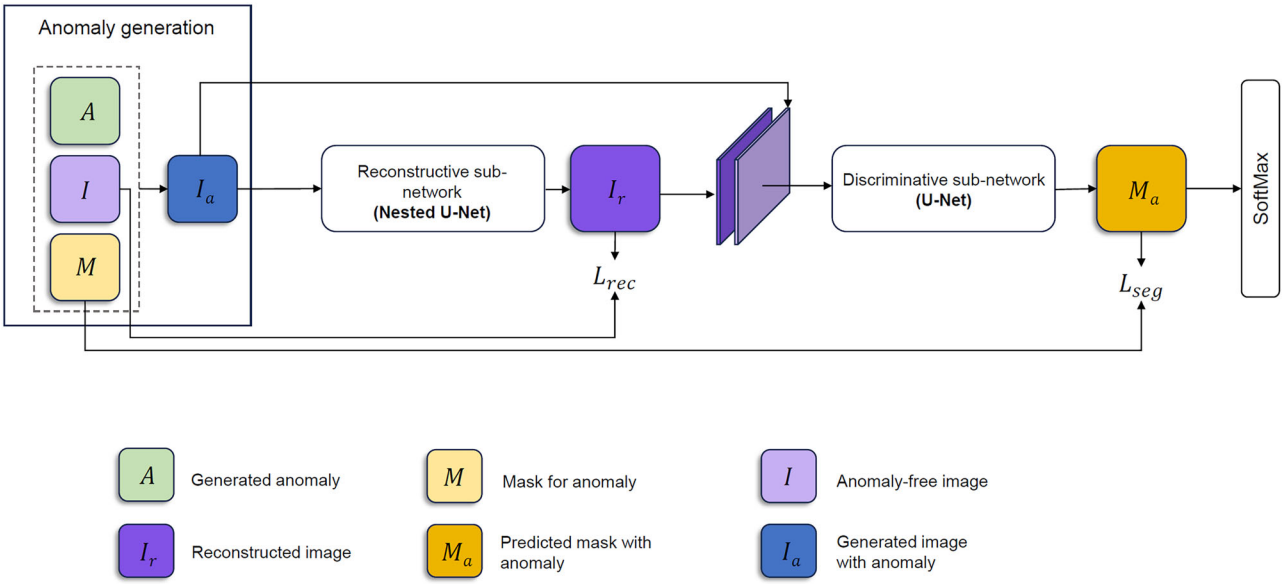


Fig. 3 | Proposed model.

original one is the employment of Nested U-Net (U-Net++)[25] in the reconstruction sub-network. The proposed model also consists of an anomaly generator and reconstructive and discriminative sub-networks. The first sub-network is aimed at detecting and reconstructing anomalies with anomaly-free content. The second sub-network learns reconstruction-anomaly embedding and produces the anomaly segmentation map. This architecture modification is unique in anomaly detection tasks and is applied for the first time to painting analysis. The overall scheme of the proposed model is shown in Fig. 3, and the pseudocode is also presented in Algorithm 1.

The main workflow consists of several steps. Firstly, the anomaly generator provides the anomaly image A containing texture and mask M for this image. After that, both images are merged with normal image I, and this step is formulated as follows:

$$I_a = I * (1 - M) + \beta * A + (1 - \beta) * I * M, \tag{1}$$

where β regulates the opacity of merge and is selected randomly from the range between 0.2 and 1.

The next step is to reconstruct the original image I from an image of artificial anomalies I_a. This step allows to learn the patterns of anomaly-free samples. This part is performed with a reconstructive sub-network. Here, the training of the model is evaluated using the loss function L_{rec}, which is based on Structural Similarity Index Measure (SSIM) value and l_2 and is defined as follows[18]:

$$L_{rec}(I, I_r) = \lambda \frac{1}{N_p} \sum_{i=1}^{H} \sum_{j=1}^{W} 1 - \mathrm{SSIM}(I, I_r)_{(i,j)} + l_2(I, I_r), \tag{2}$$

where H – height of the image, W – width of the image, N_p – number of pixels in the image, I_r – reconstructed image, $SSIM(I, I_r)_{(i,j)}$ – SSIM value of the I and I_r, centered at (i, j) image coordinates, λ – loss balancing parameter, which is equal to 2.

The definitions of SSIM and l_2 are formulated as follows:

$$SSIM(I, I_r) = \frac{(2\mu_I\mu_{I_r} + C_1) + (2\sigma_{II_r} + C_2)}{(\mu_I^2 + \mu_{I_r}^2 + C_1)(\sigma_I^2 + \sigma_{I_r}^2 + C_2)}, \tag{3}$$

$$l_2 = (I - I_r)^2, \tag{4}$$

where μ is the mean, σ^2 is the variance, σ is the covariance of I and I_r and C_1 and C_2 are constants.

Finally, the reconstructed image I_r and the image with anomaly I are processed with a discriminative sub-network, which is presented with a U-Net model to generate an anomaly map M_a. Considering that I and I_r are significantly different (in the case of abnormal images), there is enough information for a segmentation map.

To increase the accuracy of the segmentation capability of the sub-network, the focal loss function[26] is applied in this part:

$$L_{seg} = -\alpha_t(1 - p_t)^{\gamma} \log(p_t), \tag{5}$$

where p_t – the model's estimated probability for the class with label $y = 1$, $\alpha_t = 1$ – balancing factor, $\gamma = 2$ – modulating factor.

The total loss function of the model is computed as follows:

$$L = L_{rec}(I, I_r) + L_{seg}(M_a, M), \tag{6}$$

where M_a – ground truth of anomaly segmentation mask, M – predicted one.

Algorithm 1. Proposed Model for Anomaly Detection

Require: Training dataset (X_{train}, M_{train}) with normal images I and anomaly masks M

Require: Hyperparameters: α, β, γ, λ

Training

Initialize anomaly generator A_{gen}

Initialize reconstructive sub-network (Nested U-Net)

Initialize discriminative sub-network (U-Net)

for each epoch in 1, …, max_epochs **do**

  for each batch (I, M) in X_{train} **do**

    Step 1: Generate Synthetic Anomalies

    Generate anomaly texture $A \leftarrow A_{\text{gen}}(I, M)$

    Create anomalous image $I_a \leftarrow I \cdot (1 - M) + \beta \cdot A + (1 - \beta) \cdot I \cdot M$

  Step 2: Reconstruct Original Image

  Reconstructed image $I_r \leftarrow$ Nested_UNet(I_a)

  Compute reconstruction loss:

    $L_{\text{rec}} \leftarrow \lambda \cdot \left[1 - \mathrm{SSIM}(I, I_r) + \| I - I_r \|_2^2\right]$

    Step 3: Segment Anomalies

    Predicted anomaly mask $M_{\text{pred}} \leftarrow$ U_Net(I_r, I_a)

    Compute segmentation loss (focal loss):

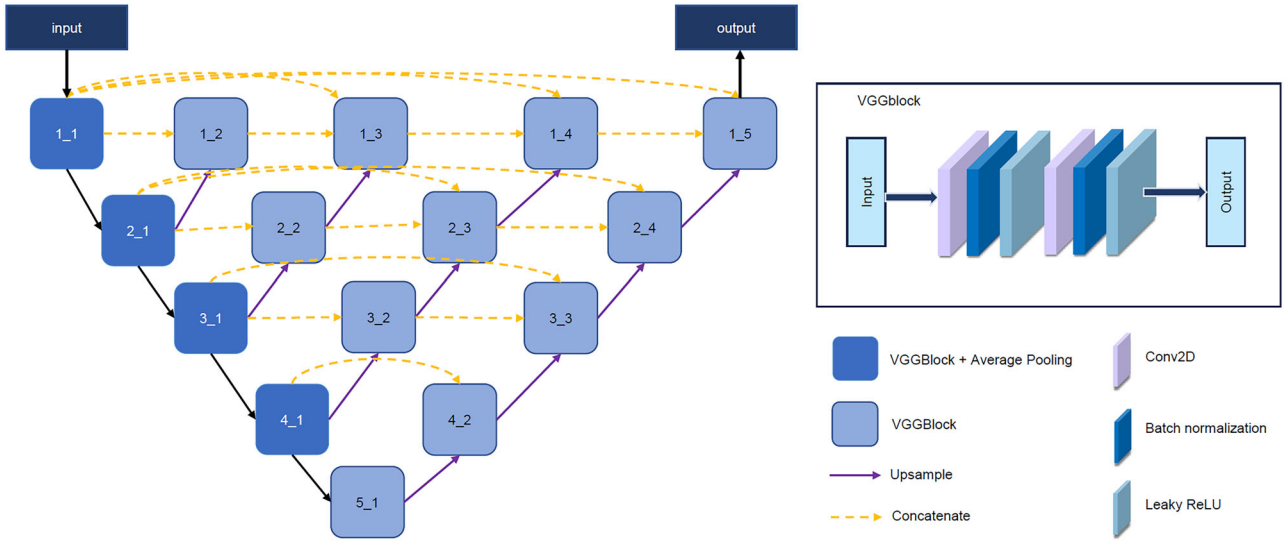


Fig. 4 | Nested U-Net.

$L_{seg} \leftarrow -\alpha_t \cdot (1 - p_t)^{\gamma} \cdot \log(p_t)$
Step 4: Compute Total Loss and Update Weights
$L_{total} \leftarrow L_{rec} + L_{seg}$
Backpropagate L_{total} and update weights
end for
end for
Evaluation:
for each test image I_{test} in X_{test} **do**
Reconstruct image: $I_r \leftarrow$ Nested_UNet(I_{test})
Predict anomaly mask: $M_{pred} \leftarrow$ U_Net(I_r, I_{test})
end for
Compute performance metrics: Accuracy, Precision, Recall, F1, ROC-AUC, create heatmaps
Output: Trained model, computed metrics, generated heatmaps

Reconstructive sub-network. The reconstructive sub-network consists of the so-called Nested U-Net, which is an extended version of the original U-Net model.

The basic encoder-decoder parts of U-Net model can be mathematically formulated as:

Encoder:

$$\mathbf{X}_l = \sigma(\mathbf{W}_l * \mathbf{X}_{l-1} + \mathbf{b}_l) \tag{7}$$

where $\mathbf{W}_l$ represents the learned weights for the l-th layer, $\mathbf{b}_l$ represents the bias terms for the l-th layer, * represents the convolution operation, σ is a non-linear activation function Rectified Linear Unit (ReLU).

Decoder:

$$\mathbf{Y}_l = \sigma(\mathbf{W}'_l *' \mathbf{X}_l + \mathbf{b}'_l) \tag{8}$$

where, $*'$ represents the transposed convolution (up-sampling) operation, $\mathbf{W}'_l$ and $\mathbf{b}'_l$ are the weights and bias for the decoder layer.

The main advantage of the Nested U-Net model is the ability to learn a representation of input data in different levels of abstraction and connect them through skip-connection paths. Additionally, low-level representations are tuned to the higher levels, which allows to fill the semantic gap between different levels. It can improve the reconstruction capabilities of the model. The detailed architecture scheme is introduced in Fig. 4.

Mathematically, the skip pathway can be formulated as[25]:

$$x^{i,j} = \begin{cases} \mathcal{H}(x^{i-1,j}), & j = 0 \\ \mathcal{H}\left(\left[[x^{i,k}]_{k=0}^{j-1}, \mathcal{U}(x^{i+1,j-1})\right]\right) & j > 0 \end{cases} \tag{9}$$

where $x^{i,j}$ – output of node $X^{i,j}$, i – index of s the down-sampling layer, j – index of the convolutional layer within the dense block along the skip connection, $\mathcal{H}$ – an activation function applied after a convolution operation, $\mathcal{U}$ – an up-sampling layer, [] – the concatenation layer. This structure allows the nested U-Net to have multiple levels of skip connections, leading to a richer set of features.

Each cell is represented by the so-called VGG block, which consists of a convolutional layer with a kernel size of 3 and a number of filters in each level (from 1 to 5): 32, 64, 128, 256, and 512, batch normalization and leaky ReLU activation function with a negative slope of 0.2. Increasing tensor dimensions between different layers is performed with the Upsampling layer with the bilinear algorithm.

Discriminative sub-network. The discriminative network is represented with the original U-Net model[27]. The overall scheme is depicted in Fig. 5. The advantage of this model is the application of the skip connections, which allow the transfer of the semantic information from the encoder part to the decoder:

$$Y_l = \text{Concatenate}(\text{Upsample}(Y_{l+1}), Z_l), \tag{10}$$

where Y_l is the output feature map after the convolution and upsampling, Z_l is the feature map from the encoder, used in the skip connection.

Each level in the encoder part consists of 2 convolutional blocks (a convolutional layer, batch normalization, and ReLU activation function) and a max pooling layer to reduce the dimensions of the input tensor and preserve only significant information.

Generally, the encoder path can be formulated as

$$Z_l = g(W_l * Z_{l-1} + b_l), \tag{11}$$

where Z_{l-1} – Input feature map from the previous layer, $*$ – convolution operation, W_l and b_l – weights and biases of the convolutional layer at level l. $g()$ – activation function.

In the decoder part, the level contains the Upsample layer and three convolutional blocks, concatenated with blocks from the respective level in the encoder part. The numbers of feature maps in the levels are 64, 128, 256

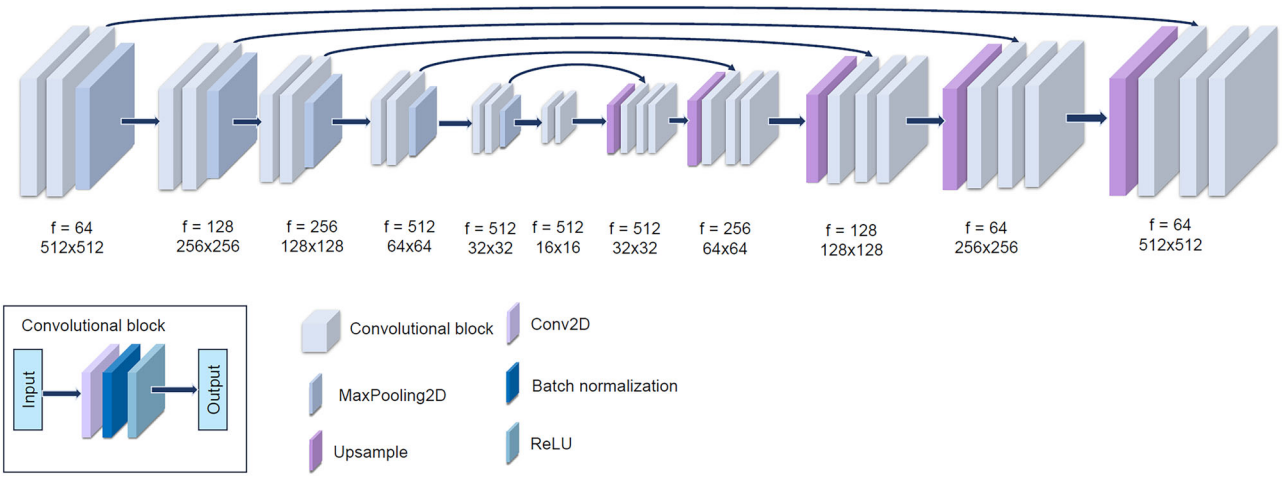


Fig. 5 | U-Net.

and 512, and the input tensor size is gradually reduced from 512×512 to 32×32 (see Fig. 5). Upsampling is also performed using a bilinear algorithm.

The mathematical formulation is the following:

$$Y_l = g(W_l^d * \text{Upsample}(Y_{l+1}) + W_l^s * Z_l + b_l^d), \tag{12}$$

where W_l^d are the weights of the decoder convolutional layer, W_l^s are the weights for the skip connection, Z_l is the feature map from the encoder, used in the skip connection, Y_l is the output feature map after the convolution and upsampling.

Results

This section represents the results obtained from the mentioned anomaly detection methods and compares them.

Metrics

Considering that the created dataset contains only labels if an anomaly is presented, the task to be solved is binary classification. Except for the subjective evaluation of generated heatmaps with abnormal areas, the following objective metrics were used to evaluate and compare the trained models[28]:

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{13}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{15}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{16}$$

where TN – True Negatives, TP – True Positives, FP – False Positives, FN – False Negatives.

Objective evaluation results

The results are shown in Table 2, where the best results are shown in a bold. Based on the objective parameters, the best-performing model is our proposed model: accuracy 0.8399, F1 0.7869, precision 0.8383, DRAEM model is better regarding ROC-AUC 0.9024. Reverse Distillation has the highest value of recall – 0.9836. On the other hand, this model achieved an accuracy of 0.4935 and a precision 0.4396. This indicates that the model just labeled all samples as positive (anomalous). The second best approach is DRAEM: accuracy 0.8192, F1 0.7773, precision 0.7636, which is worse for 1 – 7% than our model. The third best approach is CFA: accuracy 0.7422, F1 0.6784, precision 0.6748, ROC-AUC 0.7951.

The worst results are obtained by STFPM: ROC-AUC 0.4273, accuracy 0.4125, F1 0.5608, precision 0.3994, and however, recall 0.9408. The reason for this is similar to reverse distillation – major samples are labeled as anomalous.

Table 2 also provides the time in seconds to predict one patch. The fastest is FastFlow – 0.0425 s. The second is the proposed model – 0.0824 s. On the other hand, reverse distillation is the most time consuming in predicting 0.1594 s. DRAEM is also relatively time-consuming – 0.1402 s. However, this method gives the best results according to the rest of the metrics. Here, it can be concluded that the choice of model to be used for a real-world application depends on the system's priority where it will be integrated: if time processing is essential or if precious results are required. In the first case, the proposed model can be the most suitable, since it provides relatively accurate results and is one of the fastest algorithms among the tested ones. In another case, if the processing time for the input image is not important but accuracy is a crucial metric, the DRAEM can provide more accurate results, but with time-consuming computing. It can be seen that the proposed model is worse only for 3% considering accuracy, but according to other metrics, it is better and faster than the original DRAEM, which makes the model more suitable for real-world applications.

Subjective evaluation results

Figure 6 provides the results of generated heatmaps, which indicate the anomalous parts of patches. The results of each method are shown in an individual row. The first three samples in a row are supposed to be abnormal, and the last two are normal. As can be seen, the most significant detections are provided by CFA, DRAEM, and proposed models. These three methods found the pins and identified the white areas where they were not expected to appear. Notably, the results correspond with the objective evaluation, and they achieved better results than others. In addition, comparable results from FastFlow can be considered. The interesting point is that the heatmaps produced by CFA, DRAEM, and proposed models indicate more details, which can be abnormal. At the same time, normal samples are left without highlighted areas, which admits that models can differentiate anomalous parts from normal ones. Additionally, the proposed model shows more confident highlights of anomalous areas, with more precision edges of anomaly and more differing in contrast to the background.

CS-Flow, STFPM, and Reverse Distillation provide the worst results. CS flow and reverse distillation did not even differ in anomalous and normal parts in the XR, however, STFPM, as can be seen, detects at least edges for abnormal parts.

Table 2 | Results for all tested models

Method	ROC-AUC	Accuracy	F1	Precision	Recall	Time, sec per patch
CFA	0.7951	0.7422	0.6784	0.6748	0.6821	0.0991
CS-Flow	0.7021	0.6126	0.6370	0.5084	0.8525	0.1213
DRAEM	**0.9024**	0.8192	0.7773	0.7636	0.7914	0.1402
FastFlow	0.7853	0.7153	0.6567	0.6324	0.6831	**0.0425**
STFPM	0.4273	0.4125	0.5608	0.3994	0.9409	0.1180
Reverse Distillation	0.6746	0.4935	0.6076	0.4396	**0.9836**	0.1594
Proposed	0.8755	**0.8399**	**0.7869**	**0.8383**	0.7413	0.0824

On the other hand, CFA and FastFlow show the anomaly parts lightly in normal samples (see the last two columns).

To make the results clearer and more reasonable, the first row contains the macrophotography, which corresponds to the evaluated XR and presents the results in the rest of the rows. From the given macrophotographies, it is obvious that columns 1-3 contain anomalous parts, while columns 4 and 5 do not. It can also be seen that DRAEM and the proposed model successfully detected suspicious areas. In contrast, CS-Flow and Reverse Distillation failed.

Discussion

This work compares well-known anomaly detection approaches, which are used mainly in industry, for anomaly detection in XR images of paintings. The objective results are much worse than those achieved on the MVTec AD dataset. The main point is that most implementations and papers provide comparisons only on that dataset and focus only on the industrial field of research. Consequently, these methods can achieve success on benchmarks or on some limited datasets. It is also essential to note that the results of such methods in real-world applications can be worse since there are more complicated conditions, and how they will deal with them is an open question.

Despite that, these approaches can also be used for purposes other than industrial anomaly detection, for example, analysis of XR images or macrophotography of CH, as was applied in this work. This work identified several challenges.

First, there is a lack of publicly available data. It can be described as the investigation of CH is a narrow specialty and is not so popular among most researchers, such as medicine. Another problem is that fewer people can get the original paintings and, more importantly, get very specific and expensive hardware, which can provide an XR of the painting. Consequently, most works related to the processing of XR of paintings focus on the limited number of samples and do not provide these data, which makes them valuable.

Secondly, an expert in this field is required to label these data. It is not easy to determine if any weird area is anomalous or normal. For this aim, modifying the models to correspond to this problem will probably be necessary. In this work, such abnormal patches were manually identified. After such labeling, the solution to the problem is becoming similar to the problem of anomaly detection in the industry.

Third, not all state-of-the-art approaches can process the dataset after creating it due to limited hardware abilities. Several models, which are based on memorizing the features from normal patches and have a great complexity, failed because of that. This problem leads to the conclusion that it is necessary to develop a model that is independent of the size of the dataset and that can process many samples.

The objective and subjective evaluations correspond to each other in terms of the results achieved. An interesting point is that CFA, which is based on a memory bank, and DRAEM and the proposed model, which has the principle of GAN, are more successful than others. Worse results are achieved by normalization flow-based models (CS-Flow and FastFlow). The models, based on the teacher-student principle, failed to detect any anomaly in the given samples. Here, the trend can be seen: the memory bank principle and the GAN-based model are the more suitable options for this problem. Here, it is essential to note that CFA has the same principle as PatchCore. However, thanks to the reduced complexity of the model, it is possible to conduct experiments and train the model. On the other hand, normalization flow-based models should be used with caution: some of them can detect some anomalies, but others fail, such as CS-Flow. The last category used is teacher-student-based architectures. As can be seen, objectively and subjectively, they cannot learn the abnormalities from a given data set.

The proposed methodology has the potential to be applied in real-world applications. In this work, the images were also reconstructed from patches to see how they look in the whole image. Since the original image has several thousand dimensions, the cropped part of the example is introduced in Fig. 7. As can be seen, the XR image contains several white areas that have not appeared in some way in macrophotography. Figures 7c, d and e are the results of the three most successful methods: CFA, DRAEM and proposed one. As can be seen, they detected those anomalous parts mentioned above. In particular, despite the white areas in the eyes in the image, the methods paid less attention to them and did not identify them as anomalous. This example shows that the tested models can be used for real-world applications as the assistant tool for anomaly detection in XR of paintings.

The effectiveness of the proposed model illustrates how DL can be applied beyond just the detection of industrial anomalies, addressing the advanced requirements of heritage science. Unlike standard industrial uses, cultural heritage analysis requires non-invasive techniques to maintain the artifact's structural integrity. Our method meets this criterion, enabling conservators to detect hidden damages or modifications within paintings without direct contact. The effectiveness of the proposed method was evaluated using Ghent Altarpiece data, which proves that the method is suitable for analyzing fragile and valuable work of art. Additionally, the model's capability to function with minimal labeled data is especially beneficial in heritage science, where obtaining annotated data can be costly and labor-intensive.

Here, this work can be summed up as follows. This work aims to develop an approach and compare it with several methods for anomaly detection for the processing of XR images of paintings and to show the potential of their application to this field of research. First, a dataset based on images of Ghent Altarpiece paintings is created and pre-processed for the experiments. Secondly, the following models: CS-Flow, FastFlow, CFA, DRAEM, Reverse Distillation, and STFPM were trained and tested on the created dataset. Third, a novel architecture based on the DRAEM and Nested U-Net is proposed. According to the results, the best performing is DRAEM, which is better than others, at least for 15%. However, our proposed model is faster and has objective metrics that are very close to the original DRAEM, which makes this model accurate and less time-consuming. Subjectively, this architecture successfully detects anomalous parts in images. Also, the CFA model can be considered a suitable one for this task. During experiments, several complications are encountered, such as a lack of publicly available data, a lack of previous

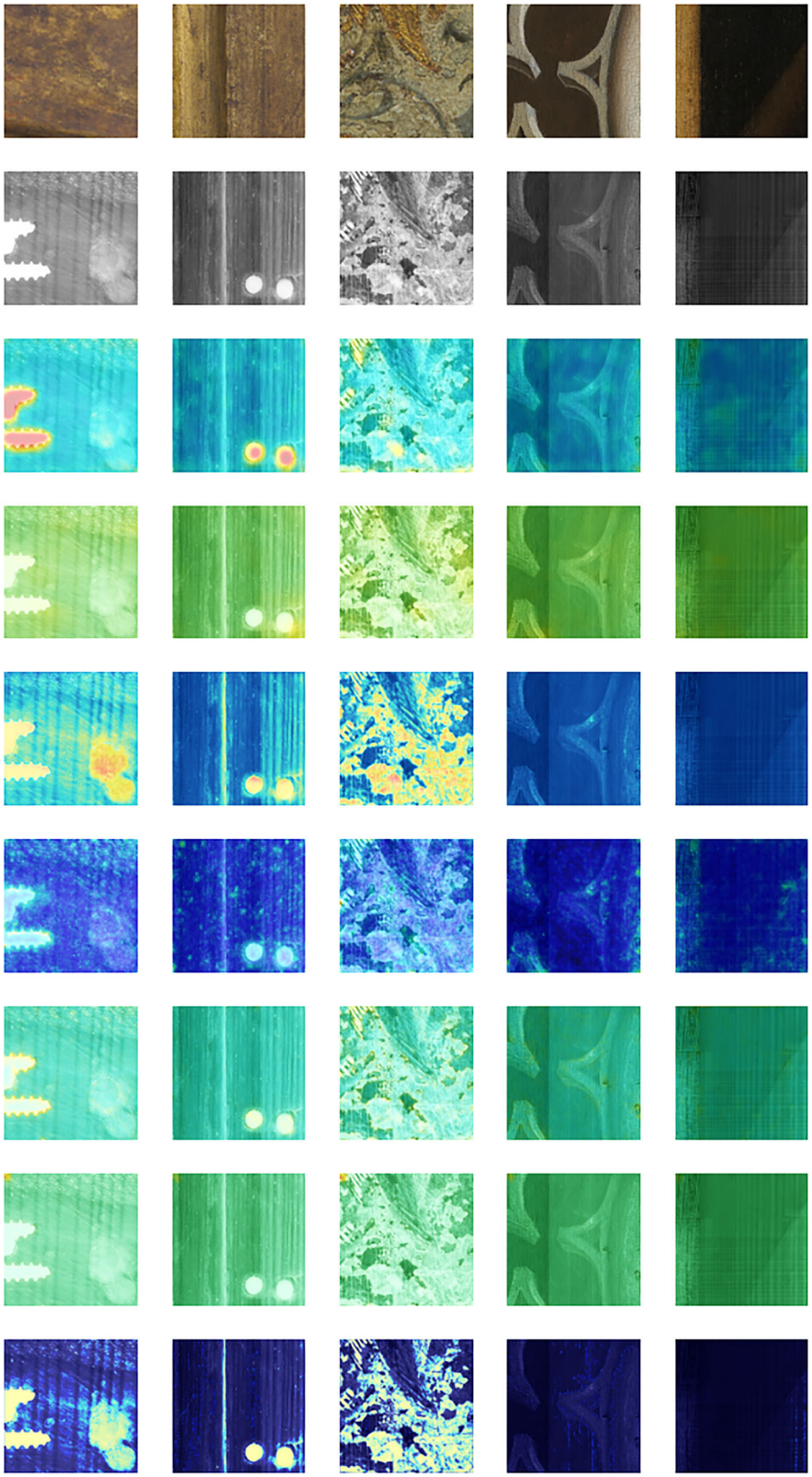
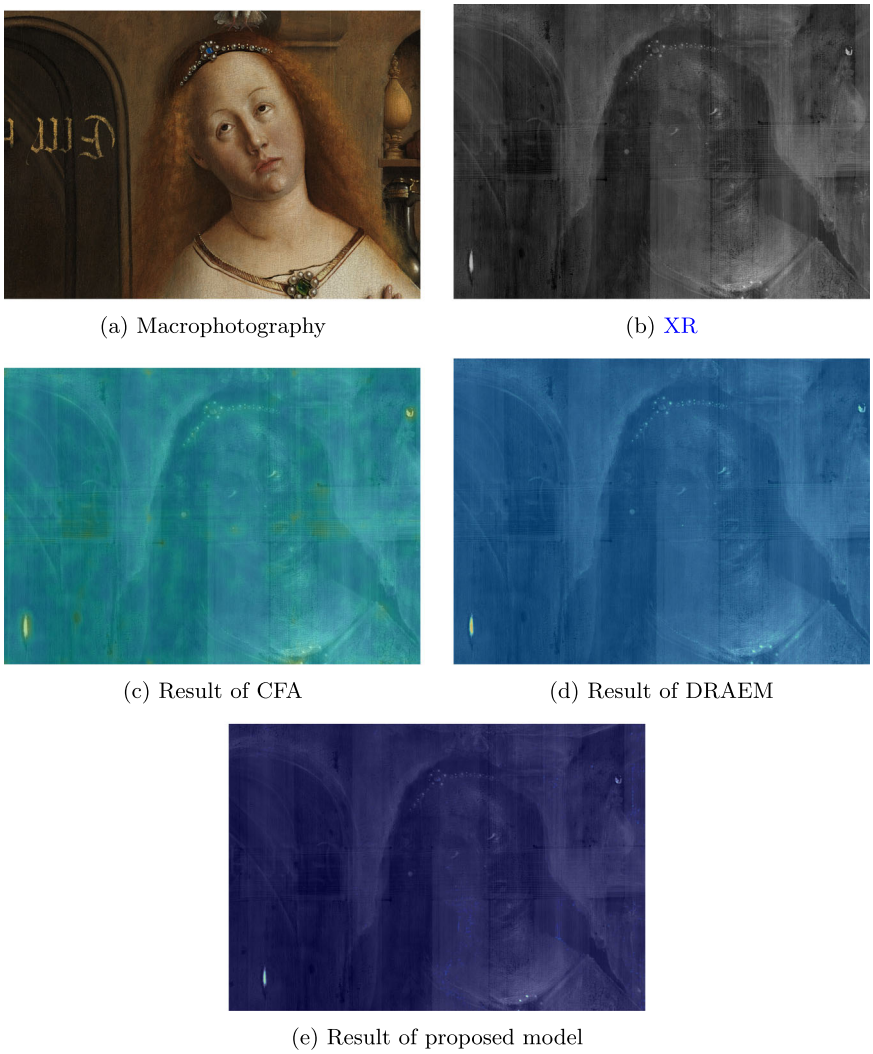


Fig. 6 | Heatmaps with detected anomalous areas for different trained models. Original macro-photography: 1st row, original XR: 2nd row, CFA: 3rd row, CS-Flow: 4th row, DRAEM: 5th row, FastFlow: 6th row, STFPM: 7th row, Reverse Distillation: 8th row, Proposed: 9th row.

research on this type of problem in this field, limitations of hardware, and so on.

As future work can collect data with the following labeling, a more detailed investigation of methods can be applied to these data. It is possible to modify existing approaches to focus more on XR images of painting, not just defects on different things, as in the MVTec AD dataset.

Our findings indicate that DL models such as the proposed one have the potential to revolutionize art conservation through the wide-scale implementation of precise anomaly detection. Specifically, the model can help recognize early signs of degradation, thus facilitating preventive conservation measures. With further advancements, this technology might lead to automatic monitoring systems that consistently evaluate state-of-the-art



(a) Macrophotography (b) XR

(c) Result of CFA (d) Result of DRAEM

(e) Result of proposed model

Fig. 7 | A comparison of models' results. a The macrophotography of the given XR image. **b** XR used as an input into the model. **c** The predicted result of the CFA model. **d** Predicted result of the original DRAEM model. **e** Predicted result of the proposed model.

collections, promptly notifying conservators of developing issues before they escalate.

Data availability

No datasets were generated or analysed during the current study.

Received: 8 November 2024; Accepted: 19 April 2025;
Published online: 02 May 2025

References

1. Sperlí, G. A cultural heritage framework using a deep learning based chatbot for supporting tourist journey. *Expert Syst. Appl.* **183**, 115277 (2021).
2. Khan, M. A. et al. Using augmented reality and deep learning to enhance taxila museum experience. *J. Real.-Time Image Process.* **18**, 321–332 (2021).
3. Matrone, F. et al. Comparing machine and deep learning methods for large 3d heritage semantic segmentation. *ISPRS Int. J. Geo-Inf.* **9**, 535 (2020).
4. Manfriani, C. et al. The forger's identikit: A multi-technique characterization of pippo oriani's fake paintings. *Dyes Pigments* **207**, 110755 (2022).
5. Sirro, S. et al. Recognition of fake paintings of the 20th-century russian avant-garde using the physicochemical analysis of zinc white. *Forensic Chem.* **26**, 100367 (2021).
6. Sindel, A., Maier, A. & Christlein, V. Craquelurenet: matching the crack structure in historical paintings for multi-modal image registration. In *2021 IEEE International Conference on Image Processing (ICIP)*, 994–998 (IEEE, 2021).
7. Sizyakin, R. et al. Crack detection in paintings using convolutional neural networks. *IEEE Access* **8**, 74535–74552 (2020).
8. Meeus, L. et al. Assisting classical paintings restoration: efficient paint loss detection and descriptor-based inpainting using shared pretraining. In *Optics, Photonics and Digital Technologies for Imaging Applications VI*, vol. 11353, 99–110 (SPIE, 2020).
9. Pu, W. et al. Mixed x-ray image separation for artworks with concealed designs. *IEEE Trans. Image Process.* **31**, 4458–4473 (2022).
10. Pu, W. et al. A connected auto-encoders based approach for image separation with side information: With applications to art investigation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2213–2217 (IEEE, 2020).
11. Sabetsarvestani, Z., Sober, B., Higgitt, C., Daubechies, I. & Rodrigues, M. Artificial intelligence for art investigation: Meeting the challenge of separating x-ray images of the ghent altarpiece. *Sci. Adv.* **5**, eaaw7416 (2019).
12. Garcia, J. A. M., Yahaghi, E. & Movafeghi, A. Improvement of the digital radiographic images of old paintings on wooden support through the anisotropic diffusion method. *J. Cultural Herit.* **49**, 115–122 (2021).

13. Ji, F. et al. Discerning the painter's hand: machine learning on surface topography. *Herit. Sci.* **9**, 1–11 (2021).
14. Chen, A., Jesus, R. & Vilarigues, M. Convolutional neural network-based pure paint pigment identification using hyperspectral images. In *ACM Multimedia Asia*, 1–7 (2021).
15. Rudolph, M., Wehrbein, T., Rosenhahn, B. & Wandt, B. Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1088–1097 (2022).
16. Yu, J. et al. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677* (2021).
17. Lee, S., Lee, S. & Song, B. C. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access* **10**, 78446–78454 (2022).
18. Zavrtanik, V., Kristan, M. & Skočaj, D. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330–8339 (2021).
19. Deng, H. & Li, X. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9737–9746 (2022).
20. Wang, G., Han, S., Ding, E. & Huang, D. Student-teacher feature pyramid matching for anomaly detection. *arXiv preprint arXiv:2103.04257* (2021).
21. KIK/IRPA. Theghentaltarpiecerestored. http://closertovaneyck.kikirpa.be/ [Accessed 15.08.2023] (2023).
22. Roth, K. et al. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328 (2022).
23. Defard, T., Setkov, A., Loesch, A. & Audigier, R. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, 475–489 (Springer, 2021).
24. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D. & Steger, C. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *Int. J. Computer Vis.* **129**, 1038–1059 (2021).
25. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, 3–11 (Springer, 2018).
26. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988 (2017).
27. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241 (Springer, 2015).
28. Mezina, A. et al. Corticosteroid treatment prediction using chest x-ray and clinical data. *Computational Struct. Biotechnol. J.* **24**, 53–65 (2024).

Acknowledgements

This work was supported by the Ministry of the Interior of the Czech Republic, under Grant no.VK01010153.

Author contributions

A.M.: Writing – original draft, Visualization, Validation, Methodology, Conceptualization. R.B.: Writing – review & editing, Validation, Supervision, Project administration. M.K.: Writing – review & editing, Formal analysis. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Anzhelika Mezina.