Draw with Me: Human-in-the-Loop for Image Restoration

Thomas Weber*
Heinrich Hußmann
thomas.weber@ifi.lmu.de
hussmann@ifi.lmu.de
Ludwig-Maximilians-University
Munich, BY, Germany

Zhiwei Han*
Stefan Matthes
Yuanting Liu
han@fortiss.org
matthes@fortiss.org
liu@fortiss.org
fortiss GmbH
Munich, BY, Germany

ABSTRACT

The purpose of image restoration is to recover the original state of damaged images. To overcome the disadvantages of the traditional, manual image restoration process, like the high time consumption and required domain knowledge, automatic inpainting methods have been developed. These methods, however, can have limitations for complex images and may require a lot of input data. To mitigate those, we present "*interactive Deep Image Prior*", a combination of manual and automated, Deep-Image-Prior-based restoration in the form of an interactive process with the human in the loop. In this process a human can iteratively embed knowledge to provide guidance and control for the automated inpainting process. For this purpose, we extended Deep Image Prior with a user interface which we subsequently analyzed in a user study. Our key question is whether the interactivity increases the restoration quality subjectively and objectively. Secondarily, we were also interested in how such a collaborative system is perceived by users.

Our evaluation shows that, even with very little human guidance, our interactive approach has a restoration performance on par or superior to other methods. Meanwhile, very positive results of our user study suggest that learning systems with the human-in-the-loop positively contribute to user satisfaction. We therefore conclude that an interactive, cooperative approach is a viable option for image restoration and potentially other ML tasks where human knowledge can be a correcting or guiding influence.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing systems and tools**; • **Computing methodologies** → *Neural networks*; • **Applied computing** → Fine arts.

KEYWORDS

Interactive Machine Learning, Image Restoration, Image Prior

*Both authors contributed equally to this research.

IUI '20, March 17–20, 2020, Cagliari, Italy

ACM ISBN 978-1-4503-7118-6/20/03.
https://doi.org/10.1145/3377325.3377509

ACM Reference Format:
Thomas Weber, Heinrich Hußmann, Zhiwei Han, Stefan Matthes, and Yuanting Liu. 2020. Draw with Me: Human-in-the-Loop for Image Restoration. In *25th International Conference on Intelligent User Interfaces (IUI '20), March 17–20, 2020, Cagliari, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3377325.3377509

1 INTRODUCTION

Image inpainting is a process that fills missing sections in images, such that the restored images are visually plausible. It can be applied to a variety of real-world applications such as removing unwanted objects in images or image restoration.

In order to distinguish from the general image inpainting tasks, we consider image restoration of damaged or corrupted art works in this paper.

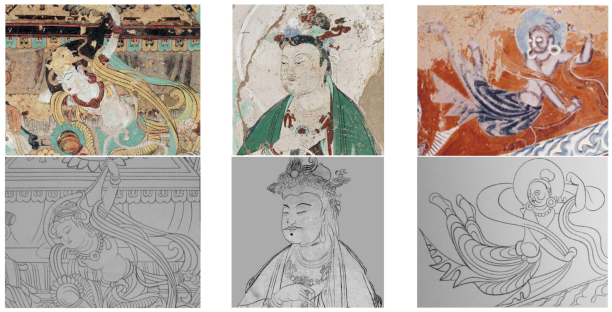

Figure 1: Manually restored murals from Mogao Grotto. Images in the first row are the damaged murals and the images in the bottom row are their corresponding line drawings by experts. Image from [24].

A typical scenario for image restoration is heritage protection. The Dunhuang grottoes dataset [31] of damaged murals from the Mogao Grottoes (see Fig. 1) which we use throughout this work is a popular example for both heritage protection and image restoration [24]. These murals were created by ancient artists between the 4th and 14th centuries. The majority of discovered artifacts in those murals are damaged in some way and continue to deteriorate making restoration essential for preserving this cultural artifact. Traditionally, the restoration requires a professional to paint manually, which requires much experience and effort. While this may remain necessary for the physical artifacts, the digitization of such murals can

be a helpful support. The ability to quickly and reliably restore digital copies of historical artifacts can for example help to determine whether a particular restoration seems plausible without having to work on the valuable original artifact. Additionally, creating a digital representation of the murals ensures that they are available even with the progression deterioration. In recent years, numerous automated frameworks have been proposed for this digital image processing. The inpaiting frameworks proposed by prior works can be categorized into two main classes: exemplar-based [4, 21, 26] and learning-based methods [17, 18, 27]. Those frameworks offer a digital restoration process which showed decent results for many image inpainting tasks while being significantly less time-intensive. However, exemplar-based methods have trouble in recovering complex images, since they only copy existing patches from the same image. And while Deep Learning (DL) works well when trained on a large dataset, DL-based approaches severely suffer from overfitting when only a small training set is available. The fact that such datasets are rarely available prevents learning-based methods from being adopted into many domains.

Furthermore, the restoration performance usually degrades when the corrupted sections become dense or large. Due to the lack of semantic information in large corrupted areas, restored images can be filled with artifacts like inconsistent texture or monotone color. Nevertheless, when missing image features are obvious from semantic but not structural context, humans can easily deduce these missing features, while many algorithms may fail. The deep body of knowledge a human has would allow such inference but is currently unavailable to the machine. A mechanism to harness human knowledge in Machine Learning (ML) is still missing. Such knowledge-based enhancement can make the restoration more robust than when only learned by algorithms alone. To incorporate human knowledge in image restoration and improve the restoration quality, we present a collaborative, interactive image restoration system which enables humans to iteratively guide and correct an automated restoration process, embedding their knowledge into the process.

Directly learning the pixel-level statistics from a small training set without severe over-fitting is unfortunately not possible. We took inspiration from Deep Image Prior (DIP) [22]. Ulyanov et al. [22] claims that the structure of a convolutional generator network is sufficient to act as image prior, i.e. knowledge available before the restoration process, for many images, making it independent of the learning process. This eliminates the need for pre-training the network on large datasets.

Combining DIP with Human-Computer Interaction resulted in an interactive Machine Learning (iML) [7] process: After running the image through the automated DIP restoration, the human operator can manually refine the image via a user interface. This image is then passed through the DIP again for polishing and the process repeats until the user is satisfied. Not only does this mean the process can be terminated anytime once a level of quality has been reached that satisfies human perception, but it also provides more frequent feedback as to the restoration progress. Since these visual media are primarily designed for human perception, tailoring the output of a process like image restoration towards perceived quality should be a key focus. Additionally, an interactive process informs the user about their impact on the restoration results as well as how well the system performs, which can lead to better system transparency.

As a combination of automated, ML elements with human interaction, this system is a "Human-in-the-Loop" ML system [13]. Those are systems that involve the human in the machines operations to inform the human and/or to improve the performance, accuracy or efficiency of the machines.

Since human-in-the-loop approaches require some work from the user, it is important though, that the interactivity and the human involvement in the process is designed with human factors in mind to ensure user acceptance and satisfaction, ideally such that even people with little restoration expertise should be able to create plausible images with the proposed system.

We present our interactive Deep Image Prior (iDIP) in Sect. 3, including how it automates the majority of the process but enables human operators to embed their knowledge by manually providing additional image information. We implemented a back-end that runs the iDIP to perform the automated steps, as well as a front-end with some drawing capabilities that allows the human to directly manipulate the intermediate increments.

We then evaluated this system with respect to two core questions:

(1) **Does the interactive approach produce higher quality reconstructed images?**
We evaluated this in terms of both objective measures like Dissimilarity Structural Similarity Index Measure (DSSIM, Sect. 4.2) and subjective perception of users (Sect. 4.3).
(2) **How do users view such a system in terms of user experience and satisfaction?**
We evaluated this with a users study (Sect. 4.4).

2 RELATED WORK

Restoration of historical artifacts and images is of course far from a new field. It has, however, gained increasing attention recently due to technological developments. The following section outlines some of the new approaches to image restoration. It also gives some context to our contribution, the interactive approach, by providing some background on iML.

2.1 Image Inpainting

Especially the development of DL and computer vision have contributed to the increased attention inpainting has received recently. There are two main categories of approaches demonstrating state-of-the-art performance for inpainting tasks. Exemplar-based restoration algorithms, such as PatchMatch [1] and PatchOffset [10], fill the missing or damaged parts by copying local patches from the same image to those regions. PatchMatch [1] quickly finds approximate nearest-neighbor matches between image patches and was adopted into Photoshop. PatchOffset [10] minimizes an energy function to select patches with dominant offsets. This approach is especially suitable for images with simple and repeating textures. However, images which contain rich semantic information are difficult to restore by simply copying local patches. To deal with complex content information in those missing parts, learning-based methods, such as EdgeConnect [18] and PartialConv [17] address the image restoration problem in a data-driven manner. These methods leverage the expressiveness of Deep Neural Networks (DNNs) and

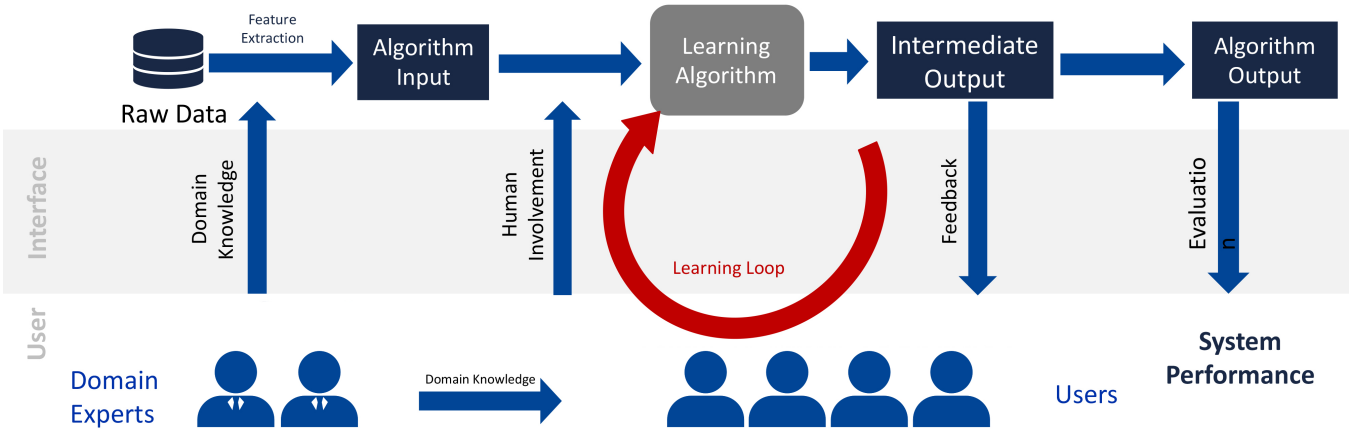

Figure 2: A framework of Human-in-the-Loop ML, where human perception and intelligence can be tightly integrated with the advances of ML.

learn the semantics in corrupted regions on massive datasets. In the past years, many variants of DNN-based algorithms have been studied for image inpainting, e.g. Deep Convolution Generative Adversarial Networks (DCGANs) [19] and Deep Convolution Autoencoder (DCA) [5], and achieve the state of the art performance on benchmarks for content-aware image inpainting tasks [14, 30]. EdgeConnect [18] proposed a two-stage adversarial model and can deal with irregular masks, while PartialConv [17] utilized partial convolutions with an automatic mask update step. Their data inefficiency requires that huge amount of data must be used for training such networks though, which is still infeasible for most practical image restoration tasks, where very few or even no ground-truth images are available.

To tackle the data inefficiency problem, Ulyanov et al. utilized DNN in a different way and presented Deep Image Prior (DIP) [22], a DNN-based multi-task solution. Compared to classic DNN-based inpainting approaches, one notable improvement is that it does not require any pre-training or training data. However, the authors report that DIP still suffers from overfitting with the increasing iteration number, since DIP utilizes an over-parameterized DNN. We will further discuss DIP as the basis of our approach in Sect. 3.1.

2.2 Interactive Machine Learning

In this paper we describe our extension of the DIP method by bringing the human in the loop, making it an iML setup. iML tools are gaining interest as a alternative to fully automated ML.

Most ML research until now has concentrated on fully automated ML, where great advances have been achieved by DL in recent years, for example, in image classification [23], natural language processing [29] and recommender systems [11]. A Classic Machine Learning (CML) process typically starts with feature engineering by domain experts or specific algorithm input for the target application. CML users need to work together with domain experts to identify and determine data patterns. Next, ML experts experiment with different ML algorithms, tune parameters, tweak features and collect more data to improve target performance metrics. However, CML and many of its applications are considered hard to approach due to two main reasons: the complexity of the algorithms and the low data efficiency.

Potential users of ML-based applications are always non-experts for the underlying task and ML, so that the tight coupling between system and users can be hardly modelled with CML when no sufficient support from professionals is available. As a result, it is not possible to apply their perception and insight such as found patterns to enhance the learning algorithms. On the other hand, fully automated approaches greatly benefit from massive data with large training sets. However, in some domains with high dimensional input data, a certain accuracy must be guaranteed while less data is available [12]. For example, ML algorithms would fail in healthcare and medical diagnostics where they would have to deal with very small datasets. This would lead to strong biases for fully automated approaches due to insufficient and unbalanced training samples. Therefore, the limited involvement of human expertise and the low data efficiency are two factors that largely prevented ML from being adopted into these domains.

iML overcomes the aforementioned disadvantages by extending ML with interactivity. One way is with systems that use interaction as an input modality for human knowledge to learn from human perception. The insights gained by this make it a popular approach that is quickly becoming widespread [6]. Another method for iML to improve both task accuracy and user satisfaction of ML systems is to introduce the human into the training loop of ML algorithms, which is refereed as Human-in-the-Loop ML.

Human-in-the-Loop ML aims to complement ML algorithms with human perception and intelligence by tightly integrating human knowledge with the power of ML. Compared to the one-time training in CML, Human-in-the-Loop ML breaks down the tasks in iterative learning loops as shown in Fig. 2. As a first step, an appropriate basis (intermediate output) is formed by using the CML algorithms. Then users are able to derive insights from this basis and contribute to them via a interactive process. The interactive

process is designed to incorporate input from the user but does not require much domain knowledge that might be necessary to work with most CML techniques. Finally, the overall performance can be boosted by leveraging the knowledge encoded in the human involvement obtained from a careful designed interactive user interface.

Recent research provides a number of case studies that show how existing ML-based systems fail to account for the user, so they instead explore new Human-in-the-Loop approaches. Caruana et al. [3] developed an interactive protein taxonomy by clustering low-level protein structures. This framework allows domain experts to critique and set new constraints for low-level protein structures for next iteration. Sanchez-Cortina et al. [20] proposed an automatic speech recognition system with an interactive interface that facilitates error correction. Fails' et al. [7] frequently cited work describes an idea similar to ours. It demonstrates a user interface where users can train a classifier incrementally by drawing on an image. With a Human-in-the-Loop ML process, even non-experts can gain insights into unwieldy datasets and contribute, regardless of their limited domain knowledge, to the use of complex, data-driven applications. This process is co-adaptive in nature and relies on carefully designed interactions between human and machine [6].

iML is well suited to be applied in image processing, when interactions can be realized by providing pixel-level feedback by directly sketching on images, which is successfully exploited in several pior research works.

3 METHOD

In the following section, we present iDIP, a collaborative and interactive image restoration process. We demonstrate its setup and the methods used to perform the interactive image restoration. Our method of image reconstruction is based on the DIP, which we extend with interactivity, making it an iML system [7].

3.1 Deep Image Prior

Unlike most DNN-based image inpainting frameworks, DIP does not directly learn pixel-level statistics on a training set. Instead, DIP considers the architecture of Convolution Neural Networks (CNNs) [15] as image prior and adjusts the network parameters so that it can recover the undamaged part of one single image. DIP can be generalized as a fully automated framework for multiple tasks e.g., image super resolution, image denoising and image inpainting. Mathematically, DIP minimizes the following loss function in image inpainting tasks:

$$\min_{\boldsymbol{\theta}} \mathcal{L} = \min_{\boldsymbol{\theta}} ||(f_{\boldsymbol{\theta}}(\boldsymbol{z}) - \boldsymbol{x}_0) \oslash \boldsymbol{m}_0||_2, \tag{1}$$

where $f_{\boldsymbol{\theta}}$ is a convolutional generator network parameterized with $\boldsymbol{\theta}$, $\boldsymbol{z}$ is a fixed input, $\boldsymbol{x}_0$ is a corrupted image, $\oslash$ is the Hadamard product and $\boldsymbol{m}_0$ is the mask for damage area. By taking a fix noise input signal $\boldsymbol{z}$ as input, DIP gets rid of the requirement of large training set while maintaining the iterative, learned improvement process of learning-based methods [14, 27, 28, 30]. DIP can iteratively capture the complex textural semantics in uncorrupted region using backpropagation algorithm [16] in an end-to-end manner. With its iterative update quality of the restored image continues to improve as the generator is trained further.

While this makes DIP quite effective and, when fully automated, very efficient, overfitting can lead to deteriorating quality especially when larger areas need to be inpainted. Additionally, when the content of missing patches is obvious from semantic but not structural context, humans, who can infer these missing semantics, have an advantage.

To leverage this human knowledge and to overcome the overfitting issue, we extended DIP to a Human-in-the-Loop ML system.

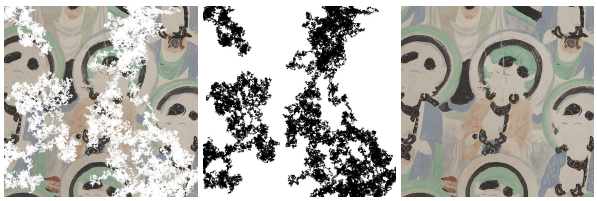

Figure 3: Left to right: the damaged image from the Mogao Grotto dataset [31], a mask specifying damaged regions, and a restoration by iDIP.

3.2 Interactive Image Restoration

To our knowledge, there exists no interactive extension of DIP for inpainting. We therefore attempt to leverage human knowledge and overcome some of DIP's disadvantages by extending DIP with interactivity that brings humans into the training loop.

iDIP restores images by alternately and iteratively exploiting the image prior and human knowledge. The underlying algorithm updates the image iteratively, incrementally, and focused onto specific masked regions (see Fig. 3). Refinement by the user can come in two forms: First, the user can edit the mask and therefore direct the DIP to include or exclude specific regions in the restoration process. Second, the user can paint onto the current increment to provide information that may not be restorable by structural information alone. This may for example be features that can be deduced from image semantics. This in particular is where human-machine-collaboration can shine, since especially features obvious from the semantic context are easily detected by humans, from the original damaged image, but especially from a first iteration when something in the image looks wrong. Blending hand painted image features into the structure of the original image can be hard though. With the collaborative approach, this can be left to the DIP algorithm.

The results of the human involvement are fed back into the DIP system which continues training – and therefore refinement of the image prior – until the next increment is reached. The human is in control of how many training iterations should be performed for the next increment, giving more control and making degradation due to overfitting less likely.

Fig. 4 visualizes the stages of iDIP:

(1) To provide some base information, the first increment x_0 is given without the user refining it.
(2) From then on the user always receives the current increment x_n restored by iDIP.

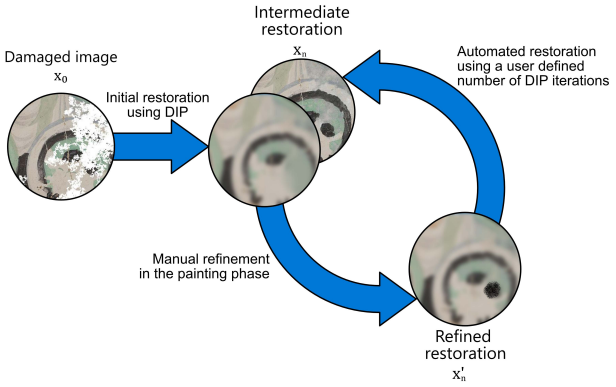

Figure 4: iDIP performs an initial restoration using DIP and then cycles through a phase of user refinement, followed by iterations of DIP.

(3) In the following painting phase, the user paints onto the image x_n to refine it, yielding the refined image x'_n.
(4) The refined image x'_n is fed back to the DIP algorithm for another training phase where a set number of training iterations are applied.
(5) After training, the system generates the next increment x_{n+1} from the further trained generator. At this point the process starts anew.

With the iterative nature we intend for DIP and human knowledge to jointly boost each other. Besides, this approach should also give users greater control on the output: by trial-and-error they can determine what impact their actions have to better gauge their actions for the next increment.

With the added interactivity and the subsequent breaking down of the whole restoration into smaller increments, the user remains in control of the algorithms progress. Frequent interactions can also contribute to greater user satisfaction. It might even give the user insights into how the algorithm performs which can further add to the user experience and transparency of the system.

By inspecting intermediate steps it can become clear early on when the algorithm fails to perform as expected and when changes are necessary. The worst case, when the restoration fails altogether and has to be stopped, can also be detected earlier, reducing the risk that comes with unsupervised long-running jobs. Likewise, the restoration can also deliberately be stopped to prevent overfitting, visible for example when textural consistency degrades, which can happen with to large a number of DIP iterations [22].

To facilitate our interactive system, we implemented the DIP as a backend that accepts requests for iterations via an API using Python. To achieve acceptable response time, as necessary for an interactive system, we ran the DIP backend on a server with the setup of 64 GB RAM, one 20-core Intel® Xeon® CPU and 8 Nvidia GeForce® RTX1080 GPUs. Our DIP implementations follows the original repository [1] and uses one GPU to avoid the complexity caused by gradient parallelization.

[1] https://github.com/DmitryUlyanov/deep-image-prior

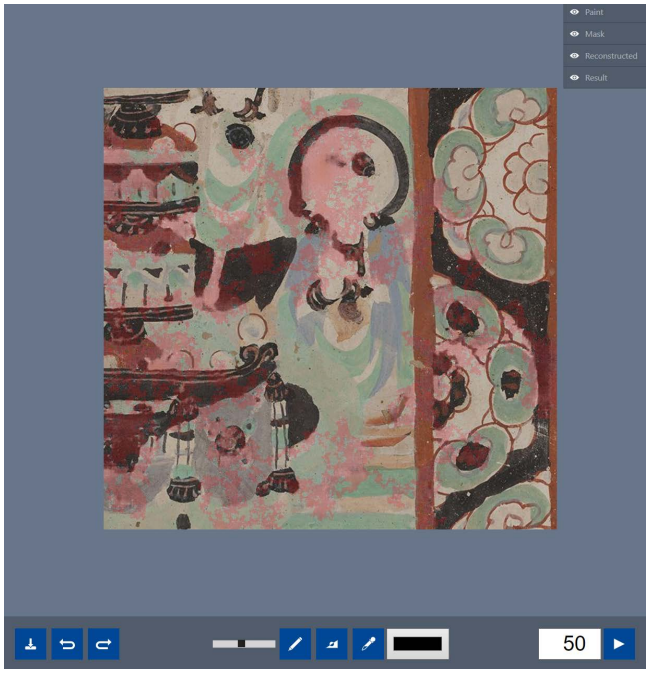

Figure 5: The user interface for the interactive image reconstruction. At the center is the image that is being reconstructed with the mask as red overlay. Top right the different layers. At the bottom the different tools.

3.3 User Interface and Interaction

While the backend is fully capable of running the DIP image reconstruction without any human intervention, our interactive system requires a frontend for a human operator to observe the progress and provide additional information. This happens via direct manipulation by painting onto the image. Therefore we create an HTML/JavaScript application for a web client, as shown in Fig. 5, which includes the following functions:

- It shows a composition of the known, undamaged image regions with the reconstructed patches filled in. The composition is necessary to give the user feedback as to how far the reconstruction has progressed while also showing the original context.
- Next to the composition the user can also toggle the visibility of the pure reconstructed image, the mask, if available, and the currently added paint as individual layers. This is a feature also commonly found in image manipulation software and allows the users to choose the context appropriate for them.
- On the paint layer the user can draw using a simple spray tool, which randomly colours pixels in a set radius in a set colour. We chose this method over a pen tool that pains all pixels within a radius because the textured images we reconstructed rarely had large patches of a single colour.
- The images from the Mogao Grottoes dataset come with a mask that should define the damaged region. For dealing with these masks we added two modes. While in the

first, constrained mode, the user could only draw inside the masked area, i.e. the area that needed reconstruction. This was ideal for when the users did not want to modify the original image, accidentally changing undamaged areas. In the second, free mode, the user could draw anywhere on the image. This was useful for corrections when the algorithm incorrectly reconstructed the image but flagged the area as already reconstructed and therefore masked. We also gave the users a tool for editing the mask.
- After finishing the painting, the user could send the image to the backend while providing a number of how many iterations of the DIP should be applied for the next increment. More iterations would lead to higher reconstruction quality but would take more time.
- There were also some additional convenience feature like undo and redo, a colour picker and an export feature which allowed the user to download the result.

To accommodate all drawing functionality, the images and masks are treated as different layer for the whole client-side process. Once the users chooses to transmit them for additional DIP iterations, they are composited on the client and submitted to the servers API, which triggers the DIP. Once the DIP has completed the requested number of DIP runs, the result is again sent to the client for additional drawing or for the users to export the image as final result. This way the performance and therefore the wait time for the DIP is independent from the client hardware. This was a prerequisite for performing the user study, since this way we are independent in terms of client hardware and therefore location and could scale the processing time down by running the backend on the aforementioned hardware.

4 EVALUATION

To ensure our interactive process fulfills its purpose, namely improving image restoration, we subsequently analyzed it. The analysis consists of two parts: regarding restoration quality and regarding usability and overall user perception of the tool and process.

The evaluation of the restoration quality firstly uses two quality measures for comparison with five baseline algorithms. However, these pixel-wise measures cannot account for the the criteria a human would use to judge the quality such as semantic correctness and consistency. As a consequence, we also asked humans to subjectively judge the different reconstructed images from our iDIP approach in comparison to those reconstructed with the five alternative methods listed below, two of which are exemplar-based, two learning-based and the pure DIP. Note that to evaluate the learning model, we used the pre-trained model on Places2 [2] [32], because it is one of the largest and widely-used scene recognition dataset.

The following section outlines the setup for these evaluations as well as the individual results.

[2]http://places2.csail.mit.edu/

Figure 6: The Mogao Grotto as an example of damaged cultural artifacts in need of restoration. The sheer volume of more than 45,000m^2 of murals and more than 2,000 statues motivates the need for automated restoration. Image taken from [31].

4.1 Evaluation Setup

4.1.1 Dataset. To evaluate a data-driven method we required a dataset to work with. We chose the Mogao Grottoes dataset [3] [31] which was readily available and matched our use-case of image restoration. It contains 500 full frame paintings with artificially generated masks for damaged regions of the more than 45,000m^2 of murals in the grottoes. The masks were generated using random walks from randomly selected points on the image [31]. One might argue that this process roughly matches how damage due to moisture or fungi spreads. We determined by visual inspection, that these masks, as seen in Fig. 3, seem like plausible damaged areas. For our evaluation we randomly picked ten sets of original ground truth image, generated mask, and resulting artificially damaged image.

4.1.2 Metrics. For the comparison, we utilized Local Mean Square Error (LMSE) [8] and Dissimilar Structural Similarity Index Measure (DSSIM) [25] as our quality measures for restoration performance. Means Square Error (MSE) is a common and easy-to-compute measure of estimation quality of the estimated values of independent variables. However, we would like each local region of the estimated images to be exactly same like the ground truth image. For this purpose, we computed the MSE of the masked region, which is equivalent to its LMSE by setting $k = 1$. The Structural Similarity Index Measure (SSIM) is a improved version of similarity measure for predicting the perceived quality of digital images or videos. The SSIM is a full reference metric, which means it is based on an initial uncompressed image (ground truth) as reference. By using

[3]The Mogao Grottoes, also known as the Thousand Buddha Grottoes or Caves of the Thousand Buddhas, consist of 492 temples spread over 25 km (16 mi) in the area to the southeast of the ancient city Dunhuang, an oasis located at a religious and cultural crossroad on the Silk Road, in the Gansu province, China. The grottoes may also be known as the Dunhuang Caves. The grottoes contain more than 10000 full frame paintings, which are consecutively created by ancient artists over a thousand years between the 4th and the 14th centuries.

Figure 7: Examples of images that can be found in the Mogao Grottoes dataset. Image taken from [31]

.

	LMSE	DSSIM
EdgeConnect	629.65***	0.2803***
PartialConv	2550.02***	02816***
PatchMatch	185.68	0.2423
PatchOffset	558.05***	0.2247*
DIP	214.23	0.2228
iDIP	**207.37**	**0.2227**

Table 1: Results for the restoration metrics. Lower values are better. Significance levels for comparison to iDIP using Mann-Whitney-U test.

$DSSIM = \frac{1-SSIM}{2}$ we let the DSSIM be inversely proportional to restoration quality as LMSE.

4.1.3 Baselines. To show the effectiveness of our interactive approach, we compared it with five state-of-the-art algorithms:

- **EdgeConnect** [18] proposed a two-stage adversarial model and can deal with irregular masks.
- **PartialConv** [17] used partial convolutions with an automatic mask update step.
- **PatchMatch** [1] quickly finds approximate nearest-neighbor matches between image patches, a process also adopted in some image manipulation software.
- **PatchOffset** [10] minimizes an energy function to find patches with dominant offsets.
- **Deep Image Prior (DIP)** [22] minimizes the Mean Squared Error (MSE) in the unmasked region while exploiting prior image information contained in the architecture of CNNs.

Fig. 8 shows a comparison of restored images from each method.

4.2 Objective Evaluation

To compare the reconstructed image to the baseline we used established measures for the comparison of images: We compute the Dissimilarity Structural Similarity Index Measure (DSSIM) [25] and the Local Mean Squared Error (LMSE) [8] between the restored and ground truth images. We used the Shapiro-Wilk test to determine that our quality metrics were not normally distributed. We therefore continued the comparison of the different methods using a pairwise comparison by Mann–Whitney-U test (see Tab. 1).

Clearly visible in the scores is the fact that, although we used the pre-trained model, the performances of two learning-based methods are worst for both DSSIM and LMSE. The reason is that the styles varied too much on training set and Mogao Grottoes Painting Dataset. This is another clear indicator that learning-based methods, especially deep learning methods, are unsuitable for inpainting task with few training samples.

Of the other methods, even though DIP optimizes on Mean Square Error, PatchMatch has the best LMSE score, although not statistically significant. Similarly there is no significant difference between the performance of pure DIP and our approach. Merely PatchOffset performed significantly worse than all other three methods ($p < 0.001$).

Results for DSSIM were similar with no significant differences between PatchMatch, DIP, and iDIP, while PatchOffset again showed significantly worse performance than our approach ($p = 0.010$). compare to the three other methods.

While significant improvement would have been more desirable, we see these results as an indicator that our approach achieves at least performance on par with these baselines regarding the objective measures.

4.3 Subjective Evaluation

While the aforementioned quality metrics can be an indicator for the quality of the image reconstruction, subjective perception remains an important factor when deciding whether a certain quality threshold has been reached or whether the reconstruction makes sense to begin with. As visible in Fig. 9, the differences between the methods can be subtle but may still be noticed, consciously or subconsciously, by humans. This does not only apply to interactive approaches, but to image reconstruction in general. Consequently we decided to conduct a user study with two goals: to evaluate the subjective quality of our image reconstruction as described below and to receive feedback on the overall usability of the tool and method, as described in the following Sect. 4.4.

Participants in this study (n = 19, 9 male, 9 female, 1 other, see Fig. 10) were people with mixed expertise with image manipulation and generally limited prior experience with image reconstruction but overall good self-reported technology skills (see Fig. 11).

To gauge the subjectively perceived image quality, we asked participants to judge restored images. We first explained the purpose of our tool and gave a quick introduction of the functionality. We then had participants play around and restore some images, described further in the following section.

After getting familiar with the capabilities of the image restoration system, we presented ten images, restored by six methods each, in randomized order. For each image the participants had to choose the two restorations they considered best, resulting in each participant choosing 20 images, yielding 380 restored images being selected.

Fig. 12 shows the absolute frequency of the participants choice. EdgeConnect and PartialConv are not shown since they had not

(a) Corrupted Image (b) EdgeConnect (c) PatialConv (d) PatchOffset (f) PatchMatch (g) DIP (h) iDIP

Figure 8: Images restored by our baselines and iDIP.

Figure 9: Left to right: the damaged image, the image restored by DIP, the image restored by iDIP, and a map showing the different regions between the two approaches. White level of the difference map has been increased by 75% for clarity.

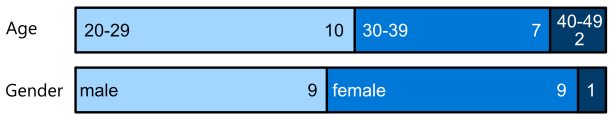

Figure 10: Demographic data of our 19 test participants.

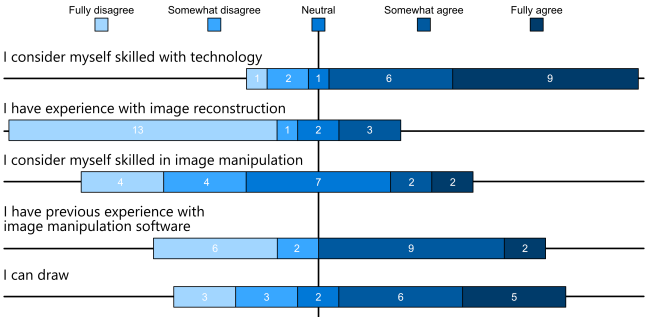

Figure 11: Prior knowledge of the participants. Self-reported on a five-point Likert scale.

been chosen at all. Interestingly, PatchMatch and PatchOffset were chosen almost equally often, 23 and 18 times respectively, even though their difference in the objective measures was considerable. This is mostly due to the worse performance of PatchMatch, which was on par with DIP and iDIP for DSSIM, even better for LMSE, but was chosen considerably less frequently, only by 12.1% of participants. DIP meanwhile was chosen by 78.9% of participants as

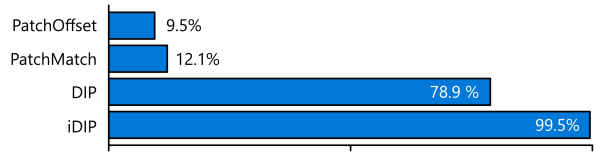

Figure 12: Frequency of how often participants chose each method in the top two restorations.

one of the top two and our approach, iDIP, was chosen for almost all images by all participants, 99.5% of times in total.

Clearly there is a difference between objective and subjective perception of the restored images. So while – by objective measure – iDIP is merely on par with some of the baselines, by the standard of human perception it outperforms all five baselines. The fact that it even outperforms the non-interactive DIP also is a strong indicator that the added interactivity improves the output quality.

4.4 User Experience Evaluation

While the improved quality is to be viewed positively, the improvements of an interactive approach can only be relevant when users choose to adopt it. We therefore asked the participants of the subjective evaluation for qualitative and quantitative feedback regarding the usability of our tool using a questionnaire including System Usability Scale [2] and NASA TLX [9]. The questionnaire also included questions regarding the benefits of machine learning for tools in general and for image restoration in particular. Participants filled this questionnaire after using our tool for restoring two images:

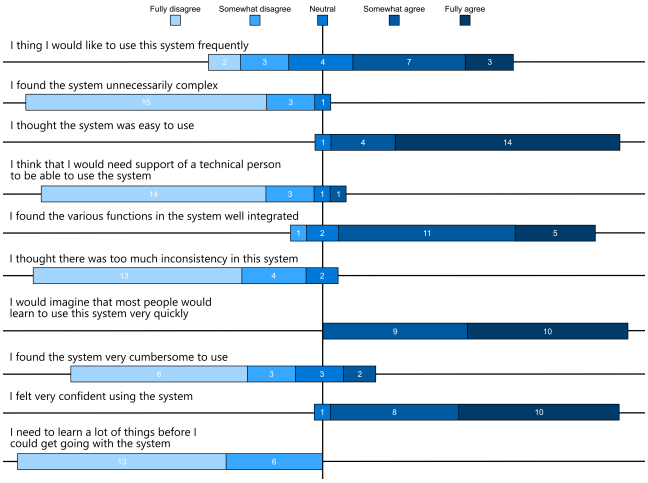

Figure 13: Individual Results of the System Usability Scale

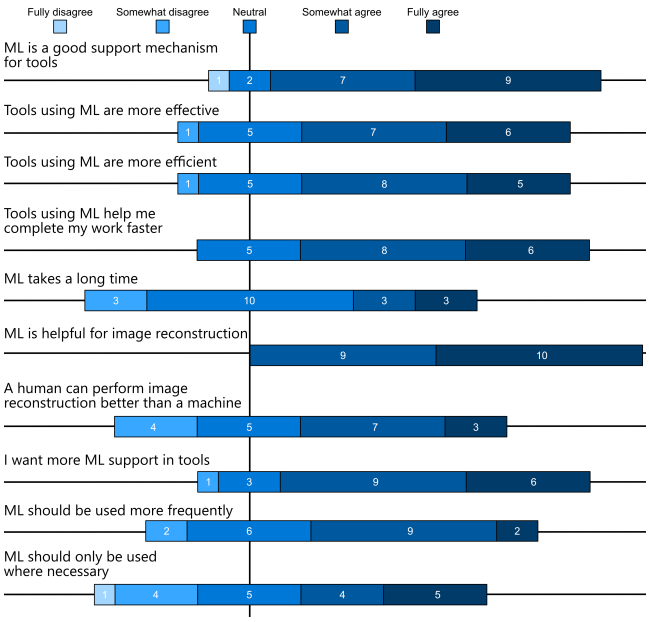

Figure 14: The participants attitude towards ML and ML in tools. Self-reported on a five-point Liker scale. Higher values mean stronger agreement with the statement.

After an introduction and explanation of the tool and its key features, each participant worked on two images. As described above, each damaged image received 500 DIP iterations initially to provide a basis for the painting. After this followed the painting phase in which the participants painted their correction and guidance. One of the two images the participants restored in a single increment with 1200 DIP runs after the painting phase. The other image was restored in two increments, the first with 500 iterations of DIP, followed by another painting phase, followed by 700 iterations of DIP. The order of those two variants was randomized between participants. In order to time-box the study we restricted the painting phases in the restoration process to five minutes each.

Regarding the user experience with our tool, three participants mentioned minor usability issues, none of which were an impediment to task-completion. In general the feedback towards the tool was very positive. Some participants requested additional features like more complex drawing capabilities. The only frequently voiced criticism was the processing time of the image reconstruction process being too long. This, however, is independent from the interactivity and could be alleviated by improvements to the underlying ML algorithms or running them on stronger hardware.

Likewise, the scores in the System Usability Scale are very positive with a total average score of 85 (out of 100). Individual results are shown in Fig. 13 Ease of use was rated very positively and the system was rated easy to learn, even though image restoration usually is a fairly specialized activity. The only relatively evenly distributed item in the SUS was "I think I would like to use this system frequently". This may very well be the case because image reconstruction is a very specialized activity and not very common in the day-to-day of our participants.

The NASA TLX showed relatively low reported values for workload on most scales. The notable exception in the TLX scales being "How successful were you in accomplishing what you were asked to do?" where our participants reported an average of 7.55 (out of 10). Given the relatively high number of iterations and amount of time necessary for a proper image reconstruction, coupled with our time constraint, it is unsurprising that participants felt unsatisfied with their results. Still, they agreed with our assessment that the cooperative, interactive approach can be a serious time-saver compared to the traditional manual restoration.

Overal the participants opinion towards ML supported tools and was very positive (see Fig. 14), with participants considering ML als support mechanism beneficial for effectiveness and efficiency. In particular for image reconstruction the participants saw the benefit of ML support. They were however critical of the time requirements, which matches the qualitative feedback, and how well the automated process compares to a restoration done by a human alone.

Participants responded positively though to adopting ML for support in tools like ours although with the qualification of adding it only when necessary or where it made sense. In summary, the feedback towards ML as support mechanism in general and towards our tool in particular was very positive, which to us is a clear sign that interactive ML is a promising direction for improving workflows that are currently work- or time-intensive or could otherwise be improved by semi-automation.

5 DISCUSSION

These results look promising, indicating that added interactivity is a positive influence on image restoration.

Since our questionnaire also included questions on collaborative iML systems for other domains, which participants responded very positively towards, it is fair to state that at least in their subjective perception, interactivity might be adopted in other settings where task are supported by machine learning. Whether a collaborative system does in fact offer benefits in these situations cannot be

inferred from the image restoration scenario though, so it remains for future work. For image restoration, the benefits are fairly clear in both quality and user satisfaction though.

Nonetheless, these benefits have to be weighed against the cost, since of course a solution based on computation alone does not require any human interaction and is therefore still more time-efficient than an interactive solution. The fact that our participants achieved good results with very little interaction suggests that the human involvement is only a small fraction of the total time the process takes. Additionally, knowing whether the algorithm is performing as intended and the option to interrupt early on can be a significant gain in cases where the algorithm fails and would otherwise have run to completion only to return an unsatisfactory result. There is also added flexibility in our approach, since the total time consumption can be adjusted depending on the requirements by choosing different settings for how many intermediate increments are created and for how many iterations the DIP runs for each increment. Figuring out the trade-off between control and time by varying the number of increments remains an open question though.

On the other side of the time-quality trade-off there is also a different consideration: there are applications of image restoration where quality is the primary objective and the time-consumption is secondary. In these cases a manual approach might still be the most viable option. As we mentioned before though, a pure digital reconstruction is often not the ultimate goal and only serves as a frame of reference or material for discussion. The interactive approach still offers this handily. The reduced entry barrier in terms of necessary expertise also enables more people to participate. This may positively contribute to debate what the ideal reconstruction may be, since the results of different people can differ to a greater degree than when only a single algorithm were used.

Of course, the quality of the restored images varies, depending on the expertise of the operator. This has the downside that inexperienced users may easily give inappropriate feedback, different from what a professionals might provide. This could be in direct contradiction to the actual true image patterns. Unfortunately, these improper interventions can not be recognized by iDIP and are therefore considered to be ground truth in the training process. Such ill-regularized DIP could easily degrade the restoration performance instead of improving it.

Also regarding the image manipulation skill, our interface was not specifically tailored towards experts, as our participants were no image restoration professionals. People with year long experience and with an interface tailored to their needs may produce even better results. The quality is also dependent on the functionality the interactive tool offers. We only implemented rudimentary painting functionality. If such a tool had a rich tool palette as found in common image manipulation software, the options for an expert are much broader, potentially leading to even better results. Inversely, it may also be possible to implement an image restoration functionality or similar interactive machine learning supported features into existing image manipulation software, where they would benefit greatly from the existing general purpose tool ecosystem.

Whether our results do apply to all images one might want to restore is also debatable. The cave painting restoration use case is a plausible but not a common one. However, these paintings were fairly abstract, since motives were not entirely clear from just the small subsections we used in our study. If human involvement is beneficial for abstract visuals, where even the human has to more heavily rely on structural information, it seems plausible that the benefit of human knowledge and semantic recognition becomes even greater for images with clearer motives. This remains to be validated in future work though.

Another issue of generalizability is the specialized nature of image restoration. As our participants reported, they were no experts for image reconstruction. While this can be seen as a positive in the sense that our tool is also usable for non-experts, it does limit the expressive power of our study for image restoration professionals and their work. It does seem reasonable though to assume that the benefits of an interactive approach would transfer to an expert interface, even it that might differ in functionality and depth.

6 CONCLUSION

In this paper we have described iDIP, our Human-in-the-Loop framework for interactive image restoration. This framework allows users to interactively contribute their knowledge to a DIP-based image restoration process such that both image prior and human knowledge are used as a collaborative iML system. We have outlined our implementation of this system as well as how we evaluated whether the interactive approach improves output quality and how it is perceived by users. Our experiments show that the interactivity positively affects the output quality as iDIP is on par with or better than the five state of the art baselines. Meanwhile, according to the user study, we achieved good user satisfaction, as participants stated their appreciation and confidence of the proposed method. They also see similar approaches for other tasks and domains a viable prospect. These positive results for our two core research questions indicate to us that iML is desirable for image restoration. Our study participants also enjoyed the collaborative setup and saw its potential. We leave the adoption to other domains and tasks as future work.

ACKNOWLEDGMENTS

This research was funded by the Bavarian Ministry of Economic Affairs, Regional Development and Energy.

REFERENCES

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, Vol. 28. ACM, 24.

[2] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[3] Rich Caruana, Mohamed Elhawary, Nam Nguyen, and Casey Smith. 2006. Meta clustering. In *Sixth International Conference on Data Mining (ICDM'06)*. IEEE, 107–118.

[4] Antonio Criminisi, Patrick Perez, and Kentaro Toyama. 2003. Object removal by exemplar-based inpainting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, Vol. 2. IEEE, II–II.

[5] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio. 2013. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 403–410.

[6] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 8.

[7] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 39–45.
[8] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. 2009. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2335–2342.
[9] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
[10] Kaiming He and Jian Sun. 2012. Statistics of patch offsets for image completion. In *European Conference on Computer Vision*. Springer, 16–29.
[11] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 173–182.
[12] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.
[13] Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M Pintea, and Vasile Palade. 2016. Towards interactive Machine Learning (iML): applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In *International Conference on Availability, Reliability, and Security*. Springer, 81–95.
[14] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 107.
[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. [n.d.]. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
[16] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 4 (1989), 541–551.
[17] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 85–100.
[18] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. 2019. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212* (2019).
[19] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).
[20] Isaias Sanchez-Cortina, Nicolas Serrano, Alberto Sanchis, and Alfons Juan. 2012. A prototype for interactive speech transcription balancing error and supervision effort. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. ACM, 325–326.
[21] Jian Sun, Lu Yuan, Jiaya Jia, and Heung-Yeung Shum. 2005. Image completion with structure propagation. In *ACM Transactions on Graphics (ToG)*, Vol. 24. ACM, 861–868.
[22] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2018. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9446–9454.
[23] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.
[24] Huan Wang, Qingquan Li, and Qin Zou. 2019. Inpainting of Dunhuang Murals by Sparsely Modeling the Texture Similarity and Structure Continuity. *Journal on Computing and Cultural Heritage (JOCCH)* 12, 3 (2019), 17.
[25] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
[26] Yonatan Wexler, Eli Shechtman, and Michal Irani. 2004. Space-time video completion. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, Vol. 1. IEEE, I–I.
[27] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. 2018. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 1–17.
[28] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. 2017. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5485–5493.
[29] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine* 13, 3 (2018), 55–75.
[30] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5505–5514.
[31] Tianxiu Yu, Shijie Zhang, Cong Lin, and Shaodi You. 2019. Dunhuang Grotto Painting Dataset and Benchmark. *arXiv preprint arXiv:1907.04589* (2019).
[32] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464. http://places2.csail.mit.edu/.