Journal of Real-Time Image Processing (2020) 17:1911–1926
https://doi.org/10.1007/s11554-020-00975-y

SPECIAL ISSUE PAPER

# Investigating low-delay deep learning-based cultural image reconstruction

Abdelhak Belhi[1,2] · Abdulaziz Khalid Al-Ali[1] · Abdelaziz Bouras[1] · Sebti Foufou[3] · Xi Yu[4] · Haiqing Zhang[5]

Received: 2 November 2019 / Accepted: 6 April 2020 / Published online: 9 June 2020

**Abstract**
Numerous cultural assets host a great historical and moral value, but due to their degradation, this value is heavily affected as their attractiveness is lost. One of the solutions that most heritage organizations and museums currently choose is to leverage the knowledge of art and history experts in addition to curators to recover and restore the damaged assets. This process is labor-intensive, expensive and more often results in just an assumption over the damaged or missing region. In this work, we tackle the issue of completing missing regions in artwork through advanced deep learning and image reconstruction (inpainting) techniques. Following our analysis of different image completion and reconstruction approaches, we noticed that these methods suffer from various limitations such as lengthy processing times and hard generalization when trained with multiple visual contexts. Most of the existing learning-based image completion and reconstruction techniques are trained on large datasets with the objective of retrieving the original data distribution of the training samples. However, this distribution becomes more complex when the training data is diverse making the training process difficult and the reconstruction inefficient. Through this paper, we present a clustering-based low-delay image completion and reconstruction approach which combines supervised and unsupervised learning to address the highlighted issues. We compare our technique to the current state of the art using a real-world dataset of artwork collected from various cultural institutions. Our approach is evaluated using statistical methods and a surveyed audience to better interpret our results objectively and subjectively.

**Keywords** Digital heritage · Image reconstruction · Low-delay reconstruction · Image inpainting · Deep learning · Image clustering

## 1 Introduction

Art and cultural heritage represent key elements that define human identity as these artifacts represent the most important medium for the transfer of history between generations and civilizations. As a result, people are more and more interested in discovering this cultural heritage. The value and the attractiveness of these artifacts are tightly tied to their physical condition and the availability of their metadata. Unfortunately, a large portion of these assets are in a degraded state or their history is lost. As a result, institutions all over the world are funding research efforts to tackle the challenges related to cultural data curation. Many databases of visual artwork and museum collections were recently opened for researchers in order to develop applications and technologies for cultural heritage promotion. Our main focus through this research is the visual restoration of damaged artwork. In this regard, we are tackling the challenge of

✉ Abdelhak Belhi
abdelhak.belhi@qu.edu.qa

Abdulaziz Khalid Al-Ali
a.alali@qu.edu.qa

Abdelaziz Bouras
abdelaziz.bouras@qu.edu.qa

Sebti Foufou
sfoufou@u-bourgogne.fr

Xi Yu
yuxi@cdu.edu.cn

Haiqing Zhang
haiqing_zhang_zhq@163.com

[1] CSE, College of Engineering, Qatar University, Doha, Qatar

[2] DISP Laboratory, Université Lumière Lyon 2, Lyon, France

[3] Le2i Lab, Université de Bourgogne, Dijon, France

[4] School of Information Science and Engineering, Chengdu University, Chengdu, China

[5] Chengdu University of Information Technology, Chengdu, China

visually completing and reconstructing damaged artwork through advanced artificial intelligence techniques.

Visual completion or as often referred to as “image inpainting” in the literature is a set of techniques used to reconstruct lost portions in visual captures. These techniques are also applied to the task of object removal found in several image editing tools such as Adobe Photoshop. Although in some cases completing a small missing region of a photograph seems trivial, performing this completion, using computer-based solutions, is a hard challenge. Moreover, completing a large or a complex region full of textures is a very challenging problem even for human experts.

Recently, a lot of research efforts were spent to tackle this challenge and many contributions were proposed. Some of these contributions focus more on completing smaller regions using diffusion methods [3–5] such as propagating neighboring pixels information to the missing regions. But these solutions become quickly ineffective for larger regions. Another type of approaches relies on statistical models to find patches either from the image itself or from a huge database that fit in the missing region and blend with its surroundings [6]. The disadvantages of these techniques are twofold: their reliance on huge visual databases (millions of photographs) and the lack of generalization (failing to find convincing results).With the proven performance of deep learning techniques, CNN-based techniques [7–11] were introduced to perform the task of inpainting. These methods were further boosted by the use of the adversarial training concept introduced in [12]. Currently, deep learning-based image completion techniques achieve state-of-the-art performance in terms of completion quality, but there are still a lot of challenges on how to make these techniques generalize for the majority of completion scenarios.

In this regard, and after the analysis of several image inpainting techniques while having an in-depth focus on those based on deep learning, it turns out that most of the existing methods are performing well at completing images having the same visual context as their training dataset. However, the majority of these techniques lack the ability of diversification in terms of completion. In fact, when we try to train most of these techniques on diversified visual contexts, the training process becomes harder (models harder to train) and takes a considerable amount of time (high delay), and the output quality is heavily degraded as the multiple contexts widen the search space for the completion [13]. As a solution, some techniques require more complex models to handle such limitations. Although some of these models are fine-tuned, their completion results look natural only when they are trained with restrained visual contexts. Also, these models underperform when trained with image datasets representing a variety of visual contexts. To overcome the highlighted limitations, we propose a framework that leverages faster architectures and better performing image reconstruction methods. Our framework aims at overcoming the lack of generalization of these methods while at the same time reconstructing damaged inputs with low delay.

Following the study of the cultural datasets at our hands, it turns out that there are thousands of art styles, and training a single completion model to handle the completion of all the visual categories is not efficient and an over-ambitious objective. Thus, the core of our cultural inpainting approach relies on the divide-and-conquer strategy, as rather than looking for complex architectures, we leverage visual data clustering to split the training data into smaller clusters regrouping samples with similar visual contexts. The resulting clusters are then used to train multiple inpainting model instances instead of training a single model instance on the whole training data.

Through this paper, we present our design of a two-stage cultural inpainting framework. In the first stage, the training data is preprocessed and filtered. Then, a clustering procedure is proposed to regroup data samples with similar visual contexts. In a second stage, and for each resulting cluster of images, a completion model is trained until achieving a good quality reconstruction. At prediction time, the clustering model is used to assign the image to complete to its corresponding cluster. Finally, the completion model trained on this cluster is used for the completion of the asset.

Our main contributions through this paper are as follows:

1. A review of recent best-performing image inpainting techniques while focusing on deep learning-based solutions.
2. Design and implementation of a low-delay image inpainting framework that leverages a clustering model and a range of specific completion models to perform low-delay good quality cultural image completion.
3. Evaluation of our framework on real-world cultural data while showcasing improvements in terms of visual quality over state-of-the-art inpainting frameworks for artwork completion and reconstruction.

The remainder of this manuscript is arranged as follows. In Sect. 2, we review the most notable contributions for image inpainting and visual data clustering focusing on deep-learning-based solutions. In Sect. 3, we present the methodology of our approach and the architecture of our framework. In Sect. 4, we present the datasets we used to test and validate our framework in addition to the experimental setup and the evaluation results. In Sect. 5, we discuss and

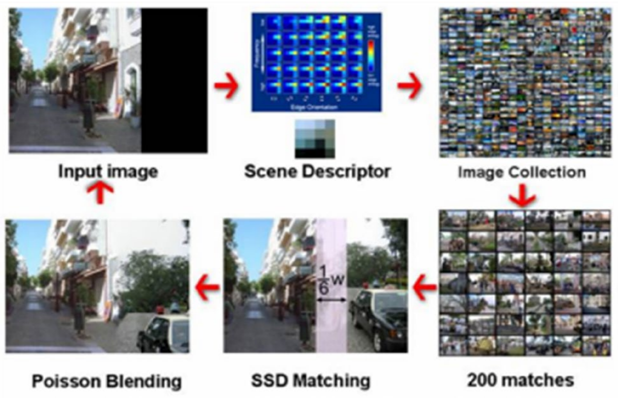


**Fig. 1** Scene completion using millions of photographs [5]

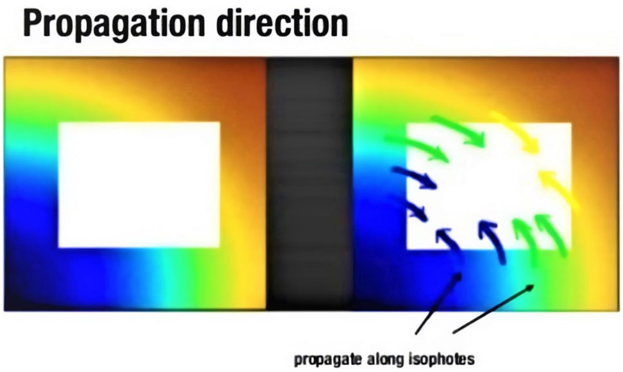


**Fig. 2** Pixel diffusion along isophotes

interpret the results and highlight the advantages as well as the limitations of our approach. Section 6 concludes the paper and gives some perspectives for future work.

## 2 Related work

### 2.1 Image completion (Inpainting)

Image inpainting or completion represents some processes used to reconstruct and retrieve missing or damaged areas in images and photographs. In the art and culture context, this task is assigned to highly skilled conservators and curators that are familiar with the history of the asset to restore. This restoration takes often a considerable amount of time and a lot of financial resources. In computer science and more precisely in computer vision, many image inpainting techniques were proposed to automatically perform the image completion task. Image inpainting techniques can also be used for watermarking and steganography in images and videos [14, 15]. For some time, there were mostly two types of image inpainting algorithms: pixel diffusion-based and patch-based statistical methods (see Figs. 1 and 2). These classical inpainting techniques are often referred to as non-learning image inpainting methods. Pixel-diffusion techniques try to propagate visual information from regions surrounding the missing area (see Fig. 2). In [16] for example, the authors used a technique called “isophote direction field”. However, these methods are known to only be effective in smaller size missing regions and fail drastically in completing larger regions [17]. The second type of non-learning approaches for image inpainting is called patch-based methods and aims at providing an alternative for completing larger missing areas. The principle is to leverage the textures present in the image itself (non-damaged regions) or from another image and then insert the relevant patch into the targeted region (missing area) [7, 11]. The process is performed iteratively and results in very high computational and space complexity. For this specific technique, one of the most notable optimizations which has been proposed is PatchMatch, where Barnes et al. present a faster patch search algorithm [6]. The technique was reported to work well for completing simple patches such as backgrounds. But due to the design of such patch-based approach and given the fact that it relies only on low-level information, it is still underperforming for complex patches [18].

Since the rise of deep learning techniques for image classification [19], super resolution [20, 21] and adversarial training [12], various approaches tackling the image inpainting challenge were proposed. The first approaches used context encoders which are a variation of autoencoders that train an encoder–decoder network on tuples taking a damaged image as input to predict the missing region. The assumption was that a square region is missing in the center of the image. However, since the introduction of generative adversarial networks (GAN) which are known to be among the most powerful generative models [22], several contributions used GANs for image inpainting tasks. In [7], the authors tried to reconstruct the data distribution of the input data using a deep convolutional GAN (DCGAN). They introduced the concept of contextual loss and used a backpropagation on the generator network to retrieve the best patch in the completion. Unfortunately, this backpropagation induces a slowdown in the completion stage.

In [23], the authors extended this concept and used an adversarial loss combining a global and a local discriminator. In [17], the authors tried to improve on the context encoder approach by introducing adversarial training and used two discriminators (global and local) to perform a consistent completion. One of the advantages of this approach is the removal of the assumption that a centered mask is needed for completion. This approach is reported to require a lot of time during training and uses a blending method to smooth out the completion. In [9] and [24], the authors implemented a new type of convolutions to address issues found when using normal convolutions in inpainting networks. The majority of inpainting solutions treat the masked pixels as valid ones which may lead to irregularities and artifacts in the completion. The same authors of [24] proposed in [2] an improved feed-forward generative inpainting network that uses a contextual attention layer and an optimized Wasserstein GAN to improve the training stability and time. In [25], the authors presented a deep fusion network that aims at addressing the problem of blending the generated content into the original image. More recently, in [1], the authors proposed a pluralistic image completion framework that overcomes one of the main issues found in the majority of inpainting networks being the ability to generate multiple plausible assumptions over the missing regions. EdgeConnect presented in [26] is a new deep learning-based technique for image inpainting. It introduces the concept of adversarial edge learning aiming at addressing major issues found in several inpainting frameworks such as the lack of fine details or blurry regions. The authors proposed a two-stage inpainting solution consisting of edge generation and image completion stagesThroughout our analysis, we found that most of the image completion frameworks are affected by the same problem being the difficulty in training and generalization over diversified datasets. Indeed, most of the cultural datasets are very diverse and regroup multiple visual contexts. As a potential solution, we propose a clustering-based framework that combines visual data clustering and image inpainting for cultural image completion and reconstruction.

A previous work published in [13] had for a goal to solve this challenge using deep convolutional generative adversarial networks. However, this approach uses gradient descent at the image reconstruction stage which leads to higher latency. The approach does not yield consistent results, which led us to do more research to improve it.

### 2.2 Image data clustering

As we aim at restraining visual contexts when training image inpainting frameworks, image data clustering is one of the most effective techniques to regroup similarly looking images into clusters. Several image clustering techniques can be used for this task, but through our analysis, we found that deep learning-based solutions considerably outperform classical solutions using handcrafted features. In our work, we compare between three deep learning-based image clustering solutions. The solutions use transfer learning from pre-trained CNNs. The CNN features are extracted from a pre-trained network in the form of vectors from all the training images and they are used to train a clustering model using the k-means algorithm [27]. Other approaches for data clustering such as Gaussian Mixture Models and Expectation Maximum can be used for data clustering, However, in this paper, our main goal is to evaluate the effect of data clustering on regrouping images having similar visual contexts on image inpainting and not on the performance of the clustering. These deeper experiments will be subject of our future research.

## 3 Methodology

Generative models and inpainting techniques are trained to reproduce visual data from a training set. The assumption is that images are samples of a high-dimensional probability distribution. The goal of these models is to learn how to generate samples from this distribution. Unfortunately, for images, we can only collect few samples from such distribution. For image completion, we noticed through our analysis that visual features of similar images can be used to find images that look the same. Also, the fact that generative models try to approximate a model that samples images from a distribution can lead to a harder convergence and the search space becomes larger. Through our state-of-the-art methods for image completion study, we can clearly see that these models cannot be used to complete visual captures from various contexts. This is mainly due to their architecture and the fact that the search space in the completion is restricted to an approximation to the data used in the training. In fact, if this distribution gets larger in dimensionality and size, it becomes very difficult to estimate the likelihood of the original data. In our approach, we mostly try to evaluate the effect of using a divide-and-conquer strategy with visual data clustering on the image completion task. By introducing clustering in the training stage of such models, we considerably limited the size of the original distribution which, regardless of the used completion strategy, always yields better results than when using these completion techniques standalone.

Figure 3 represents the overall architecture of our framework where a clustering step is required before performing the training of the completion components. In the training stage, we start by preprocessing the data and performing the proposed clustering approach which returns the visual clusters. Afterward, an inpainting model is trained for each of the clusters. At a first glance, this seems inefficient, but when we further analyze the training process, it turns out

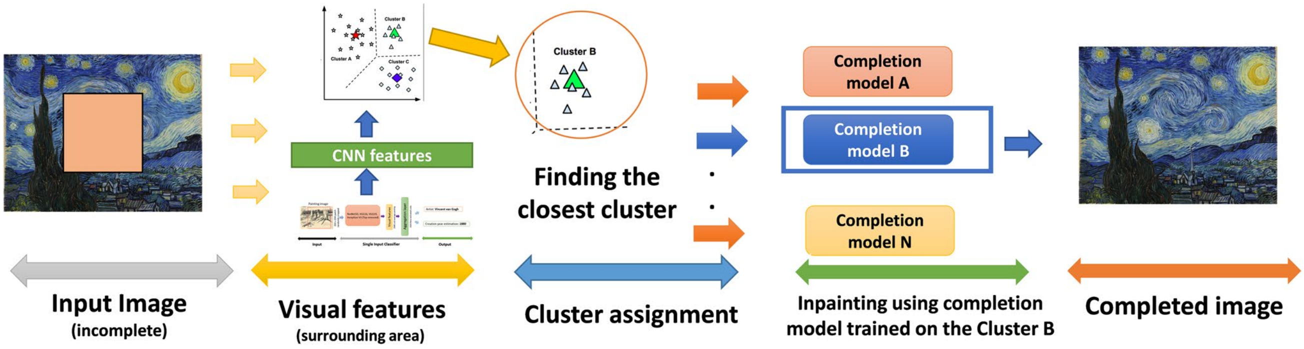


**Fig. 3** Architecture of our image completion framework. We first cluster the dataset using Visual CNN features extracted using a cultural image classification CNN. For each resulting cluster, we train an instance of the reconstruction model and use the clustering model to assign the image to complete to the completion model trained on that cluster

to be more efficient as each of the models have far less data than a general model trained on the whole dataset. In the completion stage, the clustering model is used to select the closest cluster to the damaged image using the remaining regions. Afterward, the completion is assigned to the inpainting model associated with the selected cluster.

In this work, we investigated the best clustering techniques that ensure a low intra-class variation between images of the same cluster while at the same time maintaining a higher inter-cluster distance between different clusters. For visual data clustering, we mostly compare between 3 techniques that use the *k*-means clustering approach with a variation of the visual feature extraction method [28, 29]. In the clustering stage, the only hyperparameter that needs to be fixed empirically is the number of the visual clusters which is a measure that we later explain how it was fixed in our experiments in Sect. 4.2.

Similarly, we also investigate inpainting frameworks that perform good quality completion with their optimal setup while still fail in a diversified context. We train two selected image inpainting frameworks [1, 30] using a cultural dataset with and without applying visual data clustering and compare their results.

In the following, we describe each of the studied clustering methods highlighting their benefits and their drawbacks. We then present the different image completion frameworks we trained.

## 3.1 Visual data clustering stage

Given the fact that we are designing a clustering-based image inpainting framework, a good clustering component needs to be implemented in order to ensure the best high intra-class similarity (cohesive within clusters) and the best low inter-class similarity (distinctive between clusters). We relied on three image clustering techniques in our tests. The three approaches leverage transfer learning with CNN features extracted from the ResNet50 CNN, VGG16 CNN and a CNN developed in a work that was trained on cultural images in [28, 29]. These global visual features once extracted, illustrate a high-dimensional representation of the data in an assigned feature space. A k-means clustering model is then built with an arbitrary number of centroids which depends on the level of the needed similarity in each cluster. In the following, we give details for each of the tested clustering approaches.

### 3.1.1 CNN features using pre-trained VGG16 and ResNet50

Global Features extracted from CNN are considered among the best descriptive visual features in the literature as each convolutional kernel in these networks learns how to capture a particular visual feature from the training data without any interventions from users. The networks we selected in these categories were trained on the ImageNet challenge which consists of the visual classification of millions of images into 1000 categories. These networks (VGG16 [31, 32] and ResNet50 ) were among the best networks for the visual classification task. To extract their features, we used their ImageNet weights and removed their top dense layers. The features we used are the output of the last convolutional layers. We computed the Global CNN features of our training data using the two networks and trained a k-means clustering model over these features. Figure 4 shows the features layer used in the VGG16 network.

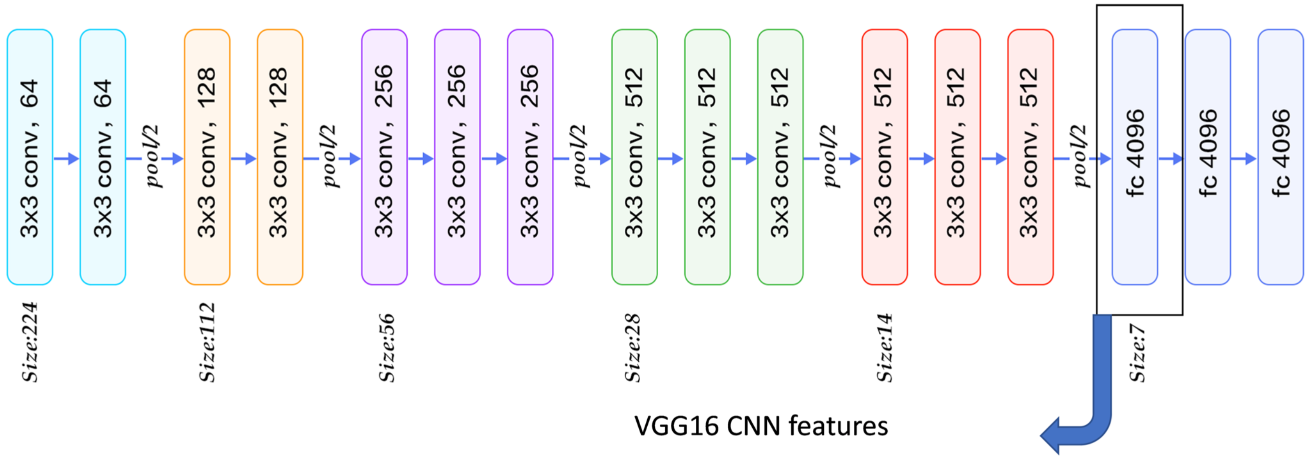


**Fig. 4** VGG16 visual features extraction

### 3.1.2 CNN features from a multitask CNN for cultural image classification

In [28, 29], a multitask approach to the classification of cultural assets was presented. The classification network that was used has the same design as the ResNet50 network except for the fact that it used multiple outputs (for different labels) and the fact that it was trained to classify cultural images. The network achieved very good classification results in contrast to a network trained on a single task due to the use of multitask learning. Indeed, the network was trained to predict multiple labels at once which allowed it to learn more relations between the different features. We find that using a multitask network trained to classify cultural assets as a visual features extractor for cultural assets results in a more accurate features representation and thus a better clustering.

## 3.2 Visual data completion stage

In the following, we present the two selected image completion frameworks used to evaluate our clustering-based inpainting framework. These frameworks were chosen mostly because they yield state-of-the-art completion performance [1, 2].

### 3.2.1 Generative image inpainting with contextual attention (GIICA)

The GIICA inpainting framework [2] was selected as it achieves state-of-the-art completion results in terms of completion quality while still lacking with diversified training data. Its authors improved the inpainting model presented in [11] and aimed at addressing common issues found in CNN-based inpainting frameworks such as boundary artifacts, distortions, and blurry inconsistent inpainting results. Extending on context encoders (CE), the authors presented a two-stage feed-forward generative network with a new contextual attention layer. The first stage uses dilated convolutions and a reconstruction loss. In the second stage, contextual attention is achieved by learning how to capture relevant information from the background of the image to complete. The authors also used WGAN instead of GAN in the local and global discriminators to stabilize the training (see Fig. 5). The main advantages of this framework are its relatively quick training time and high-quality inpainting results while still having difficulties with diversified training data.

### 3.2.2 Pluralistic Image Completion (PIC)

The authors of [1] aimed at addressing a limitation shared by most image completion frameworks being that they only provided a single assumption over the missing region and lack the ability to generate multiple results that fit semantically with the surroundings. The authors claimed that existing methods using conditional variational autoencoders (VAE) have very little variation in terms of the generated assumptions. They thus introduced a probabilistic framework that has a dual pipeline architecture (see Fig. 6). The first pipeline is a VAE reconstructive path and the second pipeline is a generative path that learns the distribution of the missing data while being conditioned on the remaining regions. This distribution is used to generate multiple inpainting assumptions. The authors compared this method against state-of-the-art inpainting frameworks and found that it achieves

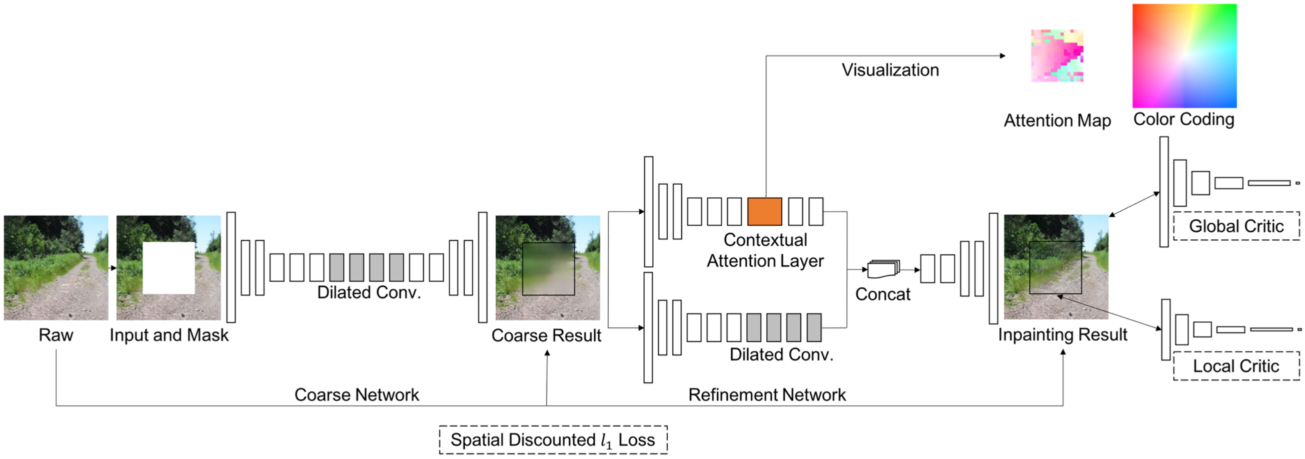


**Fig. 5** Generative Image Inpainting with Contextual Attention Architecture [2]

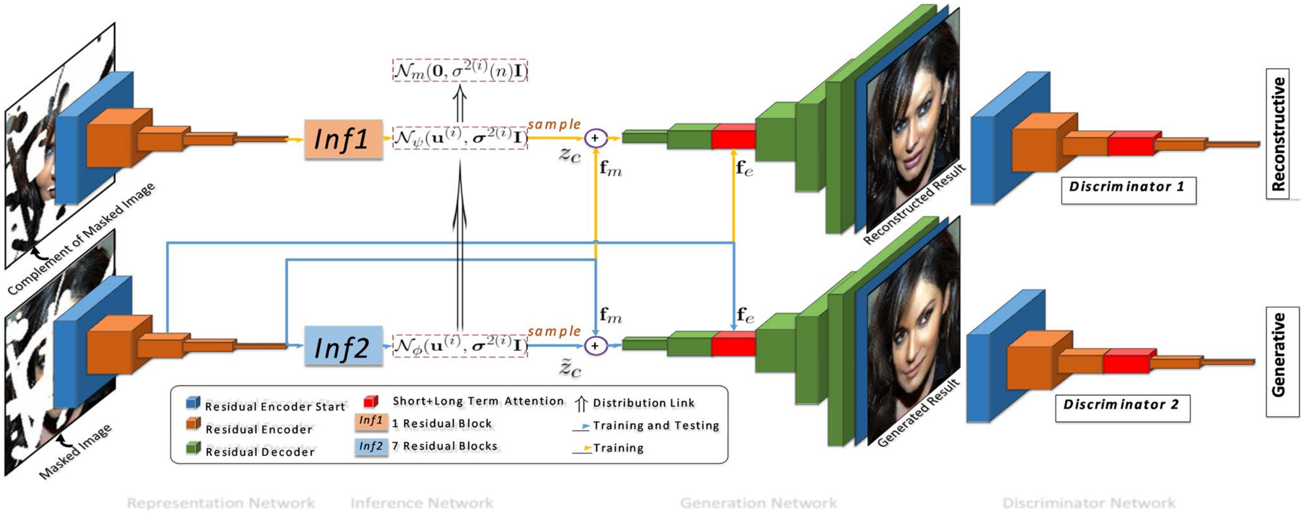


**Fig. 6** Pluralistic Image Completion framework with its two paths architecture. (Reconstructive path on the top and Generative path on the bottom) [1]

better results. The main advantages of this technique are the fact that it can generate multiple plausible assumptions over the missing area, its relatively quick training and its ability to complete larger missing regions while still lacking in diversified contexts.

## 3.3 Data collection and preprocessing

In this section, we start by presenting the datasets we used to test and validate our framework. We describe our experimental setup and the methodology we used to evaluate our clustering-based inpainting framework on the collected cultural datasets.

For our evaluation, we used three datasets collected from various institutions. These datasets contain mostly paintings. The preprocessing step consists in omitting bad and repeated data samples. It is worth noting that the WikiArt dataset is the most well-structured one.

### 3.3.1 The WikiArt Dataset

The WikiArt web gallery [33] hosts a huge collection of artworks from thousands of artists. The data is more than 140 K paintings which are not directly available for public download. The data was collected using a custom Python script written using the *beautifulsoup* library (see Fig. 7).

#### 3.3.2 The Metropolitan Museum (The MET) dataset

The MET museum of New York published half of its collection under the CC 4.0 license and provided a CSV file of the published collection consisting of some metadata and weblinks. We crawled the MET website using custom scripts to reconstruct its collection (see Fig. 8).

#### 3.3.3 The Rijksmuseum Dataset

The Rijksmuseum (Amsterdam), in collaboration with computer vision researchers, opened its collection for the public and a visual classification challenge was set. It is worth noting that the museum has a public API which we used to collect its collection (see Fig. 9).

### 3.4 Experimental setup

For evaluation, we implemented our inpainting framework in Python (Version 3.6). For the visual clustering part, we used the Keras (2.2.0) deep learning library with Tensorflow backend (version 1.13.0) for the implementations of the clustering via transfer learning. For the image completion part, we used both PyTorch (version 1.0.0) and Tensorflow to retrain the two networks on the clusters. For the sake of clarity, we only present the inpainting results for 5 clusters sampled from the 200 we used in our training data. These clusters were chosen to show the diversity of our training data. The number of samples in each cluster is 1000, on average. Training the inpainting frameworks for each cluster takes a long time (6–12 h on average), and using a single machine for the tests is not a realistic scenario. We thus used the Google Colab environment with GPU runtime. Google Colab gives the ability to run Python Jupyter Notebooks on Google Cloud. Each environment allows the user to use an Nvidia Tesla T4 GPU with 16 GB of VRAM. Training the two networks in parallel for 5 clusters took us roughly 3 days. The state of the networks after training was then transferred to a local machine with an Nvidia GTX 1070 GPU for final evaluations. Additionally, to train the inpainting models on the five clusters, we compiled a sub dataset of 1000 images by sampling 5 images from each cluster to ensure maximum visual diversity. This was done to compare the inpainting performance of these models with and without clustering the training datasets. As a result, for each of the two frameworks we trained 6 instances: 5 were each trained on the selected 5 clusters and the 6th one was trained on the mixed dataset.

## 4 Results

### 4.1 Visual data clustering

The visual data clustering techniques we used were studied originally due to their performance. But through our experiments, we saw that the best visual clustering technique that perfectly separates our training data is the one based on the

**Fig. 7** Selected artwork from WikiArt [33]

**Fig. 8** Selected fine-art paintings from the MET [34]

**Fig. 9** Some artwork from the Rijksmuseum [35]

multitask classification framework for cultural images [29]. This is mainly due to the refined filters and the fact that the network was trained in a multitask fashion.

### 4.2 Image inpainting

For the image inpainting models training, the GIICA framework converged rather quickly after 30 epochs of training over the 5 data clusters but did not achieve the same level of convergence using the mixed dataset (after 50 epochs). The convergence of the networks was measured using the loss metric. For the PIC framework, it is clear that it requires more training iterations to achieve convergence (mostly after 150 epochs), but for the sake of consistency and to test it on optimal conditions, we let the network train for 300 epochs (default setting) for the 5 clusters and the mixed dataset. The training of this network for 300 epochs takes roughly 11 h on the Nvidia Tesla T4 GPU. To evaluate the visual results of our completion, we sampled images from each cluster and tasked the frameworks to perform the completion over a centered square mask. Table 1 outlines the inpainting results for the two approaches we studied. We sampled five images from each of the clusters and tasked the models to complete a missing central region (centered mask). Table 1 reports the inpainting results (PSNR value) using the two models in the two scenarios: clustered and mixed training dataset. The first value under each image is the PSNR value between the completed image (completed region + surroundings) and the ground truth. The second value is the PSNR value between the centered completed region and its corresponding region in the ground truth image.

Additionally, since evaluating the quality of an image, especially for image inpainting tasks, is a subjective problem, as given in [36], we used an audience of 40 persons and gave them a survey (see "Appendix" at the end of the paper) to evaluate the same results on a quality scale from 1 to 5 which is often used in visual assessment surveys [37]. This methodology is used to ensure having a standardized way in assessing the quality of the reconstruction. The Mean Opinion Score (MOS) results are summarized by the box plot in Fig. 10. Methods 1 and 3 are, respectively, for the GIIAC+clustering and PIC+clustering. Method 2 and Method 4 are, respectively, for GIIAC and PIC without clustering. The details related to the audience survey we used are provided in the manuscript "Appendix".

The first observation we can get from these audience results (MOS) is that for the clustering based inpainting, the median value is close to 5 and for the mixed model, it is close to 2. Additionally, we notice that the spread of values for the clustered tests is less significant (between 4 and 5) compared to the non-clustered tests where we can clearly see that the spread of scores is much wider (between 1 and 3).

Following the training, and to showcase the improvements in the reconstruction step, we compared the execution time of the inpainting process (reconstruction) of our clustering-based framework against a previous work based on the semantic image inpainting approach presented in [13]. Indeed, the improvement both in time and complexity is high as the semantic inpainting approach uses the gradient descent method to generate multiple assumptions (thousands) in order to produce a result that is visually close to the damaged sample at hand. Table 2 describes the evaluation time of our framework and the semantic inpainting approach.

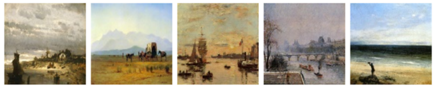
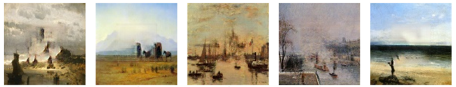
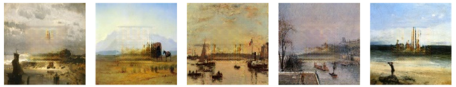
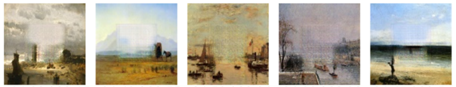
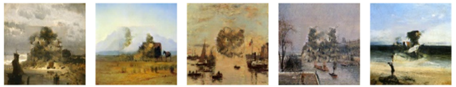
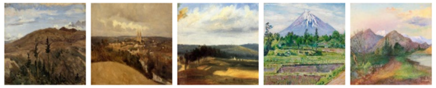
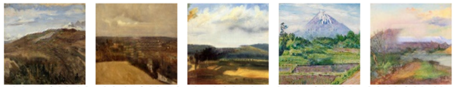
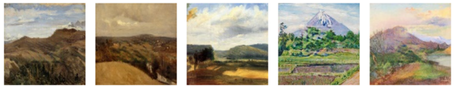
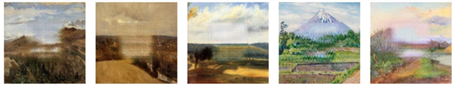
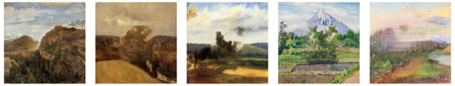
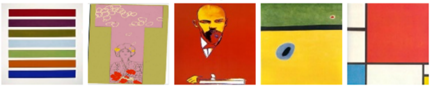
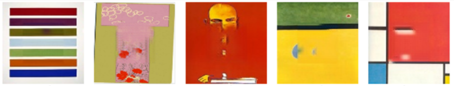
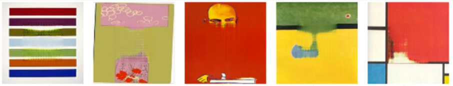
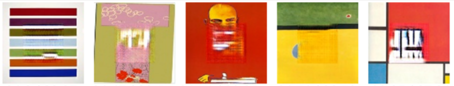
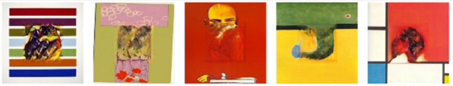
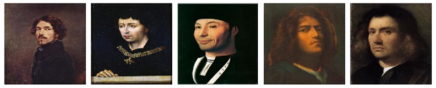
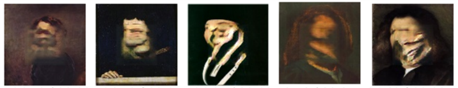
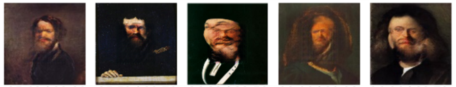
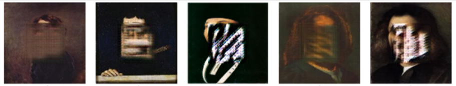
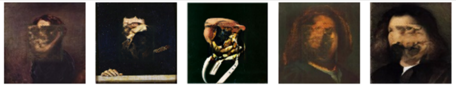
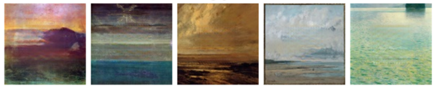
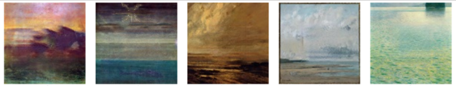
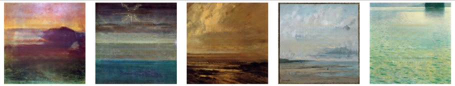
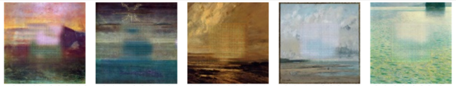
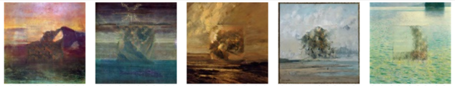
One can clearly see that the semantic inpainting approach is less performant for low-delay inpainting applications as it requires heavy processing (more than 5000 iterations) at the reconstruction stage.We find that both GIIAC and PIC can be efficiently used for low-delay applications hence their selection in our investigation.

## 5 Discussion

Through this work, we aim at reconstructing missing visual information in cultural artwork. Damaged artwork quickly loses its value and classical curation techniques are considered not effective and not suitable for low-delay applications. After our literature review, we found that deep learning-based image inpainting techniques can be used for cultural image reconstruction as these techniques are among the most powerful methods for such tasks and many optimizations were proposed such as the use of adversarial training. Following our analysis of these techniques, we noticed that they tackle the challenge from different viewpoints (using context encoders, VAEs, and GANs), but they all share the same concept of learning how to sample images from a data distribution. This data distribution is, in fact, the distribution of the training data. We also noticed that the nature and the diversity of the training data play a primordial role in the effectiveness of these approaches. If the used data has various visual concepts, the diversity of concepts will be tightly tied to the distribution's dimensionality. Additionally, some of the existing frameworks [13] require heavy processing at the reconstruction stage which is not always desirable for low-delay inpainting applications.

As a result, and through our intensive experiments, we saw that reducing the visual contexts of the training data of

**Table 1** Our framework inpainting results and PSNR values on the sampled 5 clusters and the mixed dataset over the two tested image inpainting frameworks

| Ground Truth | Training | GIIAC | PIC |
|---|---|---|---|
| C1<br> | Trained on C1 | <br>**34.32 / 28.38**<br>**35.61 / 29.68**<br>**34.61 / 28.66**<br>34.71 / 28.76<br>**34.93 / 28.98** | <br>**34.51 / 28.56**<br>**35.27 / 29.3**<br>**34.96 / 28.99**<br>**35.06 / 29.11**<br>**34.89 / 28.92** |
| | Trained on Mix | <br>34.08 / 28.15<br>34.49 / 28.55<br>34.58 / 28.64<br>**34.86 / 28.93**<br>34.45 / 28.5 | <br>34.38 / 28.44<br>35.02 / 29.07<br>34.29 / 28.33<br>34.45 / 28.5<br>34.57 / 28.62 |
| C2<br> | Trained on C2 | <br>**34.57 / 28.66**<br>**34.45 / 28.51**<br>**35.0 / 29.07**<br>**34.17 / 28.22**<br>**34.43 / 28.46** | <br>**34.64 / 28.72**<br>**34.65 / 28.71**<br>34.73 / 28.78<br>**34.12 / 28.18**<br>34.42 / 28.45 |
| | Trained on Mix | <br>34.39 / 28.47<br>34.3 / 28.37<br>34.4 / 28.46<br>33.98 / 28.02<br>34.22 / 28.25 | <br>34.47 / 28.55<br>34.4 / 28.46<br>**34.8 / 28.86**<br>34.02 / 28.07<br>**34.48 / 28.52** |
| C3<br> | Trained on C3 | <br>**36.67 / 30.74**<br>**36.81 / 30.84**<br>**36.36 / 30.49**<br>**36.61 / 30.7**<br>**37.56 / 31.67** | <br>**35.35 / 29.4**<br>**34.57 / 28.59**<br>**37.78 / 31.95**<br>**36.01 / 30.07**<br>**37.47 / 31.58** |
| | Trained on Mix | <br>34.9 / 28.97<br>34.47 / 28.5<br>34.74 / 28.85<br>34.71 / 28.79<br>33.85 / 27.91 | <br>34.19 / 28.23<br>34.26 / 28.28<br>35.39 / 29.5<br>34.54 / 28.6<br>34.45 / 28.5 |
| C4<br> | Trained on C4 | <br>34.59 / 28.69<br>**34.72 / 28.85**<br>34.05 / 28.18<br>35.2 / 29.35<br>34.**19 / 28.33** | <br>**35.16 / 29.29**<br>**35.59 / 29.76**<br>**34.39 / 28.51**<br>**36.14 / 30.32**<br>**34.64 / 28.79** |
| | Trained on Mix | <br>**34.75 / 28.87**<br>34.53 / 28.68<br>**34.07 / 28.21**<br>**35.78 / 29.95**<br>34.18 / 28.33 | <br>34.84 / 28.95<br>35.57 / 29.73<br>34.33 / 28.45<br>35.86 / 30.02<br>34.55 / 28.7 |
| C5<br> | Trained on C5 | <br>**34.85 / 28.91**<br>**35.12 / 29.21**<br>**35.6 / 29.71**<br>**35.41 / 29.43**<br>**35.31 / 29.34** | <br>**35.39 / 29.45**<br>**34.87 / 28.95**<br>**35.51 / 29.62**<br>**35.46 / 29.49**<br>**35.3 / 29.34** |
| | Trained on Mix | <br>34.43 / 28.49<br>34.5 / 28.58<br>34.43 / 28.51<br>34.47 / 28.51<br>34.86 / 28.9 | <br>34.56 / 28.61<br>34.24 / 28.31<br>34.56 / 28.64<br>34.45 / 28.48<br>34.89 / 28.92 |

The first value is the PSNR of the whole image. The second value is the PSNR of the recovered region only most of the inpainting techniques positively impacts their performance in terms of visual quality. We thus designed our framework using a divide-and-conquer strategy and a method to effectively separate the training data. In fact, instead of training a single instance of the completion model, we cluster the training data and train a completion model instance for each resulting cluster. Our first observation is that this clustering step reduces the training difficulty encountered in completion frameworks making them converge faster. In our work, we relied on transfer learning from pre-trained CNNs for image completion to extract global visual features. These features were then used to train a clustering model.

For our tests, the data we had was more than 150 K samples, and thus we set the number of centroids (K for the k-means algorithm) empirically to 200. This resulted in clusters that regroup artwork looking very similar. We then used these clusters to train instances of the two inpainting frameworks. As a proof of concept, we only trained the models on 5 clusters and added an instance trained on mixed data (without clustering) for each of the two frameworks.

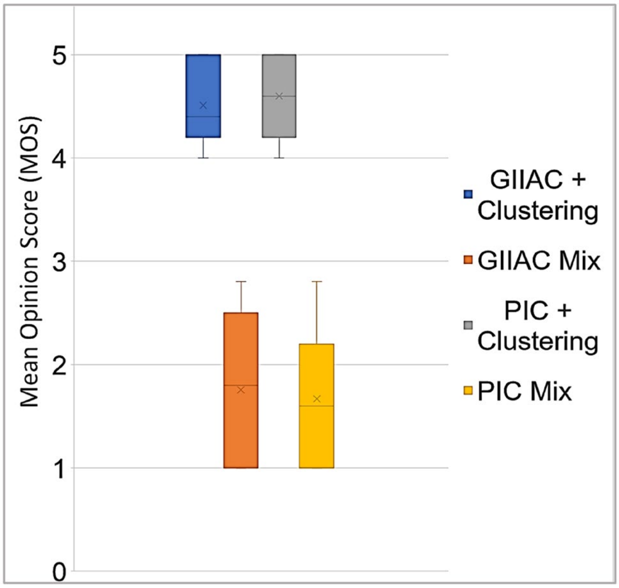


**Fig. 10** Box and Whisker plot of the survey results. The Y-axis represents the Mean Opinion Score (MOS) of the four experiments

The results we reported in Table 1 show that the completion based on the clustering procedure significantly improves the completion operation compared to the data mixture. These results were achieved using the same setup for both scenarios which consists of allowing the models in each case to achieve their best convergence (fair training between the two approaches). As mentioned previously, the observed increase in completion quality and training performance validates our claim which states that restraining the visual contexts boosts the performance of generative models and thus image inpainting frameworks. Notwithstanding the advantages of our approach, some limitations could be also observed as we can see in some cases (Cluster 4 for example). However, in this case, it is worth noting that neither of the clustering-based nor the traditional model could give a plausible result. This is mainly due to the difficulty of the completion task especially using a centered mask which is reported often to be a tough challenge and a “non-realistic” scenario.

Evaluating hallucinations generated by a machine learning model is a subjective task. Using numerical methods such as PSNR, SSIM, MSSSIM is usually encountered in the literature [18]. However, if the hallucination given by the inpainting model is visually different, these metrics report a bad result despite the fact that the completion may be of a very good quality. Suppose the scenario where an image of landscape next to a lake has a central region over the lake missing. An assumption of a continuation of the lake is close to the ground truth numerically. However, another assumption of a boat in the middle of the lake is also sound, but numerically, it is different than the ground truth. This in fact can be observed in our results for the completion of cluster C1 (see Fig. 11). Each of the GIIAC and PIC methods using our clustering approach yielded different result over the missing region. GIIAC yielded a continuation of the lake, which is close to the ground truth (high PSNR). PIC returned a boat-like shape, which is visually good but had a lower PSNR compared to GIIAC because the result was different from the ground truth.

To address this issue and include a subjective evaluation of our results, we used a visual quality assessment survey sent to an audience of 40 to evaluate the quality of the results. Following the analysis of the audience results, it was also confirmed that clustering has an additional benefit over

**Table 2** Reconstruction time and number of iterations for our framework and the semantic inpainting approach

| Model | Clustering + PIC | Clustering + GIICA | Semantic inpainting [13] |
|---|---|---|---|
| Average execution time | 0.8 s | 0.82 s | 20–30 mins |
| Iterations (No. of assumptions) | 1–100 | 1 | > 5000 |

the traditional training, which validates the performance of our clustering-based inpainting approach.

Still, the completion framework presented in this paper takes relatively more time to be trained than other solutions as many steps are required (clustering + completion models). However, since the complexity in the reconstruction stage is what matters the most, we find that our new framework is suitable for low-delay applications as no further processing of the input data is required at the reconstruction stage. This can be seen in Table 2 where the number of iterations for our framework is more or less just in single digits, whereas it is in thousands for the semantic inpainting approach.

## 6 Conclusion

Deep learning-based image completion and reconstruction frameworks can be used as a computer-based alternative for cultural artwork curation. However, due to the diversity of visual contexts found in cultural datasets, these frameworks become difficult to train as they mostly try to sample data from a high-dimensional probability distribution of images. Adding more visual contexts increases the dimensionality of this distribution and makes these networks training hard. Moreover, these frameworks require heavy processing at the reconstruction stage which makes them not suitable for low-delay applications. In this paper, we proposed a low-delay clustering-based cultural image reconstruction approach which combines supervised and unsupervised learning that to improve state-of-the-art image completion and reconstruction techniques. With visual data clustering, we demonstrated that better inpainting can be achieved using mixed training datasets. For the experiments, we compared the performance of our approach with two state-of-the-art image inpainting frameworks in the cultural image inpainting task. We also evaluated the execution time of our approach and found that it is suitable for low-delay applications as no heavy processing is required at the reconstruction stage. Additionally, since evaluating the quality of an image is a subjective task, we used a surveyed audience to evaluate the results of our approach compared to standard image inpainting. This evaluation demonstrated the performance of our approach from a subjective point of view. Overall, the obtained results are encouraging and show that our clustering-based approach improves significantly the output quality of inpainting frameworks and is suitable for low-delay applications. In the future, we aim at integrating this approach into a custom-designed image reconstruction solution that can be used for other types of cultural assets such as pottery, ceramics, and statues.

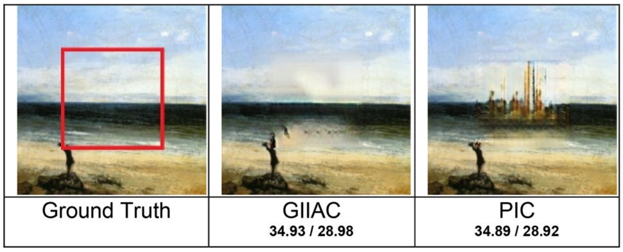


**Fig. 11** The Seaside at Palavas, by Gustave Courbet. Values represent the PSNR of the whole image and the PSNR of the completed part only

**Acknowledgements** This publication was made possible by NPRP grant 9-181-1-036 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors (www.ceproqha.qa). The authors would also like to thank the Museum of Islamic Art (MIA), the MIA Multimedia team, Mr. Marc Pelletreau, the Art Curators and the management staff of the Museum of Islamic Art, Doha Qatar for their help and contribution in the data acquisition.

## Appendix

User Evaluation Forms The evaluation form can be accessed following this link: https://forms.gle/gQATiv4HeJhRWKLj9

The following pages show part of the evaluation forms:

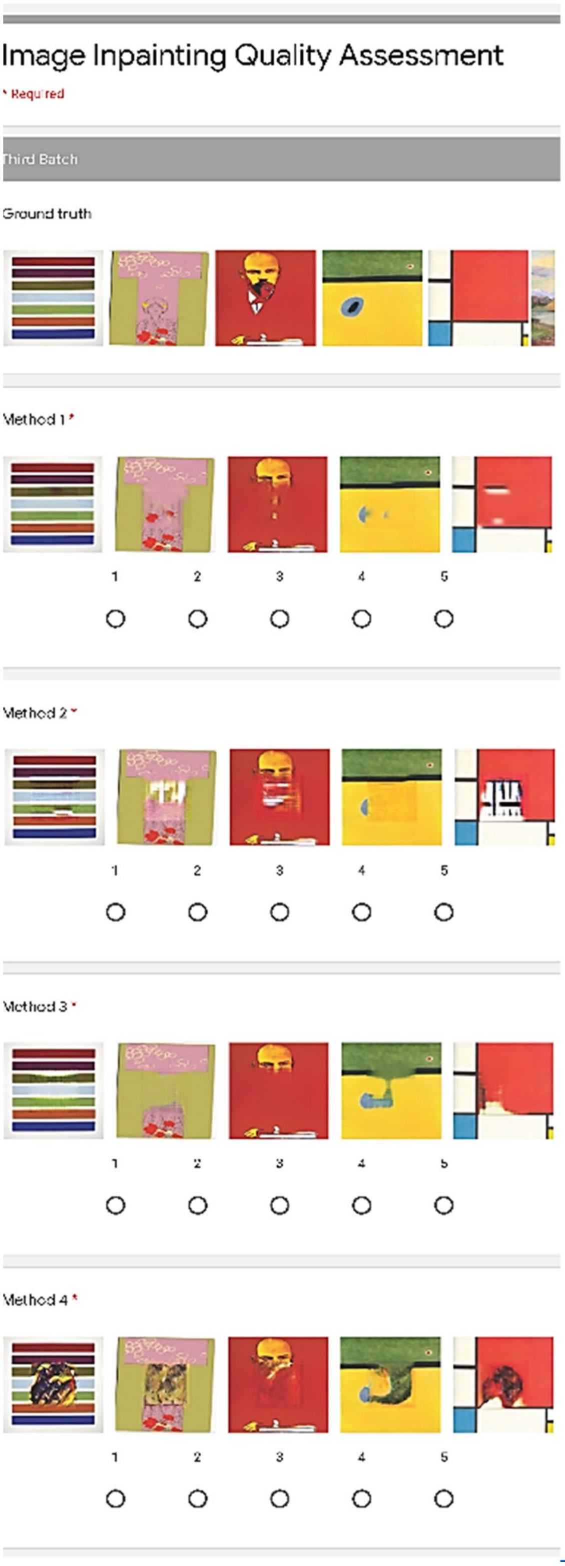


Image Inpainting Quality Assessment

* Required

Third Batch

Ground truth

Method 1 *

1 2 3 4 5

○ ○ ○ ○ ○

Method 2 *

1 2 3 4 5

○ ○ ○ ○ ○

Method 3 *

1 2 3 4 5

○ ○ ○ ○ ○

Method 4 *

1 2 3 4 5

○ ○ ○ ○ ○

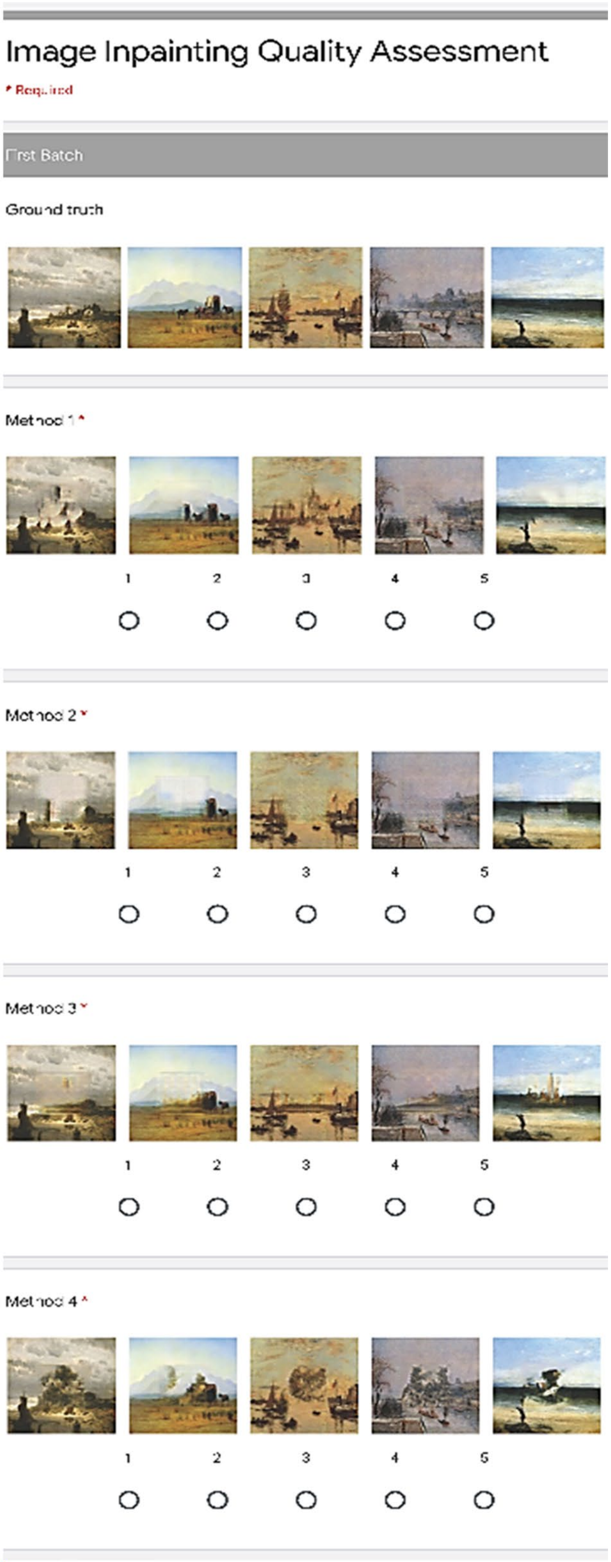


Image Inpainting Quality Assessment

* Required

First Batch

Ground truth

Method 1 *

1 2 3 4 5

○ ○ ○ ○ ○

Method 2 *

1 2 3 4 5

○ ○ ○ ○ ○

Method 3 *

1 2 3 4 5

○ ○ ○ ○ ○

Method 4 *

1 2 3 4 5

○ ○ ○ ○ ○

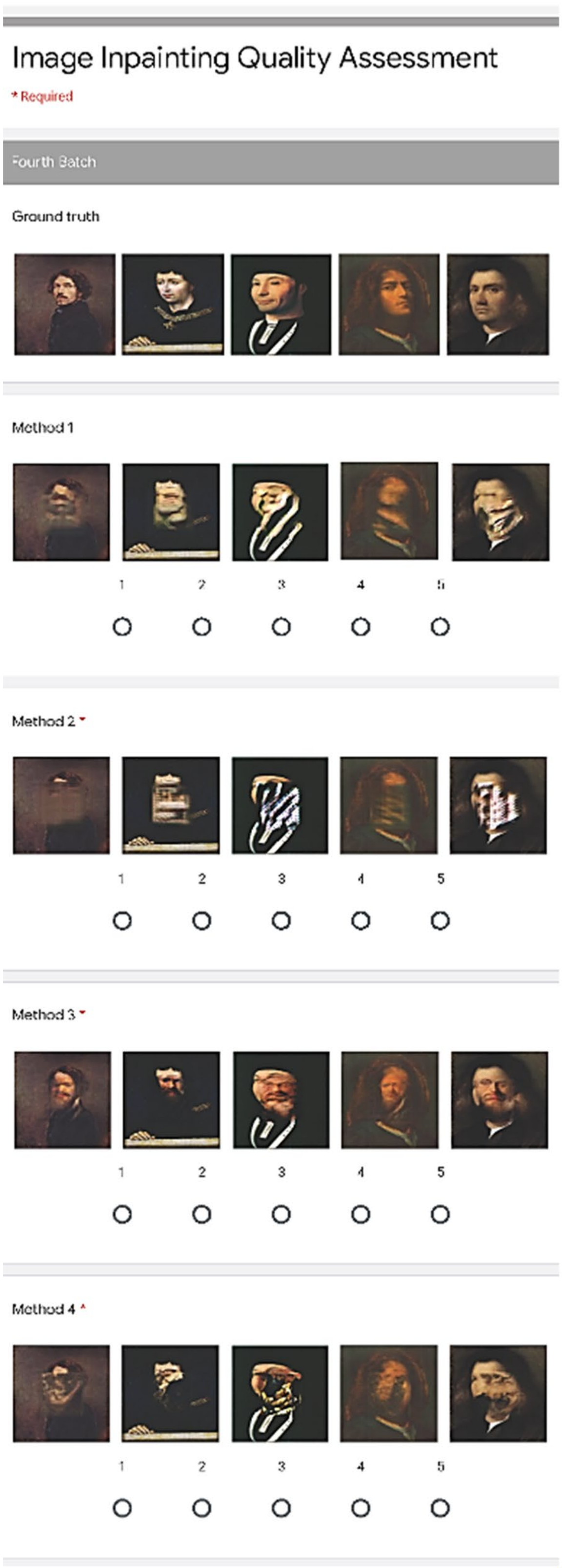


Image Inpainting Quality Assessment

* Required

Fourth Batch

Ground truth

Method 1

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

Method 2 *

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

Method 3 *

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

Method 4 *

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

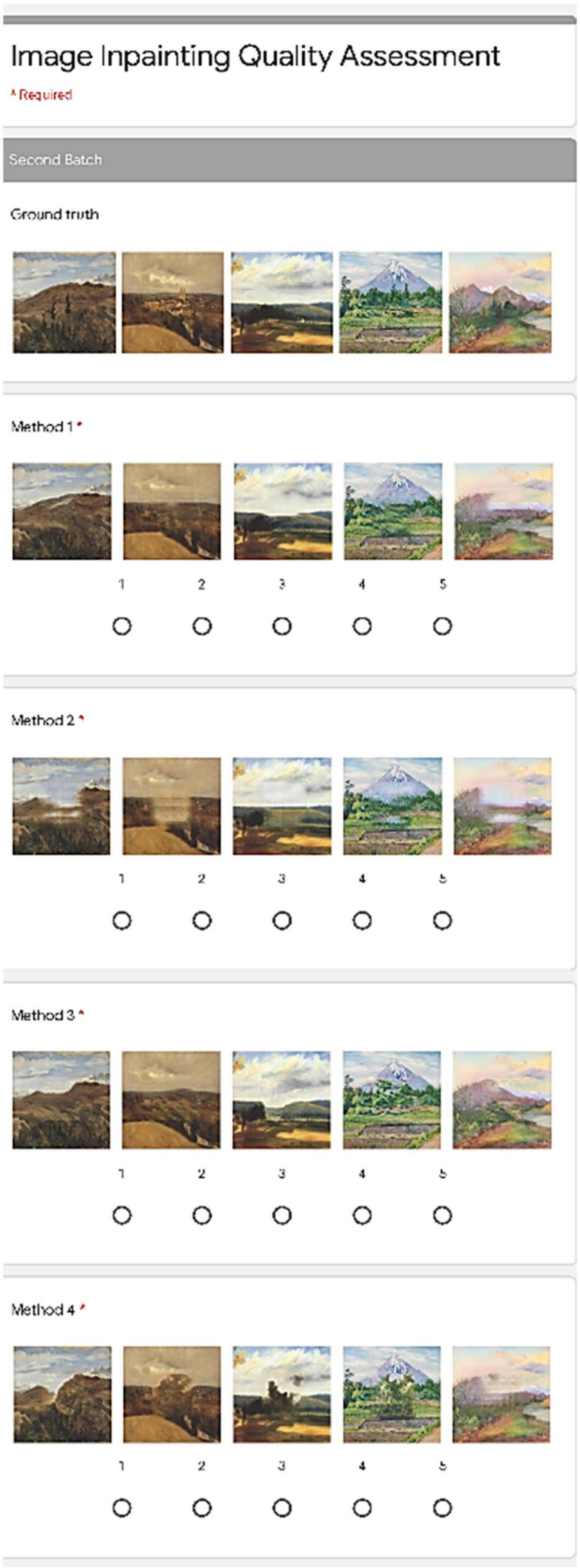


Image Inpainting Quality Assessment

* Required

Second Batch

Ground truth

Method 1 *

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

Method 2 *

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

Method 3 *

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

Method 4 *

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

## References

1. Zheng, C., Cham, T.-J., Cai, J.: Pluralistic image completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1438–1447 (2019).
2. Yu,J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5505–5514 (2018).
3. Ashikhmin, M.: Synthesizing natural textures. SI3D **1**, 217–226 (2001)
4. Ballester,C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels (2000).
5. Hays,J., Efros, A.A.: Scene completion using millions of photographs. In: ACM Transactions on Graphics (TOG), 2007, vol. 26(3), p. 4. ACM (2007).
6. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graphics (ToG) **28**(3), 24 (2009)
7. Yeh,R.A., Chen, C., Yian Lim, T., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5485–5493 (2017).
8. Demir,U., Unal, G.: Patch-based image inpainting with generative adversarial networks. arXiv preprint arXiv:1803.07422 (2018).
9. Liu,G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 85–100 (2018).
10. Xiong,W., et al. Foreground-aware image inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5840–5848 (2019).
11. Pathak,D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544 (2016).
12. Goodfellow,I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680 (2014).
13. Jboor,N.H., Belhi, A., Al-Ali, A.K., Bouras, A., Jaoua, A.: Towards an inpainting framework for visual cultural heritage. In: 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019, pp. 602–607. IEEE (2019).
14. Zhou, Q., Yao, H., Cao, F., Hu, Y.-C.: Efficient image compression based on side match vector quantization and digital inpainting. J. Real-Time Image Proc. **16**(3), 799–810 (2019)
15. Zhang, W., Kong, P., Yao, H., Hu, Y.-C., Cao, F.: Real-time reversible data hiding in encrypted images based on hybrid embedding mechanism. J. Real-Time Image Proc. **16**(3), 697–708 (2019)
16. Bertalmio,M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques, 2000, pp. 417–424: ACM Press/Addison-Wesley Publishing Co (2000).
17. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graphics (TOG) **36**(4), 107 (2017)
18. Elharrouss,O., Almaadeed, N., Al-Maadeed, S., Akbari, Y.: Image inpainting: a review. In: Neural Processing Letters, pp. 1–22 (2019).
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105 (2012).
20. Shamsolmoali, P., Zhang, J., Yang, J.: Image super resolution by dilated dense progressive network. Image Vis. Comput. **88**, 9–18 (2019)
21. Shamsolmoali, P., Li, X., Wang, R.: Single image resolution enhancement by efficient dilated densely connected residual network. Signal Process Image Commun **79**, 13–23 (2019)
22. Shamsolmoali, P., Zareapoor, M., Wang, R., Jain, D.K., Yang, J.: G-GANISR: gradual generative adversarial network for image super resolution. Neurocomputing **366**, 140–153 (2019)
23. Li,Y., Liu, S., Yang, J., Yang, M.-H.: Generative face completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3911–3919 (2017).
24. Yu,J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S.: Free-form image inpainting with gated convolution. arXiv preprint arXiv:1806.03589 (2018).
25. Hong,X., Xiong, P., Ji, R., Fan, H.: Deep fusion network for image completion. arXiv preprint arXiv:1904.08060 (2019).
26. Nazeri,K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019).
27. Jain, A.K.: Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. **31**(8), 651–666 (2010)
28. Belhi,A., Bouras, A., Foufou, S.: Towards a hierarchical multitask classification framework for cultural heritage. In 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), 2018, pp. 1–7. IEEE (2018).
29. Belhi, A., Bouras, A., Foufou, S.: Leveraging known data for missing label prediction in cultural heritage context. Appl. Sci. **8**(10), 1768 (2018)
30. Yu,J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T. S.: Generative image inpainting with contextual attention. arXiv preprintarXiv:1801.07892 (2018).
31. Simonyan,K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
32. He,K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778 (2016).
33. WikiArt.org.: WikiArt.org—visual art encyclopedia (31–01–2019). Available: https://www.wikiart.org/. Accessed on 29 March 2020.
34. MET,T.: The Metropolitan Museum of Art (2019, 31–01–2019). Available: https://www.metmuseum.org/. Accessed on 29 March 2020.
35. Mensink,T., Van Gemert, J.: The rijksmuseum challenge: museum-centered visual recognition. In: Proceedings of International Conference on Multimedia Retrieval, 2014, p. 451 (2014)
36. Tiefenbacher,P., Bogischef, V., Merget, D., & Rigoll, G.: Subjective and objective evaluation of image inpainting quality. In: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 447–451. IEEE (2015).
37. Bt, R. I.-R.: Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union (2002).

**Abdelhak Belhi** received his software engineering and information processing M.Sc in 2016 from the University of Boumerdes - Algeria. He is currently working as a senior research assistant at Qatar University. Abdelhak is also a Ph.D. candidate in computer science at the DISP laboratory, University of Lyon - France. His current research interests include image processing, data mining, machine learning, blockchain, and cybersecurity.

**Dr. Abdulaziz Khali Al-Ali** is currently an assistant professor in the Computer Science and Engineering department, College of Engineering, in Qatar University. He has received his BS and MS Degrees in Computer Engineering from the University of Miami, Florida, USA in 2008 and 2011 respectively. Following that, he earned his Ph.D. in Machine Learning in 2016 from the University of Miami. His current research interests include Machine Learning, Artificial Intelligence and Data Mining. Dr. Abdulaziz is an awardee of several research grants, among which are the National Priorities Research Program from Qatar National Research Fund.

**Prof. Abdelaziz Bouras** is professor in the Computer Science and Engineering Department of the College of Engineering at Qatar University. He is also managing the Pre-Award Department at the Office of Research Support of the university. His research focuses on software lifecycle engineering and data management. His current work deals with the application of AI and deep learning approaches for prediction and decision making. He managed several international projects in this area and published many research papers in referred journals/conferences in collaboration with his team. He also edited several books and two international journals (IJPD, IJPLM) and is currently chairing the IFIP International Federation of Information Processing WG5.1, which holds a yearly international conference on specific enterprise lifecycle information systems (PLM) with a yearly edited book with Springer.

**Prof. Sebti Foufou** holds the position of professor of computer science at the University of Burgundy, Dijon, France, since September 1998. He is currently visiting professor at the Computer Science Department at New York University of Abu Dhabi in UAE. Prior to that, he was with the Qatar University, Qatar, and the NIST, Maryland, USA. His earlier research interests include geometric modeling and image processing for face recognition. He also worked on network topologies to improve data center interconnections for green computing and quality of service. His current research activities focus on data models and algorithms for product lifecycle management in smart machining systems and digital cultural heritage. Machine learning and IA tools are applied to support product data processing and decision making in these two domains.

**Dr. Xi Yu** received his Ph.D. in the DISP (Decision & Information Sciences for Production Systems) laboratory of University Lyon 2. Currently, He is Professor in ChengDu University, Sichan province, China. His research interests are in the areas of LCA, decision-making methodology, machine learning and data analysis. He has published more than 30 research papers (including 5 scientific papers indexed by JCR (SCI)) in these fields. He is the master tutor in Chengdu University and doctoral supervisor in Chiang Mai University, Thailand.

**Dr. Haiqing Zhang** received her Ph.D. in the DISP (Decision & Information Sciences for Production Systems) laboratory of University Lyon 2. Currently, she is an associate researcher in Chengdu University of Information Technology. Her research interests are in the areas of PLM maturity models, decision-making methodology, fuzzy/rough mathematics and data mining. She has published more than 30 research papers in this field. She also holds one National Natural Science Fund, one project supported by Science and Technology of Sichuan Province, and one international project supported by ERASMUS. She also joined five scientific projects in provincial and national levels as a main researcher. As a tutor, she has guided students to win provincial and national awards for several times. She is also the master tutor in Chengdu University of Information Technology and Doctoral supervisor in Chiang Mai University.