




Artificial Intelligence for Dunhuang Cultural Heritage Protection: The Project and the Dataset

Tianxiu Yu^{1,2} · Cong Lin³ · Shijie Zhang⁴ · Chunxue Wang² · Xiaohong Ding² · Huili An² · Xiaoxiang Liu³ · Ting Qu³ · Liang Wan¹ · Shaodi You⁵  · Jian Wu² · Jiawan Zhang¹

Received: 2 April 2021 / Accepted: 27 July 2022 / Published online: 24 August 2022
© The Author(s) 2022

Abstract

In this work, we introduce our project on Dunhuang cultural heritage protection using artificial intelligence. The Dunhuang Mogao Grottoes in China, also known as the Grottoes of the Thousand Buddhas, is a religious and cultural heritage located on the Silk Road. The grottoes were built from the 4th century to the 14th century. After thousands of years, the in grottoes decaying is serious. In addition, numerous historical records were destroyed throughout the years, making it difficult for archaeologists to reconstruct history. We aim to use modern computer vision and machine learning technologies to solve such challenges. First, we propose to use deep networks to automatically perform the restoration. Through out experiments, we find the automated restoration can provide comparable quality as those manually restored from an archaeologist. This can significantly speed up the restoration given the enormous size of the historical paintings. Second, we propose to use detection and retrieval for further analyzing the tremendously large amount of objects because it is unreasonable to manually label and analyze them. Several state-of-the-art methods are rigorously tested and quantitatively compared in different criteria and categorically. In this work, we created a new dataset, namely, AI for Dunhuang, to facilitate the research. Version v1.0 of the dataset comprises of data and label for the restoration, style transfer, detection, and retrieval. Specifically, the dataset has 10,000 images for restoration, 3455 for style transfer, and 6147 for property retrieval. Lastly, we propose to use style transfer to link and analyze the styles over time, given that the grottoes were build over 1000 years by numerous artists. This enables the possibly to analyze and study the art styles over 1000 years and further enable future researches on cross-era style analysis. We benchmark representative methods and conduct a comparative study on the results for our solution. The dataset will be publicly available along with this paper.

Keywords Cultural heritage protection · Dunhuang · Artificial intelligence · Computer vision

Communicated by Takeshi Oishi.

T. Yu and C. Lin are equally contributed to this article.

Supported by National Key R&D Program of China, Grant No.2020YFC1522705 and Grant No.2020YFC1522701, and the National Natural Science Foundation of China, Grant No. 62006101.

✉ Shaodi You
s.you@uva.nl

✉ Jian Wu
wujian@dha.ac.cn

✉ Jiawan Zhang
jwzhang@tju.edu.cn

¹ College of Intelligence and Computing, Tianjin University, Tianjin, China

1 Introduction

1.1 Background of Dunhuang and e-dunhuang

Dunhuang Academy manages three World Cultural Heritage sites: Mogao Caves, Maijishan Caves, and Bingling Temple, and other sites, such as Yulin Caves, Western Thousand Buddha Caves, and Northern Cave Temple. Mogao Caves is also known as the Thousand Buddha Grottoes of the Thousand Buddhas. It consist 492 caves spread over 25 km (16 mi) in

² Dunhuang Academy, Dunhuang, Gansu, China

³ Jinan University, Guangzhou, China

⁴ Tianjin Medical University, Tianjin, China

⁵ University of Amsterdam, Amsterdam, The Netherlands

the area to the southeast of the ancient city Dunhuang, which is an oasis located at a religious and cultural crossroads on the Silk Road, Gansu Province, China. The paintings were created by ancient artists over a thousand years in between the 4th and the 14th centuries. At present, more than 45,000 m² murals and 2000-plus painted sculptures are preserved. The murals are of great value for historical, artistic, and technological research with the earliest ones dating back to over 1600 years ago. Some photos of the sites are as shown in Fig. 1. In 1987, the Mogao Caves is recognized as the United Nations world heritage .

The mural of Dunhuang heritage covers a wide range of scenes and activities in ancient time, including daily lives, historical/political events, the use of technologies and art creation etc. Countless elements with different colors and pigments were drawn on the murals. Dunhuang heritage is a valuable historical asset for understanding ancient lives because it depicts a wide range of themes, from famous diplomatic missions to economic activities and technological demonstrations, from daily life to holiday celebrations, and from fictions to religious stories in classic scripts.

e-dunhuang Preserving cultural heritage objects is crucial (Ikeuchi, 2013). In the last 20 years, digital preservation (Banno et al., 2008; Bok et al., 2011), which measures such objects in three dimensions and recording them in digital forms, is one of the best ways for permanent preservation. A notable example is the Bayon project led by Katsushi Ikeuchi Banno et al. (2008), which created a holistic geometric and photometric model of the Bayon Temple in Cambodia.

In comparison, Dunhuang Academy has developed a full solution of digitized grotto cultural treasures and huge data findings after 30 years of investigation and study on digital technology, equipment, and techniques for the permanent preservation and sustainable use of the information of the Dunhuang Grottoes. Great effort and resources are invested in data acquisition, which is a race against time as the heritage sites are threaten by various deterioration. New data acquisition instruments have been developed by Dunhuang Academy, as shown in Fig. 2: (a) shows the work in 1960–1970s by using twin-lens reflex camera, the best capturing equipment in Dunhuang Academy, which has larger film to capture more details of the paintings. (b) shows the work in 2000s by using a complete digital solution. It includes hardware, such as new lighting device, tracking and capturing device; high color accuracy monitor, image processing narrow spaces) by using an ultra-large frame gigapixel digital camera developed by Microsoft Asian Research Institute. (d) shows our automated data acquisition instruments.

The high-accuracy data collection of 258 caves has been accomplished. The captured images have 300 dpi, and the distortion errors are well-controlled under 2 mm. As for the 3D reconstruction of color sculptures in the grottoes, forty-

five color sculptures have been captured and reconstructed by using laser scanner and structural light. Furthermore, the six major cultural heritage sites managed by Dunhuang Academy and the surrounding area of the Dunhuang city are also digitized. The online free database (a.k.a. *e-dunhuang* database) is a significant milestone in the digitizing Dunhuang Cultural Heritage (DunhuangAcademy, 2021). The database has high-resolution photos and a panoramic view of 30 caves. Besides by providing a virtual tour on *e-dunhuang*, the digitized heritages are also used in developing impressive artworks for exhibition using latest technologies, i.e., holographic projection, 3D sphere screen, embedded apps on social networks, immersive exhibition, and IMAX screen. These new forms of exhibition have attracted the public attention and gave the viewers different experiences in appreciating the splendid and culture-rich Dunhuang heritage. Subsequently, the digitization works have significantly promoted the public awareness of heritage reservation and protection. More importantly, digital Dunhuang is providing data access for academic research, fine art, and education

1.2 Challenges in Traditional Cultural Heritage Protection

Although the e-Heritage helps in efficiently recording and modeling the heritage sites very well, some problems, such as restoration, archaeology research, and creative material production, still need to be solved. Such digital archiving data are useful not only for the preservation in digital form of our irreplaceable treasures for future generations but also for the physical/virtual restoration and cyber archaeology. Cyber archaeology uses those modeling results (*e-Heritage*) for investigation and analysis with computer in the archaeology research area (Ikeuchi, 2013).

1.2.1 Virtual Restoration

The *e-Heritage* data from Dunhuang Academy are frequently used in exhibitions in various forms. The practical demands from the public exhibitions are different from archaeological demands, which require the presented content to be completely genuine. Regardless the artworks are printed in hard copies or on digital screens, the public exhibitions consider elegance of the *e-Heritage* more, and the exhibition goes would like to appreciate the artworks as intact as at the time they were created. In such scenarios, it is appropriate to present virtually restored *e-heritage* in image form if the global and local contextual consistency can be guaranteed. With the development of computer technology, image processing, graphics and other techniques have been increasingly applied to the preservation of cultural relics. Virtual

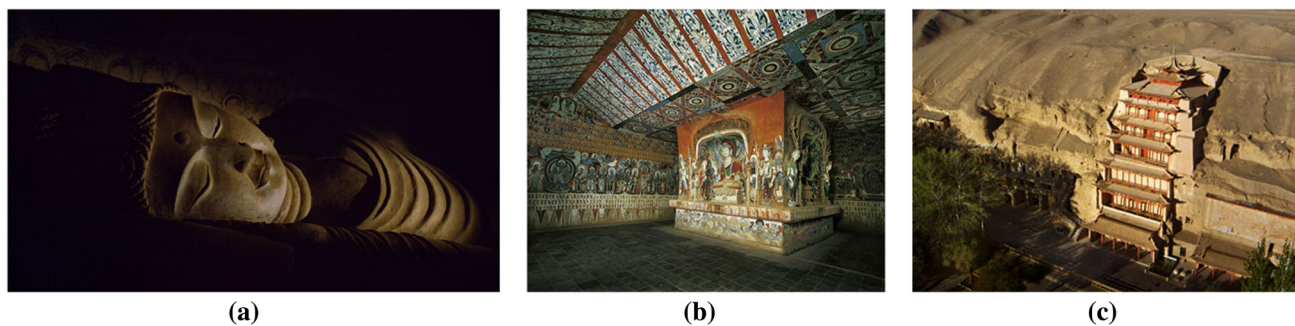


Fig. 1 The Dunhuang Grottoes at the edge of the desert preserves the large amount of cultural heritage in different forms. **a** Buddha sculpture in Cave No. 158 at Mogao Grottoes, **b** rich paintings on the wall and roof inside of a grotto, and **c** outside view of the Grottoes and its external ancient constructions

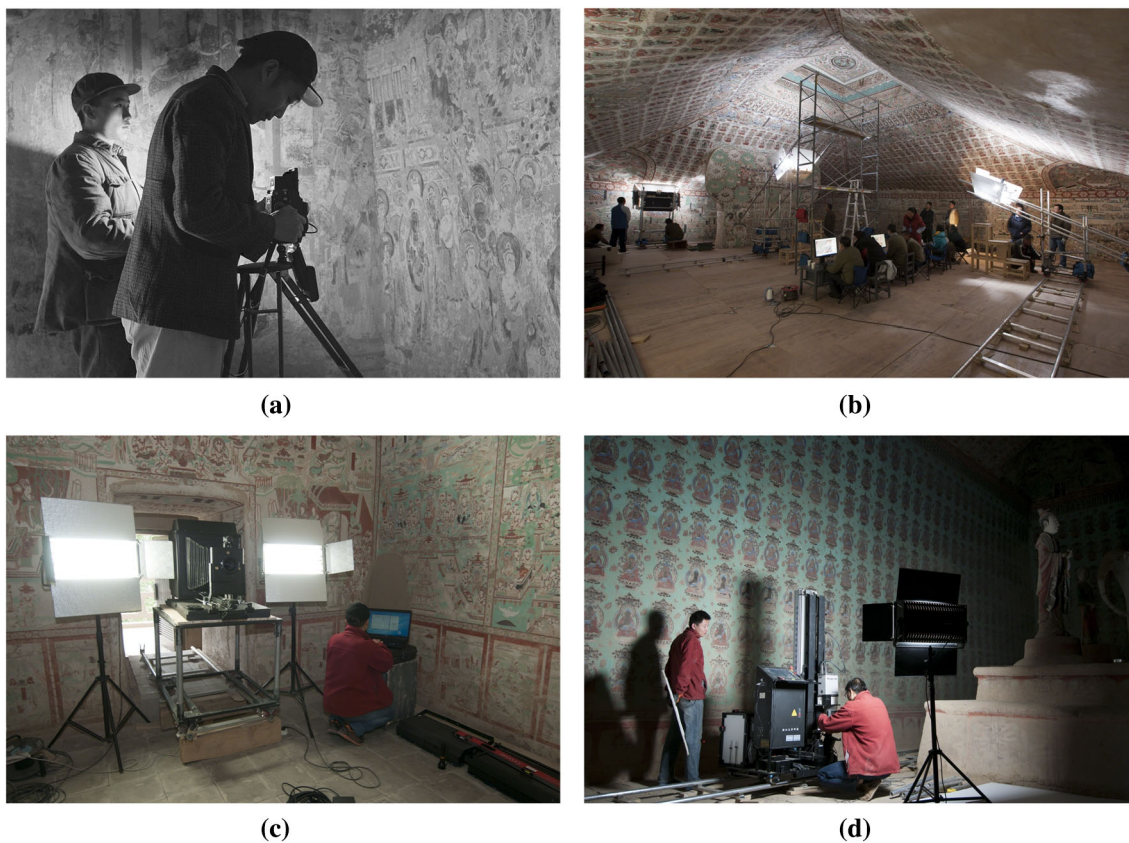


Fig. 2 Data acquisition in different periods: Utilization of **a** a large film camera, **b** a manual digital method, **c** an ultra-large frame camera, and **d** an automated digital method

restoration has become a research hotspot of digital preservation.

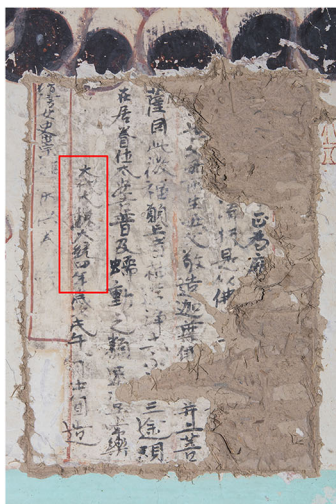
In virtual restoration, the operation targets are images instead of objects themselves. This approach can avoid damages to the original ancient paintings, and minimize the risk of physical protection and repair process. In the past, e-Heritage images are manually repaired and enhanced by artists, as shown in Fig. 3. However, the manual virtual restoration by human is prohibitively costly. The artists are required to be well-trained and familiar with the drawing skills of ancient time, and the process involves careful structure sketching and

color painting. The training and practice for skilled artists consume a substantial amount of time and human resources (Hou et al., 2018).

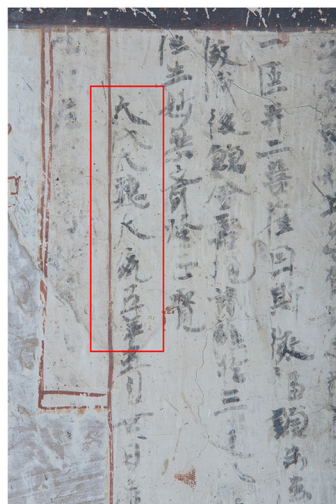
1.2.2 Historical Property Retrieval

Archaeologists carry out their work mostly by inspecting the e-Heritage, including knowing the overall statistics of e-Heritage in very detail, such as humans, trees and Apsaras. Instead of reviewing numeric images one by one and scanning the detailed content by raw eyes, which is unrealistic, an object detector can accelerate archaeologists' work of

Fig. 3 An example of result of manual restoration. **a** Wall painting damaged due to aging. **b** Partially restored by experts (the upper lost regions are carefully in-painted with, and with slightly color enhancement)



(a)



(b)



(c)



(d)

Fig. 4 Inscriptions with deterioration. **a** The inscription in the red frame is inscribed as “Dadai, Dawei, and Datong fourth years (AD 538) in Mogao Cave No. 285”. **b** The inscription in the red frame is inscribed as “Dadai, Dawei, and Datong fifth years (AD 539)” in Mogao Cave No.285. **c** Mogao Cave No. 290 depicts a benefactor of the Northern

Zhou Dynasty wearing a high crown, a long robe, and an overcoat on the shoulders (AD 557–AD581). **d** In Mogao Cave No. 409, the donors in the Uygur Era wore robes with lapels, tiny peach-shaped crowns on their heads, and red robes all over their bodies (AD 907–AD960) (Color figure online)

finding out the targeted elements and provide important clues for further archaeological research. For example, the inscription in the mural is extremely important for determining the content of the mural. Archaeological researchers can analyze in which year the cave was excavated by combining the year recorded and the events described by the inscription shown in Fig. 4, with the characteristics of ancient time, such as the clothes, decorations, and makeup of the donor. If the different elements in the Mogao Grottoes can be provided to archaeologists for comparative analysis, then the dating of the Mogao Grottoes will be more accurate, and some new content will even be discovered, as reported by Wang (2019).

1.2.3 Artistic Style Transfer

Cultural heritage carries the art from the past. The artistic style is one of the most important aspects of an artwork, which is the summary of characteristics from a set of images, including color distribution, drawing skill, and texture. Different artistic styles created with unique sketch skill and arrangements of elements and painted in various color distributions are used to support different themes. Some artistic styles may share some subtle similarities, while others show evident differences. Examples of different styles are shown in Fig. 5, the theme, elements and color distribution of those styles are

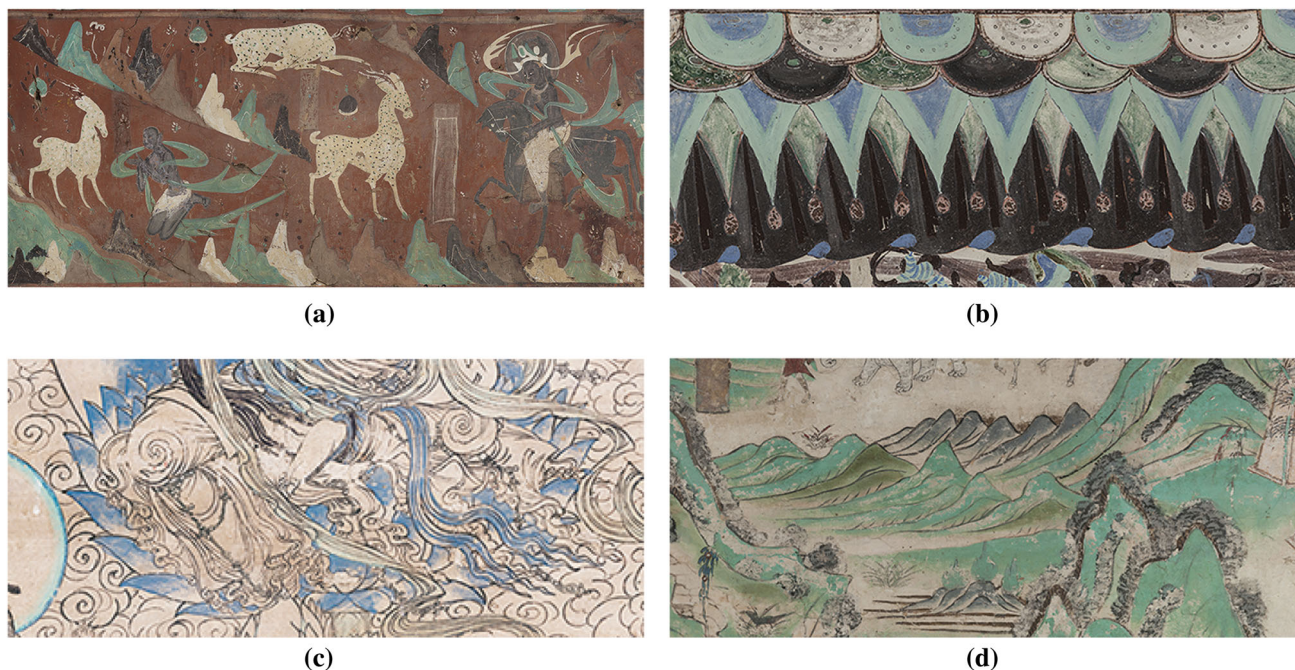


Fig. 5 The murals are created by artists from the 4th to 14th centuries. Murals that created in different era have significantly different styles. **a** Northern Wei Dynasty (AD 386–AD 534), **b** Sui Dynasty (AD 581–AD 618), **c** Western Xia Dynasty (AD 1038–AD 1227), and **d** High Tang Dynasty (AD 705–AD 781)

different, while the green color in Fig. 5a and d and the blue color in Fig. 5b and c are very similar. To make cultural art alive in modern time and attract public interest, Dunhuang Academy is working on artistic style transfer that could convert the artistic characteristics of e-Heritage images into daily photos, videos, and even sketches. For example, the interactive design project Dome to Home in Mogao Caves Cloud Museum mini program has attracted an enormous amount of page views, and it needs to provide users with enough image design options with lower entry of creation and higher efficiency. The system provides rich elements that extracted from domes, with user friendly design system. It encourages users to participate and create a unique scarf of their own. With the page view over 40 million, it had increased the awareness of the cultural heritage from Dunhuang.

1.3 Introducing Artificial Intelligence to Dunhuang Cultural Heritage Protection

In recent years, the heritage sites started to apply AI in protection, restoration, and archeology with the advance of artificial intelligence and computer vision technology. For example, AI is used for ancient character recognition (Haliassos et al., 2020), automatic recognition of archaeological pottery from excavations around the world (Anichini et al., 2020), and generate sketch for ancient mural paintings automatically (Pan et al., 2018; Sun et al., 2018). The Dunhuang Academy is proposing to collaborate with multiple research institutes

over the world, aiming to adopt the state-of-the-art computer vision technology for preservation and restoration. Among these collaborating research institutes are Tianjin University, Zhejiang University, Wuhan University, Jinan University, and University of Amsterdam, which have outstanding computer vision labs in the region. Many tasks are carried out, including virtual restoration (Yu et al., 2019b), cultural property retrieval, and artistic style transfer.

With the large scale of data collected, Dunhuang Academy is trying to apply the state-of-the-art artificial intelligence methods into solving problems and meet the demands in various directions (*e.g.* archaeology, education and exhibition). In this work, we focus on three tasks:

1.3.1 Virtual Restoration

The Dunhuang Academy has been trying to use data driven approach to in-paint the deteriorated regions with human label masks. The archaeologists judge what kind or where the deteriorated regions should be restored, while machine learning model in-paints the regions with texture learned from other similar intact regions, and the restored results are reviewed by artistic experts. In such collaborative way by expert guidance and machine intelligence, Dunhuang Academy can ensure high quality virtual restoration. The virtual restoration started from easier challenges rather than difficult ones. The early attempts focused on restoring the lost content in deteriorated regions with simpler structures

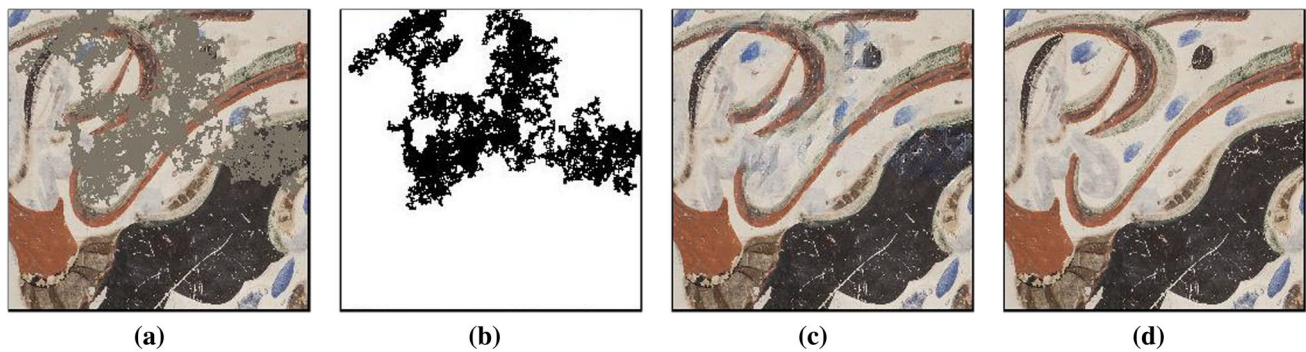


Fig. 6 Inputs, output, and groundtruth of the virtual restoration method: **a** simulated deteriorated e-Heritage image; **b** simulated deterioration mask; **c** restored result; **d** groundtruth

and textures. Restoring a complex structure with rich details (e.g., faces of Buddhas) is challenging. Figure 6 shows the examples of an input and an output for the trained inpainting model. The expected virtual restored results are required to be locally and globally consistent and aligned with human perceptions. Specifically, the color should be correctly in-painted, and the new textures should be similar to the surrounding well-maintained regions, and viewers should feel natural about the line segments and structure of the in-painted objects.

At present, Dunhuang Academy is working on virtual restoration for the simple deteriorated regions using a deep learning model with partial convolutional modules (Nazeri et al., 2019a). During this initial attempt, we found that the structure of the output result may not fit the context well in some complex cases, or the generated edge may not agree with the understanding of experts. To solve this problem, Dunhuang Academy proposed to use the EdgeConnect (Nazeri et al., 2019b), which firstly generates optimal structural edges that perceptually fit well in the local context, and then in-paint color and textures.

1.3.2 Heritage Property Retrieval

The demand for heritage property retrieval or automatic content understanding in archeology is high.

To meet these practical demands, Dunhuang Academy is trying to use object detection technology to build up an intelligent retrieval model; the model can detect the pre-defined elements and classify some patterns using multi-modal data. In other words, if an archaeologist labels a target object in an image and provides some archaeological information as prior, the an intelligent retrieval model should find out the similar elements in the same image and the given related images. Multi-modal data (e.g., visual information, dynasty label, and descriptive text), which are key characteristics, can be used to describe the target objects.

Currently, the pre-defined objects are limited to common elements. In the future, a special detector can also be built for specific objects, and it can be used as an analyzer to search these objects in all mural paintings. The understanding of the local regions enable the archaeologists to locate similar patterns/elements in a mural painting or cross different paintings, which can deepen the understanding of association of the elements and may lead to new findings. However, we still face some challenges. First, the elements in e-Heritage are naturally unbalanced in number. Elements in some categories have more samples, while some have quite few. Another insight of the unbalanced data is that some elements exist across multiple mural paintings, while some elements only appear once in a single mural painting. Second, some elements can be highly non-rigid in the e-Heritage. The objects in same categories may dramatically change their forms and appear in different shapes with diverse artistic style, causing the objects in the feature space to be sparse. The apsaras are the cases in point. The apsaras were usually painted in a romantic style with a spanning shape and long decorated ribbons in various bright colors to render the apsaras different from ordinary humans while share some humanoid characteristics. Third, labeling data by the professional with domain knowledge is costly. Relatively fewer available annotations causes the retrieval model prone to over-fitting, resulting in poor test time performance. Finally, noises and deterioration can affect model performance. The e-Heritage covers almost all the mural paintings, including some seriously deteriorated ones. How to handle these samples from deteriorated paintings is a challenging decision.

In some complex cases, a light-weight CNN model with a fine-tune routine and few-shot learning is also considered. The archaeologists can label a new category of objects and output the similar object in the order of similarity rankings. The model should be able to quickly learn from limited labeled data. In another case, new data are still acquired using updated instruments, and the continuous labeling is costly.

Few-shot learning aims to provide a more automatic labeling process.

1.3.3 Artistic Style Transfer

How can the present life incorporate the extracted artistic styles? The answer can be style transfer. A typical example is that an ordinary photo captured in daily scene are cast-ed with a Dunhuang e-Heritage style. The input photo provides the content (i.e., structural information), while the set of e-Heritage image implicitly contains the style (i.e., summary characteristics). The output photo is a fusion of the rendered e-Heritage style and the photo content.

Dunhuang Academy is working on fine-grained style definition. The artistic styles are defined by experts using knowledge of history and art. These styles are clusters of image with similar function in painting, scene, or era. After a style is well-defined, data are collected to form a style dataset. The state-of-the-art CNN-based artistic style transfer methods for digital image are used to extract the summary of characteristics from the training data in the Dunhuang style domain, and the style are casted onto images from another domain (i.e., daily photos).

There are main challenges in this task. First, there is no existing paired data. The scenes in the ancient paintings have no corresponding counterparts in the other domain. Second, the artistic style should be pre-defined by human experts in history and art, who do not have knowledge in machine learning. The cost of exchange is high in a building dataset. Third, the evaluation criteria and quantitative analysis on the results are not well established and sometimes require human visual inspection and subjective perceptions during result evaluation. To address the problem of unpaired dataset, Dunhuang Academy proposed to use the unpaired image-to-image style transfer methods(e.g.,Cycle-GAN) (Zhu et al., 2017).

1.4 Structure of this Paper

The rest of the paper is organized as follows. In Sect. 2, we introduce technologies and projects that are related to heritage protection, with particular focus on recent AI-based technologies. In Sects. 3 to 5, we present how Dunhuang Academy is using AI-based technologies to propose solutions for Dunhuang heritage protection. Each section covers one of the three major tasks, namely, virtual restoration, visual heritage property retrieval, and artistic style transfer, respectively. In each of these sections, we introduce technical details about the formation of the datasets, the proposed baseline methodologies, implementations and experiments, and comparative analysis and discussion on the results. The final Sect. 6 concludes this paper.

2 Related Works

2.1 Virtual Restoration

Image restoration methods are divided into two types. One type is based on traditional image processed techniques, while the other type is based on deep learning methods. The latter is current advance and research focus.

Conventional methods are based on the diffusion techniques. Given the to-be-restored holes, the diffusion-based methods (Ballester et al., 2001; Bertalmio et al., 2000; Levin & Weiss, 2003) in-paint the holes by propagating the local contextual visual information from the surrounding peripheral regions. These methods focus on how to propagate semantic information, e.g., extending the isophote direction field (Ballester et al., 2001; Bertalmio et al., 2000), or relying on statistical features or visual features (Levin & Weiss, 2003). These conventional methods can only properly in-paint small regions and simple texture. Accordingly, their applications are limited. Some methods are used in dust removal task in film scanning and only coarsely in-paint small regions with molds or thin scratches. The more sophisticated patch-based methods outperform the diffusion-based methods. The patch-based methods can restore larger deteriorated regions with better local consistence in texture. The first patched-based method (Efros & Leung, 1999) was proposed by Efros et al., who used a texture synthesis framework for a novel copy-paste scheme. The copy-paste scheme searches possible patches from images in a source dataset, retrieves the patch with highest similarity, and pastes the patch into loss region in the target image. Later, the patch-based framework is employed by several methods (Barnes et al., 2009; Huang et al., 2014; Kwatra et al., 2005; Simakov et al., 2008; Wexler et al., 2007), and these methods further introduced optimization process to smooth the inconsistency between in-painted and original textures. Particularly, PatchMatch (Barnes et al., 2009), proposed by Barnes et al., increases the processing speed to the sub-real-time level by greatly reducing the computational cost. A major disadvantage of patch-based methods is that these methods have difficulties in restoring textures or structures if proper matching occurs between the restoring regions and the candidate patches in a prior knowledge dataset. Rather than synthetically generating a learned texture, these patch-based methods rely on matching local pixels or their texture features. The in-painting content is associated to a set of given template images, which is a library where candidate patches are fetched from. Thus, if the matching does not occur, or the measurement is not suitable for the pair-wise match, then the restored results can be unrelated to the local context.

The state-of-the-art image restoration techniques are driven by the statistical learning-based methods. In particular, convolutional neural network (CNN)-based methods,

which have a set of new base-lines in many signal processing areas, are the most advance in recent years. Early methods (Pathak et al., 2016; Yang et al., 2017) in this category focus on in-painting a rigid rectangular blank region at the center of a target image. Pathak et al. (2016) first proposed a Context Encoder, which uses an asymmetric end-to-end CNN. The input is an 128×128 image with a large blank region, while the output is an estimated 64×64 restored content for the blank region in the input image. The Context Encoder encodes visual information of non-blank region and maps it onto the groundtruth content underlining a blank region at the center by taking the advantages of the powerfully feature embedding capability of the CNNs. Yang et al. (2017) introduced a multi-scale neural patch synthesis into the Context Encoder by adding a joint optimization of the image content and texture constraints. The improvement results in finer semantic details in the restored results. Song et al. (2018) proposed a stacking multi-scale inference and used a Patch-Swap operation to optimize the semantic texture and post-process restored regions. It too rigid and strong to assume that the images are mostly corrupted by a rectangular patch in the center. Iizuka et al. (2017) and Yu et al. (2018) dropped this concept and proposed the assumption that the restoring regions can be in non-rigid shapes. The new assumption is more flexible and similar to daily corruptions in images. The regions to be restored are generally given in the form of mask. One of the side-product advantages is that these methods reduce the risk of over-fitting the rectangular blank regions. Iizuka et al. (2017) used a generative adversarial network (GAN) framework with two discriminators. The discriminators try to correctly classify the local restore texture and global generated image, and the wrongly classify results are used as GAN loss for updating the generator. Yu et al. (2018) extended the GAN-based method (Iizuka et al., 2017) by integrating an attention mechanism. Lately, (Liu et al., 2018) proposed Partial Convolution. The partial convolutional layer is designed to gradually fuse image and the mask for restoring regions in down-sampling process of the encoder. In the decoder, the merged low-level features are up-sampled onto an output image. The local in-painted texture in the output image is further extracted and merged with undeteriorated texture to form a final restored image. Nazeri et al. (2019b) proposed EdgeConnect, which consists of two generator models. One model is to predict the structural information in the restoring regions (i.e. edge map). Meanwhile, another generator uses the edge map as the with guidance to in-paint the color and texture.

2.2 Historical Property Retrieval

A good object detector can improve the efficiency of historical property retrieval. Here, we mainly review some works related to object detection. Object detection is a computer

technology on computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or vehicles) in digital images and videos. Well-researched domains of object detection include multi-category detection, edge detection, salient object detection, pose detection, scene text detection, face detection, and pedestrian detection. Object detection, as an important part of scene understanding, has been widely used in many fields of modern life, such as security, military, transportation, medical, and life fields. Furthermore, many datasets have appeared and played an important role in the object detection field to date, such as Caltech (Dollar et al., 2011), KITTI (Geiger et al., 2012), ImageNet (Russakovsky et al., 2015), PASCAL VOC (Everingham et al., 2010), and MS COCO (Lin et al., 2014).

The existing domain-specific image object detectors can be divided into two categories. One is the two-stage detector, such as Fast R-CNN (Girshick, 2015). The other one is a one-stage detector, such as YOLO (Redmon et al., 2016) and SSD (Liu et al., 2016). Two-stage detectors have high localization and object recognition accuracy divided by an ROI (region of interest) pooling layer, whereas the one-stage detectors achieve high inference speed. For instance, in Faster R-CNN (Ren et al., 2016), the first stage, called RPN, a Region Proposal Network, proposes candidate object bounding boxes. In the second stage, features are extracted by RoIPool (RoI Pooling) operation from each candidate box for the following classification and bounding-box regression tasks. Furthermore, the one-stage detectors propose predicted boxes from input images directly without region proposal step; thus, they are time efficient and can be used for real-time devices.

2.3 Artistic Style Transfer

Artistic style transfer is to cast the summary of artistic characteristics of an image or a set of image on to target image without affecting the semantic content of the target. The traditional style transfer algorithm (Hertzmann et al., 2001; Portilla & Simoncelli, 2000) for each style image manually establishes a mathematical or statistical model and changes the image or video frame to be migrated to make them better conform to the model. Researchers began to combine neural networks with exploring new style representation and transfer algorithms. Gatys et al. (2016) first applied the deep neural network to the field of style transfer. The current research of neural style transfer method is mainly divided into two directions. The first one is the slow online image optimization method based on sum statistics or nonparametric Markov random fields (MRFs). The other one is the fast offline model optimization method based on single style single model, multi-style single model, and arbitrary style single model. The outstanding feature of a single style single model fast neural algorithm is that each model can only carry

out the migration task of a single style. Although the single style and single model method can realize the rapid stylization process, each model can only target a specific style, and the network needs to be retrained for other styles. The flexibility of the model is very poor, which will consume a lot of time and computing resources. Some methods based on intermediate feature transformation (Li et al., 2017; Lu et al., 2019; Yoo et al., 2019) were proposed to solve this problem. The main idea is to train a network of coders and decoders to reconstruct the content image. Then, the features of the content image obtained by the coder were given some features of the style image through some intermediate feature transformation operations. Subsequently, the style transferred image style transfer was reconstructed through the network of decoders.

3 Virtual Restoration

Virtual restoration technology helps artists in repairing the decaying regions of the e-Heritage images, which can be better presented in public exhibitions or artistic re-creation. In this section, we introduce how a dataset for virtual restoration for Dunhuang e-Heritage is generated. The process includes the generation of two types of masks, which simulate two common deteriorations in the heritage (i.e., the decaying and physical damages). Two state-of-the-art methods are comparatively studied and vigorously tested using the proposed dataset.

3.1 Dataset for Virtual Restoration

Given that virtual restoration is an ill-formed task that groundtruth texture/color of the deteriorated regions no longer exists, we use the mostly intact images or regions along with the simulated deterioration as training images. The training samples consist of two items: an intact image and a mask. The mask marks the region of deteriorated content, and it is used to tell the machine learning model that the content of the underlying pixels is no longer available.

The datasets consist of intact images and the masks, which are used as groundtruth and to simulate deteriorated samples. Building such a dataset for virtual restoration that is similar to the practical task is difficult. On the one hand, deteriorated regions might be caused by a variety of factors and have a variety of forms, therefore the corresponding simulated masks should be similar to deterioration. On the other hand, the images should be of high quality, with few sounds and enough visual information. This visual information would be embedded in the model. Thus, its information distribution should be consistent with the targeting test set, which contains deteriorated images and regions to be restored. However, it is unavoidable that the virtual restoration model will

learn noises from the e-Heritage data, which always contain some noise. The archaeologists are employed to determine if the area of paintings can be used for training set. This manual evaluation includes the type, density, and level of deteriorations. The selected images go through a preprocess to generate sample patches. The preprocess includes random cropping and resizing. The heavily deteriorated patches are useless for training and are manually filtered, further reducing noises in the training set. The validation set is generated in the same manner as the training set.

Two types of masks that simulated the different types of deteriorations are generated for evaluating the performances of restoration results and the similarity of generated content to the groundtruth. Here, we called them dust-like and jelly-like masks according to their morphological characteristics. The dust-like masks simulate deterioration from molds or salting erosion, while the jelly-like masks simulate physical damages or sabotages. The dust-like masks are generated through following steps: Step (1) A canvas is initialized as a square blank image by setting all its pixels in white color, i.e., of value 1. This canvas used in drawing pixels show deteriorated regions. The canvas size is 256×256 . Step (2) We randomly pick a start point P_0 on the canvas. Step (3) We perform a random walk in an iterative way. Once a pixel is walked on, its value is set to 0. A pixel is allowed to be walked on multiple times. The default number of walk steps is 10,000. The latter type of jelly-like mask is generated based on the dust-like masks. We further remove small noises and reserve the irregular block-like regions using open-close functions and image erosion.

3.2 Methodology

An end-to-end neural network, which outputs a map containing the synthesized content in the corresponding regions without additional post-processing, must be created to automatically in-paint the deteriorated regions with synthesized content. The end-to-end neural network takes in the original image and the mask at one end. Then, a predicted image is outputted as the same size of the original image at the other end. The method avoids hand-designed pre-processing optimization for the restored semantic content using the powerful capability of regression and prediction of CNN. Let I_{input} , M , I_{out} , and I_r be the original input image, mask, output image from the networks, and final restored image. Supposed that the width and height of the network input are N , and the input color image I_{input} with regions to be restored is in the size of $N \times 3$. If the input image is not square, then it will be re-scaled to be square to ensure that the ratio is consistent with the network input. The mask M , which is of size $M \times N$, labels whether the pixels belongs to deteriorated or intact regions. The labeled deteriorated regions are generally in irregular non-rigid shapes. The output image contains tex-

tural and stylistic information of both the labeled deteriorated and the unlabeled regions. The end-to-end network serves as a function that maps from an input color image and a mask to an output color image of same size: $f : (I_{input}, M)_{out}$. The final restored image I_r , containing both original intact content and synthesized content, is the pixel-wise combination of I_{input} and I_{out} . Given the mask M , restored image I_r is computed as $I_r = I_{input} \circ M + I_{out} \circ (1 - M)$, where \circ is the pixel-wise multiplication. The key of this process is to find out and train an effective end-to-end neural networks $f(\cdot)$, whose output content of masked regions has minimal perceptual difference from the groundtruth. In following subsections, we will introduce the models, implementation details, and experimental results of two baseline models: (1) PConv and (2) EdgeConnect.

3.2.1 Baseline 1: PConv

We take the advantages of the U-like end-to-end network with skip connections and partial convolutions layers (PConv) (Liu et al., 2018) to generate the synthesized content for deteriorated regions of the Dunhuang Grottoes. The network architecture, a variant of encoder-decoder configurations, is shown in Fig. 7 (Yu et al., 2019a). The encoder and decoder are not mirrored in a symmetric structure like the conventional U-nets (Ronneberger et al., 2015). The ReLU activation function is used in each layer in the encoder except for the first layer. Between layers, the batch normalizations are used on the feature maps. The decoder has similar structure to the encoder but are in reverse order. The skip connections directly pass feature maps from the i th layer of encoder to the $(8-i)$ th layer of decoder.

The key components in a partial convolutional network are the partial convolutional blocks, which make use of the down-sampling to fuse the mask into the deep features. The mask serves as a conditional inverse attention for the networks. The mathematical formula of partial convolution operation is given as follows:

$$x'' = \begin{cases} W^T(X' \circ M') \frac{1}{sum(M')} + b & \text{if } sum(M') > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

where X' and M' are the input feature maps and mask at the entry of a partial convolutional block, respectively; \circ is the element-wise multiplication; W_T and b are the weight and bias of a filter, respectively; and x'' is the output value of partial convolution. When being passed down in the encoder, the mask gradually decayed by merging with the neighboring regions in each layer. The mask is updated with a decaying process as:

$$m'' = \begin{cases} 1 & \text{if } sum(M') > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Each partial convolutional block down-samples the feature maps and the mask. When the mask reaches the bottleneck, the value in the mask will be all ones, indicating that all the masking information has been fused into the embedding low-level features. During the down-sampling and decaying, the decayed mask and partial convolution not only smoothen the feature maps but also fill the vacant regions in the subsequent feature maps, in which is of all zeros in the first layer. The overall loss for training the proposed end-to-end partial convolutional neural networks is linear combination of multiple loss terms that take account of different considerations, including content differences, style differences, and smoothing constraint. The overall loss is given as follows:

$$L = \lambda_{content} L_{content} + \lambda_{style} L_{style} + \lambda_{TV} L_{TV}, \quad (3)$$

where $L_{content}$, L_{style} , and L_{TV} are the content loss, style loss (Johnson et al., 2016), and total variation (TV) loss, respectively; and $\lambda_{content}$, λ_{style} and λ_{TV} are the balancing coefficients for the corresponding loss, respectively. With regard to the content and style losses, we employ VGG-16 ImageNet pre-trained network (Simonyan & Zisserman, 2014) as the loss network ϕ to extract the deep feature maps.

3.2.2 Baseline 2: EdgeConnect Nazeri et al. (2019b)

The fundamental idea of EdgeConnect is to use a CNN model to predict the edge or contours of objects (Xie & Tu, 2015) (Canny, 1986) in the deteriorated regions, and then, use another CNN model to in-paint the color and texture with guidance of the predicted edges. The summary of the EdgeConnect method is presented in Fig. 8. The method contains two major generator models, namely, G_1 and G_2 . G_1 is the edge generator, while G_2 is the in-painter. The inference process can be given as follows:

$$I_{out} = G_2(I, M, C|\theta_1), \quad C = G_1(g(I), M, c(I)|\theta_1), \quad (4)$$

where G_1 predicts the edges or object contours C in the masked regions given grayscale image $g(I)$, mask, M , and current edge $c(I)$. Function $g(\cdot)$ calculates the grayscale image from RGB image, and $c(\cdot)$ computes the edges on unmasked region. C is the new edge map with an estimated edge in the deteriorated regions.

A combination of the different loss functions is used during the training to properly constrain the output. A hinge variant of GAN loss and feature-matching loss are employed for training the edge generator G_1 . The feature matching uses the VGG pre-trained model (Simonyan & Zisserman,

Fig. 7 Architecture of the PConv networks

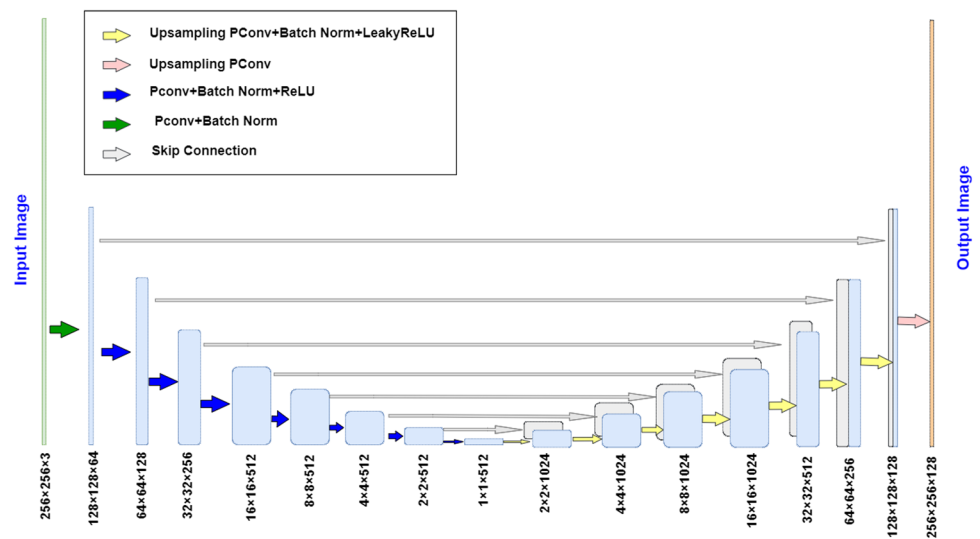
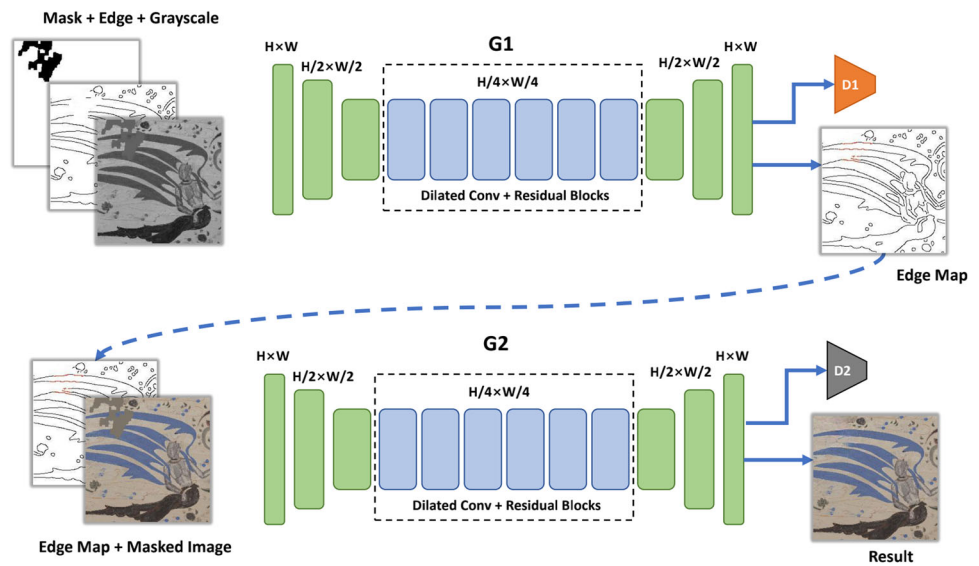


Fig. 8 Architecture of the EdgeConnect baseline



2014), and the feature matching is based on the deep features that extracted from the VGG model. More constraints are considered when training the in-painter G_2 because of the complex nature of the measuring difference of the content in the images. The hybrid loss for G_2 consists of L_1 loss, the hinge loss L_h , perceptual loss L_{prec} , and style loss L_{style} (Johnson et al., 2016), which is given as follows:

$$L_{G2} = \lambda_1 L_1 + \lambda_h L_h + \lambda_{prec} L_{prec} + \lambda_{style} L_{style}. \quad (5)$$

3.2.3 Implementation

The tested methods are implemented in PyTorch. The input and output sizes of the end-to-end networks are 256x256. The images of the different sizes and scale ratios are re-scaled to fit the input size of the network. The implemented method is trained and tested on X86 PC powered by Intel i5 CPU@3.7

GHz, 16 GB RAM, Ubuntu 16.04 OS, Nvidia GTX Titan Xp with 11 GB memory.

We use two stages to train the two baseline end-to-end networks: (1) the first stage is to pre-train a partial convolutional network for diverse low-level feature-extracting capability. The pre-training stage is carried out on the Place2 dataset (Zhou et al., 2018). (2) The second stage fine-tunes the pre-trained model to fit in our grottoes restoration task. The pre-training allows the network to generalize its low-level filters on a diverse dataset to ensure that it could extract various deep features for the latter stage. The Place2 dataset contains 10 million images, which cover numerous different kinds of texture. Given that the Place2 dataset is large enough to contain diverse visual information, data augmentation is not used in the first training stage; and all training images are randomly sampled from the huge Place2 dataset. In the second training stage, data augmentation techniques are used

in pre-processing the input images to avoid over-fitting the style and dynamically generate more training samples. The augmentation includes random vertical flip, random horizontal flip, random 90-degree rotation, random change of saturation, random adjustment of Gamma value, and random adding Gaussian Noise. The loss value is back-propagated through all parameters in the network in the first and second stages. The parameter settings for training the two baselines are as follows.

3.2.4 PConv

In the first training stage, the partial convolutional network is trained on the Place2 dataset (Zhou et al., 2018) to obtain a pre-trained model. During pre-training, we use an Adam as optimizer and set the learn rate to $2e-4$. The size of the mini batch is 16. The weighting coefficients $\lambda_{content}$, λ_{style} , and λ_{TV} for the corresponding loss functions are set to be 0.05, 1000, and 0.1, respectively. In the second training stage, the network is fine-tuned by fixing some weights in low level filters. Specifically, the parameters of batch normalization layer in the encoder of the network are frozen and no longer updated. The parameter settings of fine-tuning are similar to those in the pre-train stage except that the learn rate decreases to $5e-05$.

3.2.5 EdgeConnect

The pretrained model is obtained directly from the public resources, and it is trained by the author using the Place2 dataset. The training involved 2 steps, each of which has different learning rates and interactions between the G_1 and G_2 models. In the second stage, we freeze the parameters

in the edge generator G_1 , while fine-tune the performance of G_2 with a learning rate of $1e-5$ on a mixed dataset. The mixed dataset is a combination of the training set from Dunhuang and place2 dataset with 0.5 sampling probabilities for each subset.

3.3 Experiment

In the experiments, the implemented methods are tested on image samples cropped from raw Apsaras data, which is delicately photographed and provided by Dunhuang Academy. After generating Apsaras dataset, we conducted two experiments for the two baselines. Then, we conducted quantitative and qualitative comparison on the results of the two kinds of masks.

The results from the two types of masks are evaluated by four different criteria: (1) total variation (TV), (2) peak signal-to-noise ratio (PSNR), (3) structural similarity (SSIM), and (4) HaarPSI. Among these criteria, total variation only accepts single image input, while others are pair-wise metrics that measure similarities between results and their corresponding groundtruths. PSNR indicates the local statistical similarity; SSIM measures the local structural similarity; and HaarPSI measures the perceptual similarity in wavelet domain. The TV metric only indicates the a characteristic of the image and is not related to quality of the image. Meanwhile, a higher value in other three metrics means better results.

Figures 9 and 10 present the comparison results using dust-like and jelly-like masks, respectively. The results of PConv and EdgeConnect are aligned with the GT for easier comparison. The restored regions are similar to the groundtruth in terms of style and color. Although the in-painted texture is

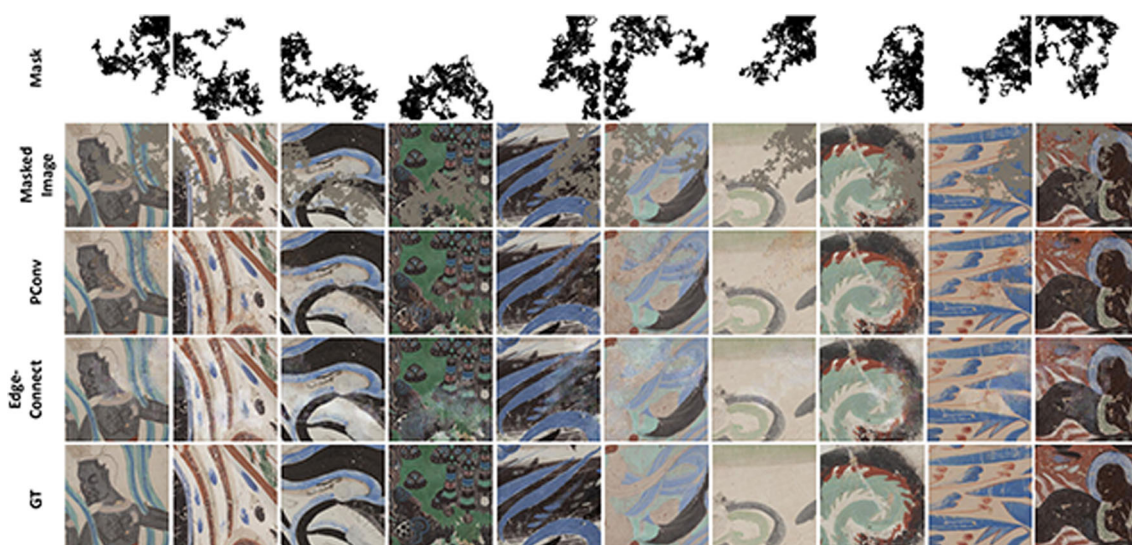


Fig. 9 Virtual restoration results on the Apsaras test set with dust-like masks

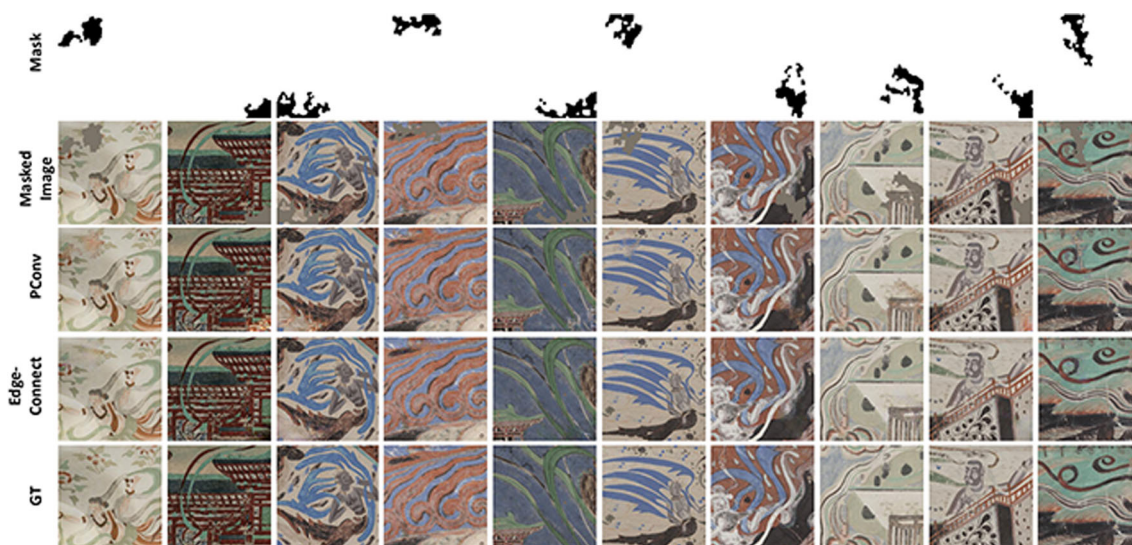


Fig. 10 Virtual restoration results on the Apsaras test set with jelly-like masks

different from the original content to a certain extent, the in-painted texture is locally and perceptually consistent. In Fig. 9, which presents results restored from the dust-like masks, we can notice the different behaviors from the two comparing methods. EdgeConnect seems not able to restore the tiny single-pixel deteriorations by leaving the white dots in the result image. PConv, which seems able to restore tiny deteriorations, has some noticeable local inconsistency that some pixels are not well blended in the region in color.

In Fig. 10, the jelly-like mask poses more challenges for PConv than the EdgeConnect. Although PConv in-painted more details, certain artifacts unrelated to the local regions tends to exist in some results, which disagree with the groundtruth. Moreover, the boundaries of segments are not well restored. Meanwhile, the EdgeConnect in-paints better texture. Although some edges are generated not sharp enough compared with the groundtruth, the perception of these restored regions are real, and it is good enough to cheat human eyes. In both methods, the details may not be sufficient to fully recover the large block of masked regions.

The proposed methods can be adopted in practice as assistive software and work together with human artists. From the perspective of e-Heritage protection, we conclude that the computer vision-based methods can restore decayed contents.

Tables 1 and 2 present the quantitative results of the two different masks under different performance criteria. The results are from Apsaras testset. It shows that the two baselines almost have opposite performance on the same test set but with different types of masks. The result data are consistent with the qualitative results. In Table 1, the EdgeConnect has much higher TV value due to the white dots, which are pixels that are not successfully interpolated. In the

Table 1 Quantitative results on the Apsaras test set with dust-like masks by different criteria

Method	PSNR	TV	HaarPSI	SSIM
EdgeConnect	21.721	23602.29	0.6829	0.6497
PConv	24.850	17443.60	0.7325	0.7410

Table 2 Quantitative results on the Apsaras test set with jelly-like masks by different criteria

Method	PSNR	TV	HaarPSI	SSIM
EdgeConnect	28.169	17960.88	0.8710	0.8301
PConv	27.843	17318.04	0.8568	0.8404

other measured results on the dust-like mask, PConv shows better performance than EdgeConnect because of better local consistency. In Table 2, the value of TV metric indicates more local variation in the restored regions of results of the PConv than that of EdgeConnect. This is because the PConv generates some artifacts in the restored regions, which has higher value variation. In the HaarPSI metric, EdgeConnect outperforms PConv, which indicates a better perceptual similarity to the groundtruth. We review the qualitative results and found that the correct edge prediction might contribute to the higher performance.

3.4 Results on Real Deteriorations

To show the effectiveness for practice, we prototype a virtual restoration software and conducted a test on real deteriorated e-heritage data using EdgeConnect model and jelly-like

mask. Shown in Fig. 11, a Tkinter-based graphic user interface (GUI) is designed to interact with the artist for importing image, obtaining mask, and showing and comparing in-painted result. After importing an image, we employ an interactive segmentation method (Sofiuk et al., 2020) for semi-automatically segment the deteriorating regions. The interactive segmentation method only required the artist to give some cues about where a deteriorated region is with 1–3 mouse clicks. Comparing to manual drawing the mask, this semi-automatic process for obtaining deteriorated regions saves much time for artists, especially when there are many smaller scattered physical damages in the image. The mask can be saved or used as input for inpainting method. The output of the interactive segmentation is morphologically very similar to the jelly-like mask. After invoking the inpainting methods, the final result is showed side by side with the original image.

Figure 12 shows some of the restoration results on real deteriorations using EdgeConnect trained with jelly-like masks. The segmented masks that mark physical damages are consistent with actual deteriorations. It is easy to observed that the in-painted textures on the small deteriorated regions is better integrated with the surrounding region, compared to the large deteriorated regions. For the large deteriorated regions, we find that the inpainting model has difficulties in predicting the structural information in the large deteriorated regions. This is because larger deteriorated regions and their external contexts are less related, and the inpainting model cannot find out the structural connection between them. As there is no groundtruth for real deteriorations, we can only judge results with subjective reasoning and feeling. Most of virtual restoration results on real deteriorations give us a better overall feeling compared to the original image.

4 Heritage Property Retrieval

Object detection method is the core technology for Dunhuang property retrieval. On the one hand, retrieving properties can facilitate archeologists in searching properties across different paintings by automatically locating the target objects in the large e-Heritage images, and help archeologists in investigating target objects by placing them side by side. On the other hand, the object detection method enables heritage protection institute to obtain statistical information of properties in e-Heritage, and similarity matrix built up for a category allows people to understand the heritage from a global view. In this section, we will introduce the built up of a dataset for Heritage Property Retrieval, the methods that benchmark, implementation details and evaluation settings, and finally, the experimental results and discussion.

4.1 Dataset for Heritage Property Retrieval

The murals of Dunhuang Grottoes have a large scaled area with substantial contents, which consists of various subjects, for example, apsaras, donors, animals, inscriptions, bridges, music instruments, pagodas, trees, and thousand Buddhas., as shown in Fig. 13. The corresponding quantity of the above subjects varies from tens, hundreds, thousands, or even tens of thousands.

The image identification of Dunhuang mural dataset is constructed by selecting five types of representative subjects, which have large number of amounts, easily recognized pictures, and high attention, following the process steps of the whole wall digital image acquisition (as show in Fig. 14), random cutting, object annotation, and others.

Figure 14 (a) shows the images of Thousand Buddhas in the north slope of the main chamber of Mogao cave No. 237. (b) The paintings of “Mount Wutai” from Mogao Cave No. 61, one of the existing paintings of Mount Wutai with the largest area, the most complex content, the richest perspective, and the most detailed cartouche, has the size of 13.45×3.42 (in meters) with a total area of around 46 m^2 and 195 inscriptions to explain the details of the paintings of “Mount Wutai”. (c) Apsaras are popular mural elements, and the most representative of Sui Dynasty apsaras are selected from Mogao cave No.305. (d) There are nearly two thousand donors in Mogao cave No.428, which locate in the middle of the south wall, west wall, and north wall. (e), (f) In addition to Mogao Grottoes, there are also abundant murals in many other cultural heritage sites, such as the east wall of the main chamber of the Yongle temple in Shanxi Province and Pulu temple in Hebei Province, with different non-Buddhist figures in mural. A large number of subjects lay the data foundation for the construction of the dataset.

Each digital mosaic image of above murals with 30dpi resolution is randomly cropped by using fixed size rectangular areas (600×600), and the cropped images are used to eliminate the image without Thousand Buddhas and inscription within the cutting range.

At the end, the above five chosen subjects in the cropped images are tagged with rectangle box (e.g., Thousand Buddha is “Thousand Buddha motif”, list cartouche is “instruction”) by experts from Dunhuang Academy, and the label file was saved in XML format. Each sample in dataset consists of one image and one label file. The label for the detection task contains the bounding boxes and the corresponding categories. Some examples in the dataset with visualized label are shown in Fig. 15.

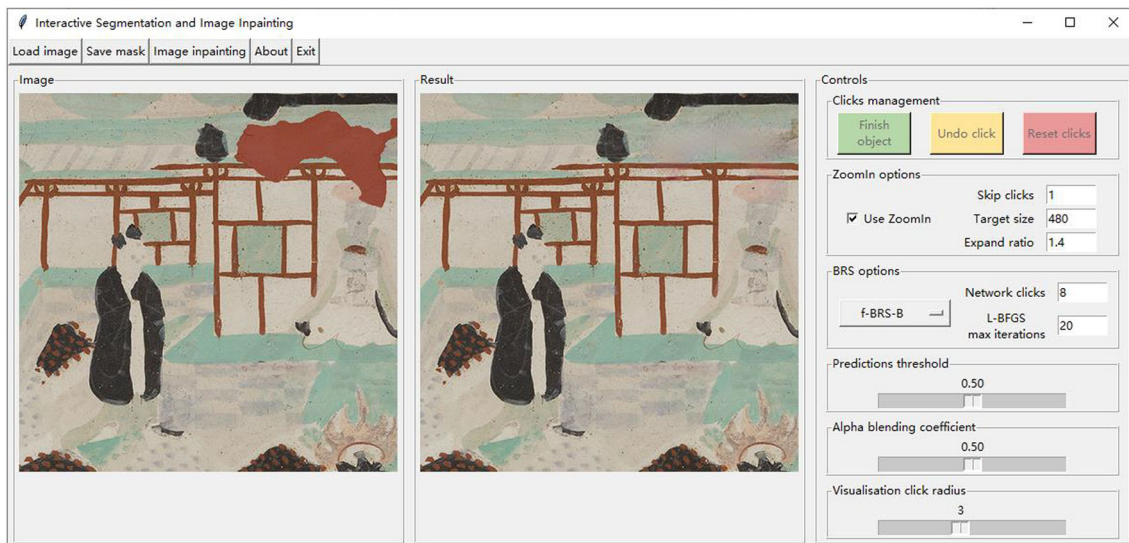


Fig. 11 GUI of the virtual restoration software which integrates interactive segmentation and inpainting method

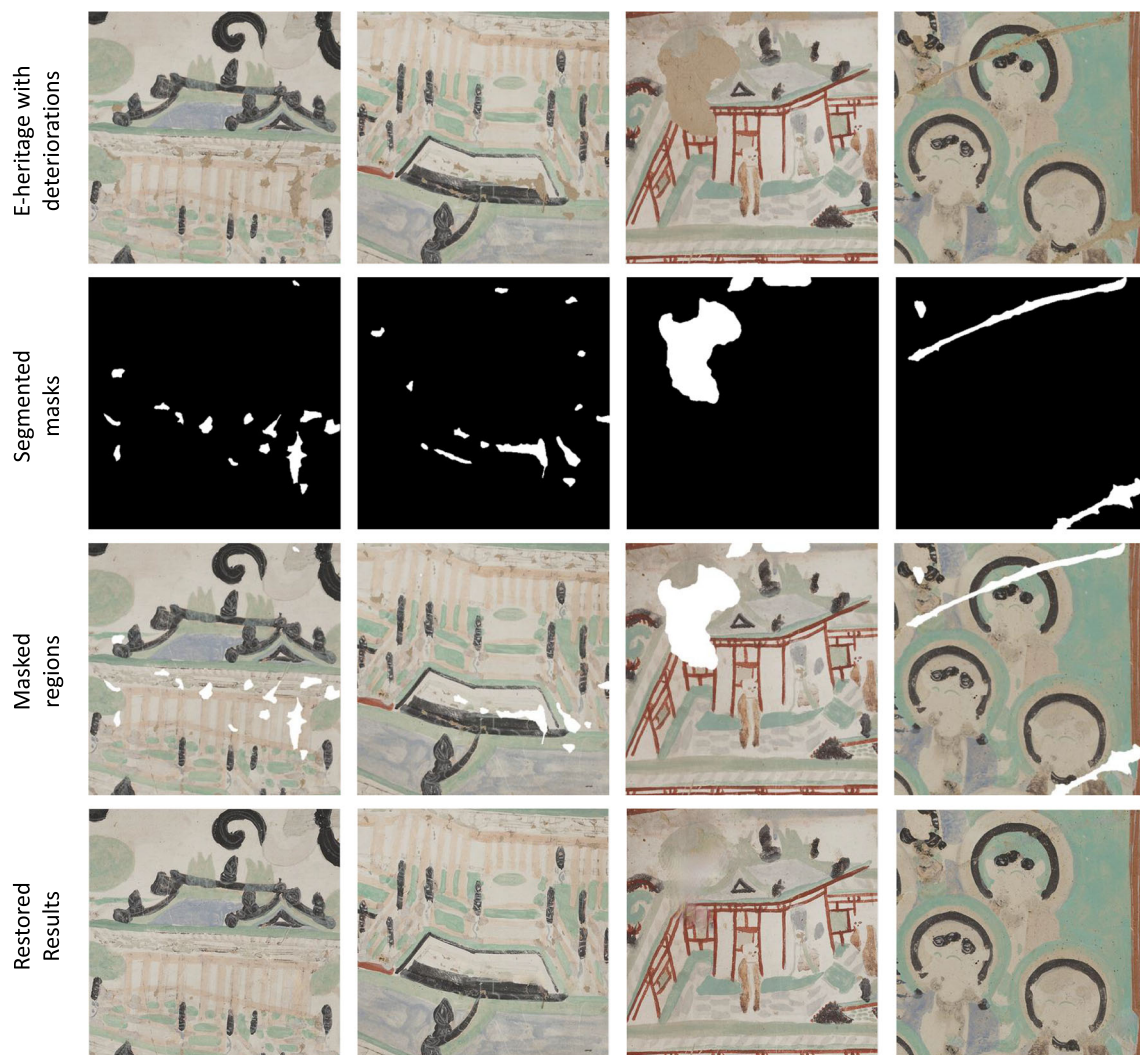


Fig. 12 Some restoration results on real deterioration using EdgeConnect with jelly-like masks



Fig. 13 Examples of objects that frequently appeared. Although some of them have relatively stable shapes and structures, others may significantly vary

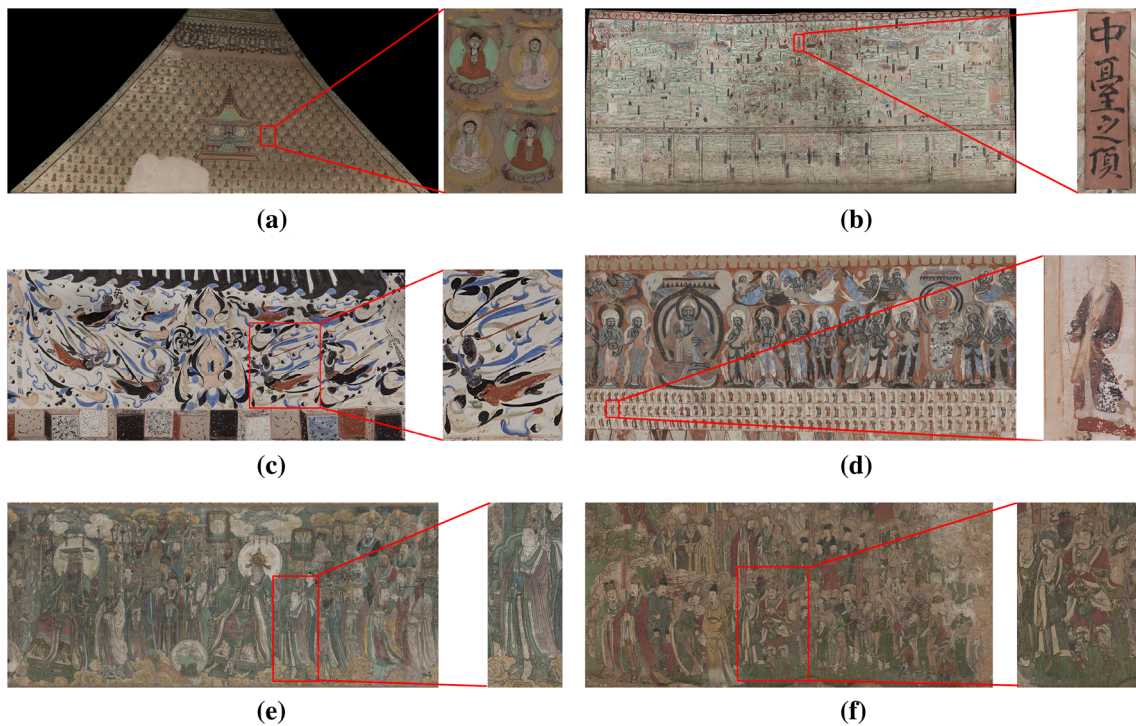


Fig. 14 Examples of selected grand murals by artists. Each of the mural may contain 10–3000 objects. The proposed detection dataset are created by manually selecting objects from them



Fig. 15 Examples from the Dunhuang object detection dataset. From top to bottom are dataset samples from categories of Apsaras, Inscriptions, Donors, Thousand Buddha Motif, and Non-Buddhist Figures, respectively

4.2 Methodology

We use the object detector as the key technology to implement heritage property retrieval applications. The Dunhuang Academy designed a top-down pipeline for the property retrieval on very large e-Heritage images and tested multiple state-of-the-art object detectors on the Dunhuang Object detection dataset. In the future application, the test patches will be sampled by a sliding window with steps of half of its width and height given a very large image. The location of each patch is recorded. After all test are inferred, the

local bounding boxes are projected back on the original large image.

We first tested eight models on the dataset to determine the detector with good performance. These test detectors are SSD300 (Liu et al., 2016), RetinaNet (Lin et al., 2017, 2020), GHM RetinaNet (Li et al., 2019; Lin et al., 2017, 2020), FSAF (Zhu et al., 2019), Faster RCNN (Girshick, 2015), Libra Faster RCNN (Girshick, 2015; Pang et al., 2019), Libra RetinaNet (Lin et al., 2017, 2020; Pang et al., 2019), and YOLO v5-S (Jocher et al., 2020). We used the classic SSD300 and Faster RCNN for the baselines in eval-

uating performance for this experiment. RetinaNet, GHM RetinaNet, FSAF, Libra Faster RCNN, and Libra RetinaNet used the popular Resnet-50 (He et al., 2016) as the backbone, which balanced performance and computational cost. We employed the improved version of RetinaNet and FSAF in our experiment. The improvements include the integration with Gradient Harmonized Mechanism (GHM) (Li et al., 2019) and Libra framework (Pang et al., 2019). The GHM is utilized to optimize the gradients during back-propagation and effectively use of the gradients. The Libra framework (Pang et al., 2019) is an effective framework to solve the class imbalanced in the dataset, which is not a serious and common issue for the dataset in the cultural heritage field. YOLO v5-S is a version of the YOLO v5 model, where the S indicates a small model with fewer parameters.

4.3 Implementation

The experiment is carried out on a regular x86 PC with Intel i5-8600 CPU, 16 GB RAM and a NVIDIA Titan Xp graphic card with 12 GB memory. We set the learning rate as the default value. The epoch value is 16, which is sufficiently large for convergence for our dataset. Eight methods are implemented and tested on the proposed e-Heritage dataset. Among these methods, SSD300 (Liu et al., 2016), RetinaNet (Lin et al., 2017, 2020), GHM RetinaNet (Li et al., 2019; Lin et al., 2017, 2020), FSAF (Zhu et al., 2019), Faster RCNN (Girshick, 2015), Libra Faster RCNN (Pang et al., 2019; Girshick, 2015), and Libra RetinaNet (Pang et al., 2019; Lin et al., 2017, 2020) use the fine-tune routine, which uses the ImageNet pre-trained backbone and adjusts the parameters in the model head to fit the current training data. YOLO v5-S does not use the pretrained-backbone and are trained from the scratch. Table 3 shows the performance of the tested detectors in details. Various evaluation criteria are employed, which includes mean average precision (mAP), average recall, and average precision and recall for each category. The first column in the Table 3 is the key performance indicator, which the evaluation criteria in MS-COCO. The MS-COCO criteria computes mean AP across IoU range of [0.5:0.95] with step of 0.05. The second column in the Table 3 is the performance in the Pascal-VOC criteria, which is more relax compared with the MS-COCO.

4.4 Experimental Results and Discussion

The quantitative benchmark results are shown in a detailed comparison way in Table 3. In this quantitative comparison, YOLO v5-S yields better performance than other detectors. The mAP@IoU0.5-0.95 for the results of YOLO v5-S is as high as 0.68, which is significantly higher than other detectors. Among other detectors, the results of FSAF, whose mAP@IoU0.5-0.95 is above 0.6, is closest to the YOLO v5-S

Table 3 Detailed performance of the tested models on the Dunhuang e-Heritage detection dataset

Model Name	mAP@ IoU0.5-0.95	mAP@ IoU0.5	ARI@ IoU0.5-0.95	ARI@ IoU0.5	ARI00@ IoU0.5-0.95	ARI00@ IoU0.5	AP: C1	AP: C2	AP: C3	AP: C4	AP: C5	AR: C1	AR: C2	AR: C3	AR: C4	AR: C5
YOLO v5-S	0.6809	0.9530	0.2609	0.7504	0.7512	0.7139	0.7917	0.7283	0.6545	0.5162	0.7814	0.8435	0.7828	0.7191	0.6293	
SSD300	0.5685	0.9107	0.2305	0.6465	0.6494	0.6384	0.7537	0.6034	0.4857	0.3565	0.7191	0.8142	0.6795	0.5670	0.4674	
RetinalNet	0.5338	0.9005	0.2193	0.6144	0.6197	0.6192	0.7521	0.5605	0.4427	0.2945	0.6851	0.8099	0.6477	0.5334	0.4226	
GHM RetinalNet	0.5627	0.9189	0.2275	0.6386	0.6426	0.6533	0.7591	0.6044	0.4633	0.3333	0.7168	0.8171	0.6847	0.5466	0.4479	
FSAF	0.6051	0.9339	0.2402	0.6757	0.6788	0.6814	0.7699	0.6667	0.5171	0.3907	0.7536	0.8279	0.7314	0.5831	0.4979	
Faster RCNN	0.5531	0.9080	0.2220	0.6359	0.6466	0.6375	0.7456	0.6373	0.4512	0.2941	0.7016	0.8025	0.7105	0.5479	0.4707	
Libra Faster RCNN	0.5718	0.9235	0.2263	0.6482	0.6570	0.6475	0.7517	0.6499	0.4765	0.3332	0.7112	0.8058	0.7197	0.5542	0.4939	
Libra RetinalNet	0.5553	0.9136	0.2253	0.6346	0.6375	0.6240	0.7559	0.5910	0.4827	0.3228	0.6941	0.8121	0.6754	0.5592	0.4470	

in overall performance. The performance of SSD300, RetinaNet, GHM RetinaNet, Faster RCNN, Libra Faster RCNN, and Libra RetinaNet is about the same level. We consider that the factor behind the better performance of YOLO v5-S is its sophisticated data augmentation techniques, which is not applied to other methods in default settings.

In Table 3, we can also notice that all methods show relative same trend in the performances for different categories. For the five categories, all benchmarked methods have best results on category 2 (Thousand Buddha Motif), while perform worst on category 5 (Non-Buddhist Figure).

The category 2 (Thousand Buddha Motif) is considerable relatively easier to detect because of its stable patterns. Objects in this category are generally used in illuminating the ambience for the main theme of a paintings. They are painted in batch, which are of almost identical shape, in limited color scheme, and layout one by one in an aligned and non-overlapped way on the mural. For detectors, these repeated similar objects pose fewer challenges. As shown in Fig. 16, for first 3 categories, even though some predicted bounding boxes may not be completely aligned with the ground truth, YOLO v5-S gives almost perfect results.

Category 4 (Apsaras) is less challenging than expected. As shown in Fig. 17, the silk ribbon in the Apsaras usually spreads widely in the figure with diversified shape. This

makes the silk ribbon to be mis-detected easily and is often not included in the bounding boxes. We thought non-rigid and dramatically varying shapes of Apsaras would make the object the most difficult to detected. However, through the experiments, it was the second challenging and was significantly less challenging than category 5. We carefully inspected the category 4 and the corresponding results and found that the detectors may have learned a key visual characteristic in this category: a more complex structure surrounded with stripe like patterns (i.e., a human shape character surrounded with silk ribbon).

Category 5 (Non-Buddhist Figure) is much more challenging to be detected than other four categories. Figure 18 shows two examples of category 5 with predicted bounding boxes, and all people in the figure are highly compact. For category 5, when inspecting the data in this category, two possible reasons might contribute the lower mAP: (1) Non-Buddhist figures are different from each other in shape and, sometimes, are overlapped. Thus, it's much more difficult for the detectors to determine the boundaries accurately for each character. (2) The objects in category 5 contains more detailed and different internal structures (e.g., higher visual diversity than other 4 categories).

Yolo v5s use mosaic data augmentation, which create more highly diversified training images. As a result, the

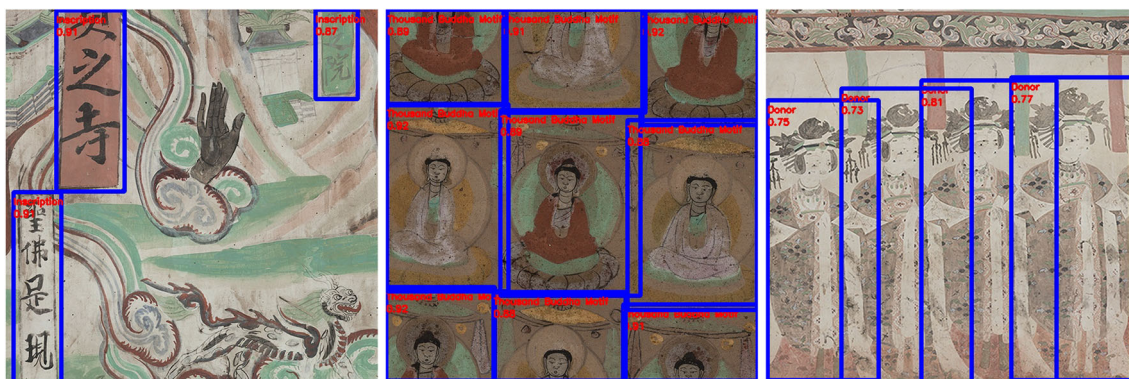


Fig. 16 Examples of Inscription (left), Thousand Buddha Motif (middle), and Donors (right)

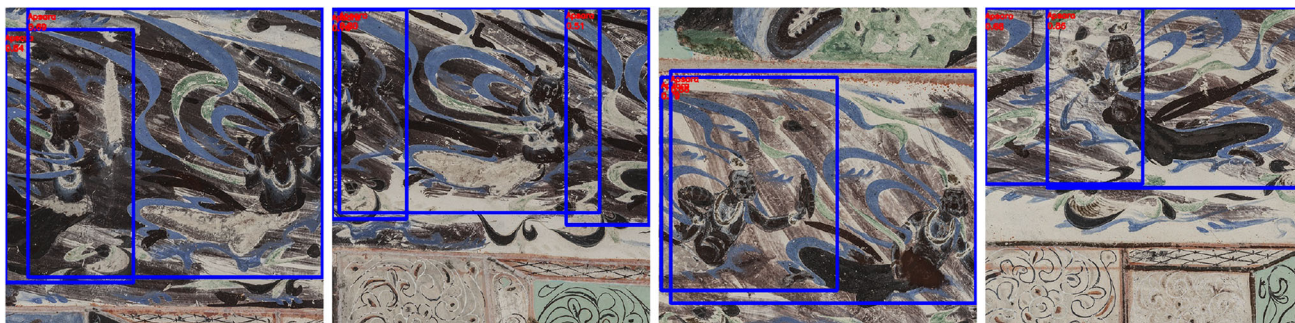
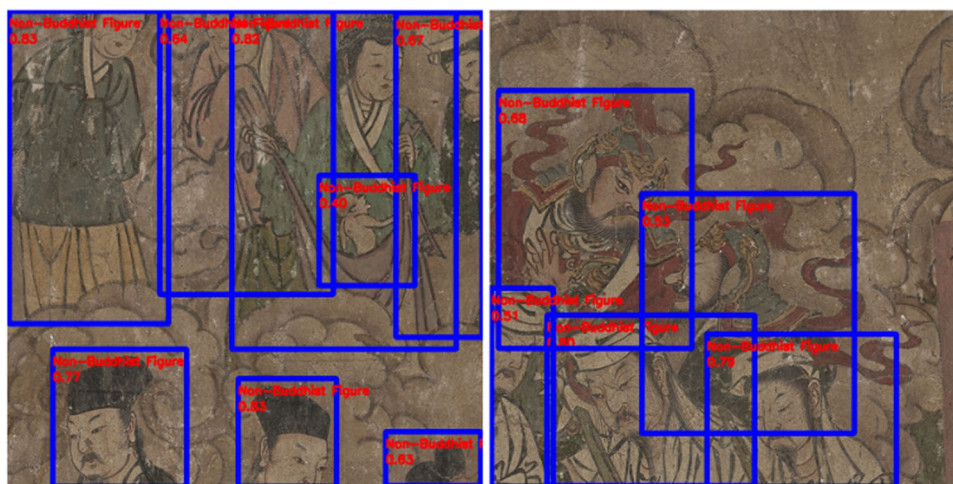


Fig. 17 Four examples of Apsara. Bounding boxes can detect the bodies of the Apsara characters accurately, even though they might miss some parts of the silk ribbon around the characters

Fig. 18 Two examples of Non-Buddhist Figure. Bounding boxes shown in the figures are compacted and overlapped and the style of characters is highly diversified



model works good at small training set like AI for Dunhuang. Besides, these highly diversified images from mosaic data augmentation makes model perform much better at Non-Buddhist figure and Apsaras figure. The example shows that YOLO v5-S is a good choice for our future application. Even though, the positions of bounding boxes do not match with the ground truth perfectly, the detection is accurate enough for practical use. To increase the detection accuracy for those difficult samples, two approaches could be employed. Firstly, since the Dunhuang Academy has large scale of e-Heritage images, more samples can be added into the coming enlarged dataset. It provides more knowledge for detectors to learn. Secondly, more dedicated data augmentation can be designed and used in training so that more challenging samples can be generated.

5 Artistic Style Transfer

In this section, we explore style transfer for Dunhuang paintings. The motivation are twofolds. First, the Dunhuang painting are created over 1000 years, the artistic style will vary significantly. And thus, style transfer based analysis may help to align and compare the artistic styles. This will enable archaeologist to further study how the styles are evolved. Second, similar as many other museums, converting natural images to Dunhuang style are very attractive to the tourists, and can boost the user experience. In detail, we introduce a Dunhuang *Shanshui* artistic dataset, which is defined and scrutinized by artist in Dunhuang Academy. The abstract framework of end-to-end style transfer methods is presented, and some methods are introduced. After benchmarking eight state-of-the-art methods, an experimental analysis and discussion are presented.

5.1 Dataset for Artistic Style Transfer

The first e-Heritage dataset (so-called *QingLvShanShui* dataset) based on the murals of Dunhuang Grottoes is built up by using unpaired image-to-image method. The *QingLvShanShui* dataset consists of two types of images: the first type is formed by the murals of Dunhuang Grottoes in a specific well-defined style (so-called *QingLvShanShui* style), which is classified and defined with artistic experts and domain knowledge; the second type is formed by natural image.

QingLvShanShui style is one of the most characteristic style of Tang dynasty's murals in Dunhuang Grottoes. In this style, the green color is the most noticeable color, and the major contents include mountain, river, tree, and sky, whilst the spatial relationship with relative distance in the scene is the composition of above. Nowadays, the existing *QingLvShanShui* style of Tang dynasty's murals can only be seen in the murals of Dunhuang and it is very rare, therefore, it is the most important and valuable for the research work of Tang dynasty's Shanshui paintings.

In Fig. 19, *QingLvShanShui* dataset is constructed from the murals of several Mogao caves such as No.23, No.103, No. 172, No. 217, No. 320, and No. 321. Four values of DIP, namely, 150, 75, 30, and 15dpi, are chosen to improve the training feature efficiency under different scenarios. For each chosen *QingLvShanShui* style murals, it is randomly cropped into several images with the size of 256x256. Furthermore, the useless images are ruled out by experts from Dunhuang Academy.

Finally, there are a total of 3455 samples images in the *QingLvShanShui* dataset, and some of them are shown in Fig. 20. Given that the unpaired image-to-image method is used in our experiments, it only needs to concern about the projection from ordinary photo domain to Dunhuang Shanshui image domain, and a separate test set is not arranged

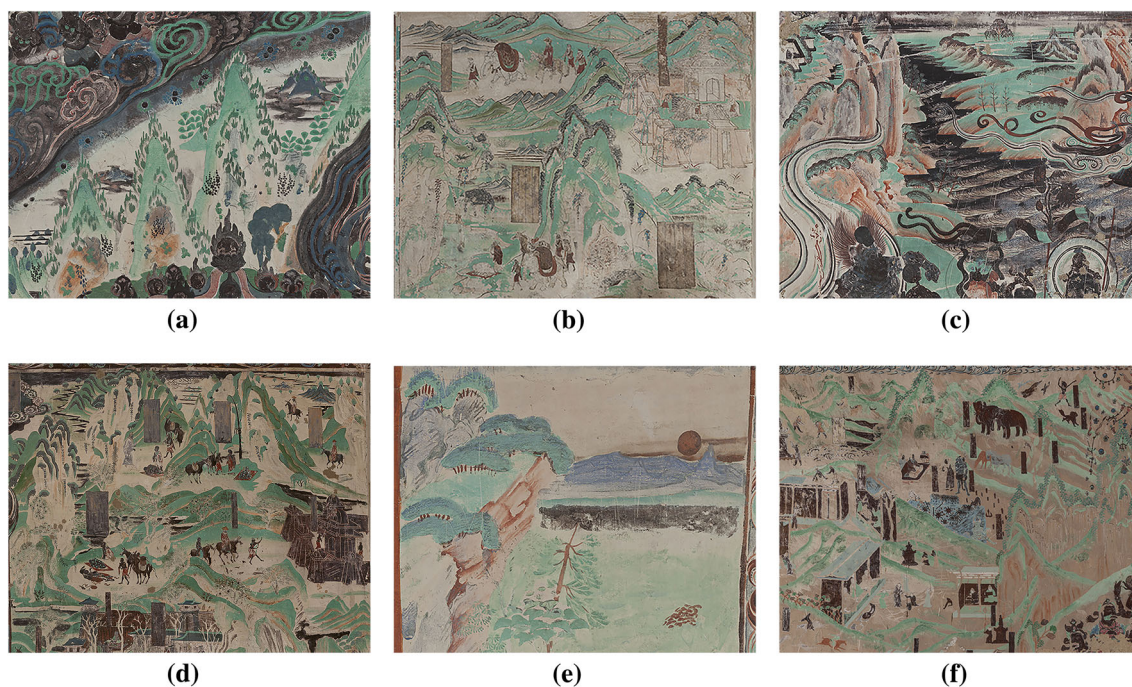


Fig. 19 Manually selected e-Heritage by artists for the source of Dunhuang Shanshui style dataset. **a** The painting images are from north wall of cave No. 23 at Mogao Grottoes, **b** south wall of cave No. 103 at Mogao Grottoes, **c** east wall of cave No. 172 at Mogao Grottoes, **d**

south wall of cave No. 217 at Mogao Grottoes, **e** south wall of cave No. 320 at Mogao Grottoes, **f** and south wall of cave No. 321 at Mogao Grottoes

for the experiments. Specifically, all of 3455 images are used for training set. In the inference stage, same training set of *QingLvShanShui* dataset is used as its the dummy test set when necessary.

5.2 Methodology

Dunhuang Academy is working with Jinan University to implement the latest unpaired image-to-image style transfer methods using the Dunhuang Shanshui style dataset. The tested methods are Cycle-GAN (Zhu et al., 2017), CUT (Park et al., 2020), DualAST (Chen et al., 2021), GATYS (Gatys et al., 2016), DFP (Wang et al., 2020), ADAIN (Huang & Belongie, 2017), WCT (Li et al., 2017), and Avatar-Net (Sheng et al., 2018). Theoretically, the unpaired image-to-image style model can be expressed as a function \mathcal{F} that is capable of processing an image I_{in} in one style domain and output another image I_{out} in another domain. The expression is as follows:

$$I_{out} = \mathcal{F}(I_{in}|\theta), \quad (6)$$

where θ is the parameters of the trained model \mathcal{F} . Variable θ is obtained by the optimization process, which can be abstractly given as:

$$\tilde{\theta} = \arg \min_{\theta} E[\mathcal{G}(\mathcal{F}(I_{in}|\theta), I_{in}) + \mathcal{H}(\mathcal{F}(I_{in}|\theta), I_{in})], \quad (7)$$

where E is the expectation, \mathcal{G} is the distance measure of the content or structural information, and \mathcal{H} compute the distance of style or image characteristics. In both tested methods, the input and output images are in the same size of 256×256 .

In the cycle-GAN, the methods simultaneously train two encoder–decoder-like generator models. Shown in Fig. 21, one generator, denoted as P , is to project a sample a in domain D_A onto domain D_B . Another generator Q reverses this projection, which projects a sample b from domain D_B to domain D_A . In our scenario, we set domain D_A as daily photo domain, and domain D_B as Dunhuang image domain. The dataset A and B are the sample collections of domain D_A and D_B , respectively. The goal is to obtain a useful generator model P for the style transfer, and Q is the helper model for training.

During the training, the key to avoiding the requirements of paired dataset is in bidirectional cycle-consistency loss. The cycle-consistency loss used the two generator to compute the distance between the a and the $\hat{a} = P(\hat{b}) = P(Q(a))$. In this manner, we avoid the requirement for existence of corresponding b in dataset B for a , and \hat{b} is merely an implicit intermediate variable generated during training. Thus, given a daily photo I_{in} , the corresponding groundtruth of I_{in} in

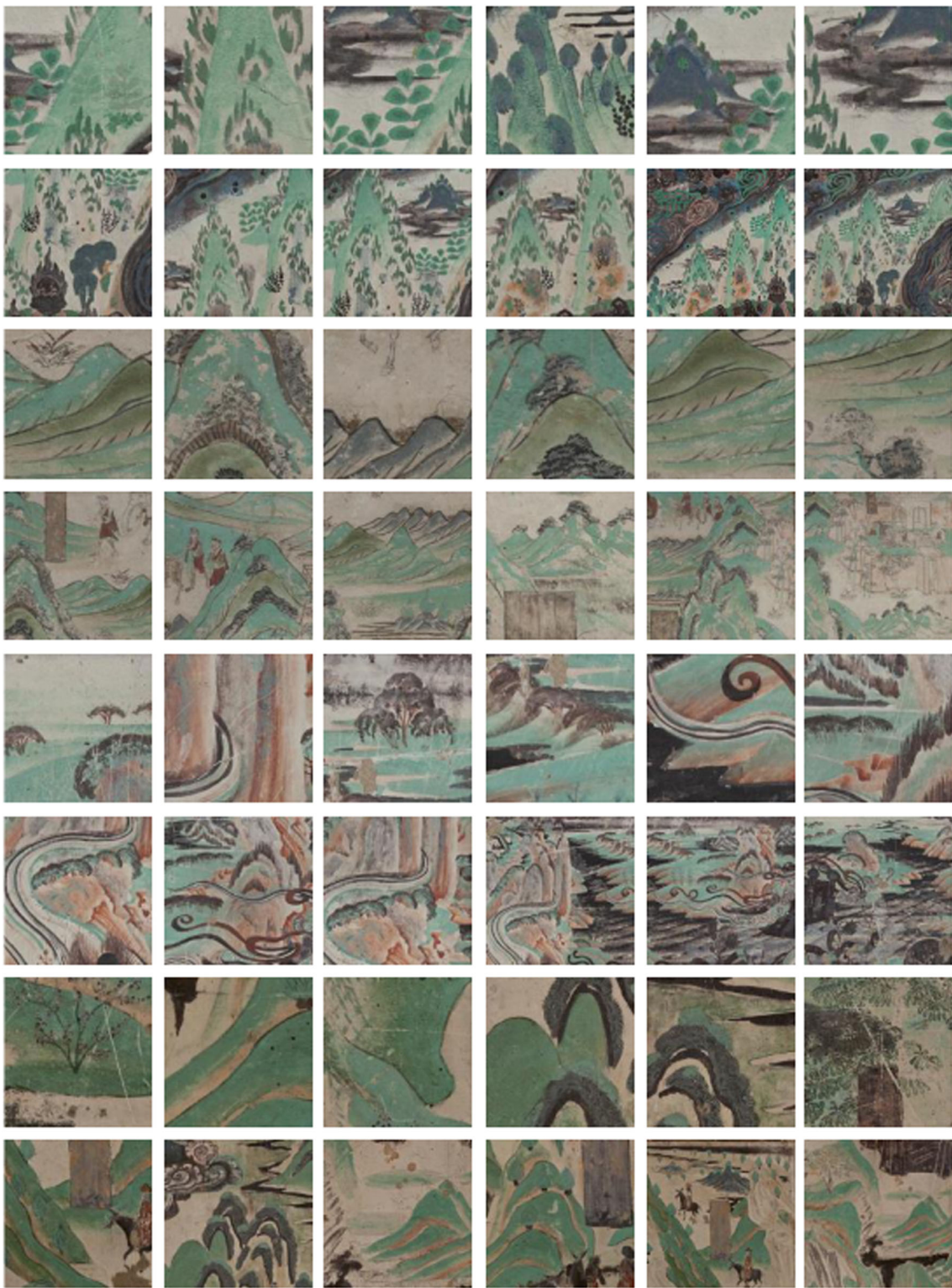


Fig. 20 Training samples from Dunhuang Caves *QingLvShanShui* style dataset

Fig. 21 Illustration of cycle-consistency loss

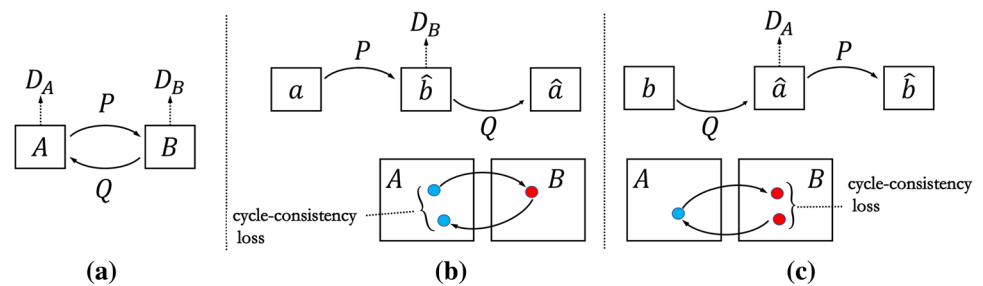
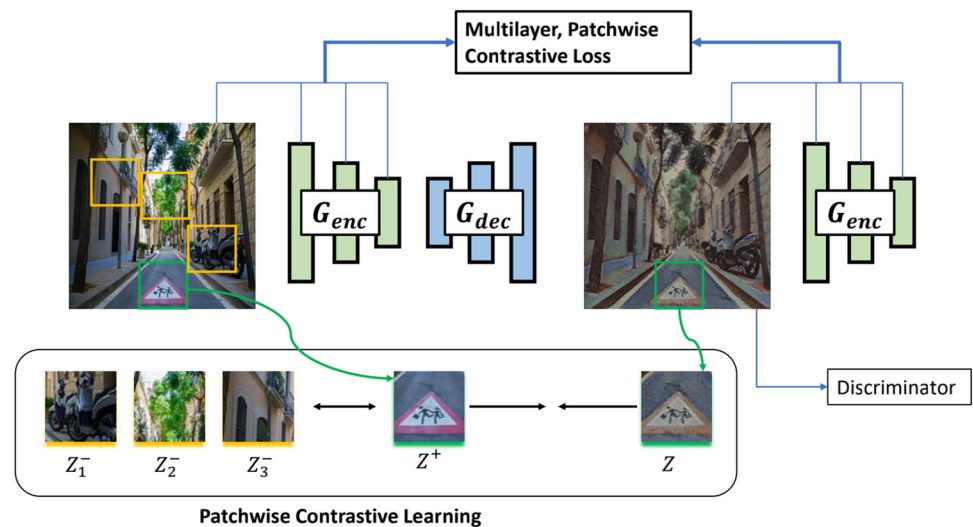


Fig. 22 Illustration of CUT and Patchwise Contrastive Loss



Dunhuang style domain, which does not exist in real world, is not required.

The CUT method, which is based on Cycle-GAN, has dropped the cycle-consistency loss and only use a single generator. Considering the of underlying bijection of cycle-consistency loss as too restrictive assumption, CUT has proposed a PatchNCE loss based on contrastive learning. In Fig. 22, the PatchNCE loss is to build up the association of textures between the input and the output domains by maximizing mutual information.

5.3 Experiment

The experiments are conducted on a regular x86 workstation with Intel i7-7800X CPU @3.50 GHz, 16 GB RAM, and a graphic card of NVIDIA 1080Ti with 11 GB graphic memory. The training process of the two experiments takes approximately 8 h, and the inference speed is in real-time. The models become effective after training for approximately 2 h and continue to improve its performance. The training and test sets for the input domain are from *yosemite* imageset of Zhu et al. (2017). The Dunhuang Style dataset is used as the training set of the output domain. Some images from Dunhuang Shanshui Style dataset are randomly picked for the test set of the output domain. The test set of the output

domain is trivial due to the projection from the Dunhuang e-Heritage image to the daily photo is not our focus.

Some qualitative results from Cycle-GAN (Zhu et al., 2017), CUT (Park et al., 2020), DualAST (Chen et al., 2021), GATYS (Gatys et al., 2016), DFP (Wang et al., 2020), ADAIN (Huang & Belongie, 2017), WCT (Li et al., 2017), and Avatar-Net (Sheng et al., 2018) are shown in Fig. 23.

The key summary characteristic is the extensive use of green color on mountains and rivers, and the strong contour along the mountain ridges. The figure shows that all methods have successfully translated the daily scenery photos into the Dunhuang *Shanshui* images, rendering the output images in the target style and preserving their content.

From the perspective of artistic analysis and human perception, the quality of results generated by the eight style transfers can be classified into three tiers. Here, the Tier 1 has the best quality, while Tier 3 is the worst.

The Tier 1 results are yielded by Cycle-GAN, Dualast, and Avatar-Net. Their results are different in styles to a certain extent; accordingly, it would be difficult to tell which one is better. The Cycle-GAN tends to generates more realistic results with mural feels. This phenomenon occurs because the structures (i.e., lines and contextual segments) are well preserved in the results. The cloud, humans, and trees can be better recognized compared with the results of other methods. The noise from the mural images is also transferred into

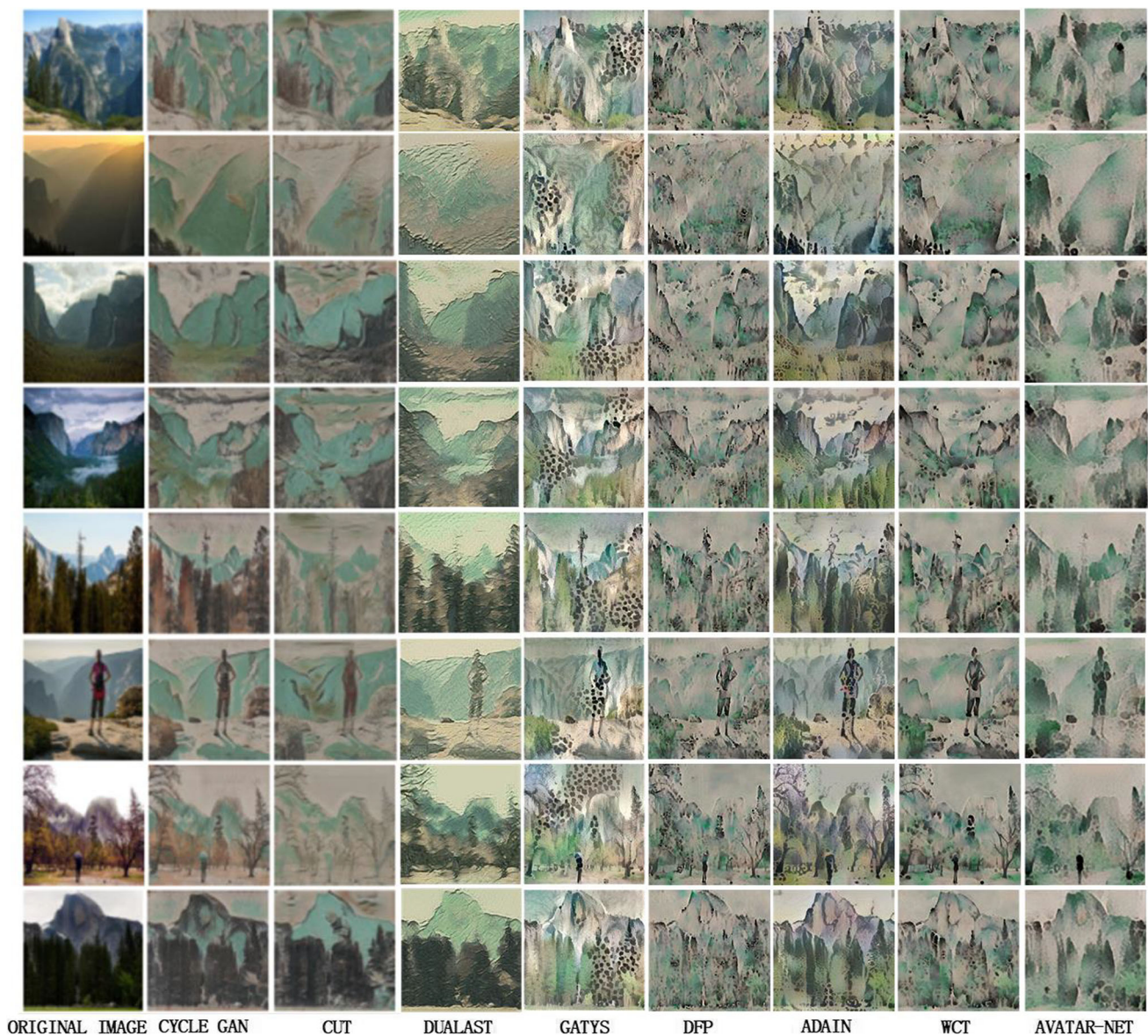


Fig. 23 Style transfer results generated by classical algorithm on the Dunhuang *Shanshui* style dataset. From left to right, they are original landscape photos, Cycle-GAN (Zhu et al., 2017), CUT (Park et al.,

2020), DualAST (Chen et al., 2021), GATYS (Gatys et al., 2016), DFP (Wang et al., 2020), ADAIN (Huang & Belongie, 2017), WCT (Li et al., 2017), and Avatar-Net (Sheng et al., 2018)

the result images but without adding too many other artifacts. The results of Avatar-Net appear more like traditional paintings from the northeast Asia, which is more impressionistic rather than realistic. Dualast generated results that looks more like traditional oil paintings from Europe. Although certain neural artifacts exist in some plain segments, the global stereo perception remains strong.

Tier 2 includes CUT, ADAIN, and WCT. The results of these methods are acceptable because the style is successfully transferred, and some details of the original photos are preserved. However, some drawbacks are apparent and easily observable. The style of CUT is somewhat similar to that of

Cycle-GAN, which the CUT based on. However, the results are not improving on the Dunhuang *Shanshui* dataset. The qualitative results show over polarized color distributions and loss of textural details. The contrast of the CUT results has shifted from the original image, making them look like different scenes. The results of ADAIN have strong global stereo feelings. However, it generated too many additional contextual segments, which is not consistent with the original, making the image difficult to be understood. The results of WCT are reasonably good, but the randomized color renderings and substantial unrelated noises leads to falling in Tier 2.

The results from Tier 3 methods, namely DFP and GATYS, are not great. The DFP generates irregular color distribution and unrelated noises, resulting in a messy image which affects human understanding. The results of GATYS appears abnormal because the method usually generates a group of unrelated dot-shape artifacts in the area that they should not appear. There are few well-defined quantitative evaluation criteria for style transfer method that are unavailable in literature. Accordingly, it is quite difficult to tell which method better translates the style in a very objective way. In the following section, we try to quantitatively analyze the results.

Assessing artistic style transfer results could be a highly subjective task. Here, we adopt two quantitative evaluation metrics: deception rate and user study, to better evaluate classical method. Deception rate was introduced by Sanakoyeu et al. (2018) to quantitatively and automatically assess the quality of the images generated by stylization method. First, Karayev et al. (2013) trained a VGG-16 network to classify 624 artists on WikiArt from scratch. Then, the pre-trained network was employed to predict which artist the stylized image belongs to. Finally, the deception rate was calculated as the fraction of times that the network predicted the correct artist (Chen et al., 2021). We report the deception rate for classical eight baseline models in Table 4. User study has been widely adopted by previous works (Chen et al., 2021) to investigate user preference over different visual results. Here, we conduct user studies to evaluate the user preference of classical eight methods in terms of visual quality. Given various photographs, we stylize them in the style using classical eight baseline methods. Then we show the randomly ordered stylized images produced by eight compared methods to participants and ask them to select the image that best represents the style of the target artist. We finally collect 1000 votes from 50 participants. We report the percentage of votes for each method in the third row of Table 4. The quantitative results in Table 4 are mostly consistent with our qualitative analysis, except that Avatar-Net rated lower than expected. The evaluating criteria tends to emphasize more on the structures rather than the correctness of color distribution. Cycle-GAN is rated highest in the benchmark. However, in future application, Dunhuang Academy consider the Tier 1 methods as candidates for deployment, and plan to further optimize the methods and results.

6 Conclusion and Discussion

To make use of these data and facilitate the development of various intelligent applications, Dunhuang Academy plans to: (1) develop an industrial standard for e-Heritage data; (2) make a dataset public for researchers; and (3) conduct an inte-

gral technical demonstration of using intelligent applications in the near future.

In this work, the project on Dunhuang cultural heritage protection using artificial intelligence is introduced. A new dataset namely, AI for Dunhuang, is created to facilitate this research. Version v1.0 of the dataset comprises of the data and labels for the restoration, style transfer, and retrieval. Given the dataset, three e-Heritage using AI tasks have been proposed. First, a deep network is proposed to automatically perform the restoration. Second, given that the grottoes were build over 1000 years by numerous artists, style transfer is proposed to link and analyze the styles over time. Lastly, deep neural network-based detection and retrieval are proposed and benchmarked to further analyze the tremendously large amount of objects, which are unreasonable to manually label and analyze. The experiments demonstrate that the performance and efficiency in archaeology has been significantly boosted by using modern computer vision technologies.

For e-dunhuang users can only browse the entire mural content, but cannot browse, categorize and compare of the thematic elements. Based on the historical property retrieval technology described in this paper, the computer can automatically find five types of elements in all murals, so as to quickly build a library of thematic elements of murals for archaeology and Dunhuang researchers to study. For the five types of elements, based on the virtual restoration technology described in this article, the broken parts of the five types of elements can be automatically restored. With the restored images of the elements, the style transfer technology described in this article is used to generate more stylistic images to provide a reference for artists. According to the results of archaeological research and content value mining of the experts on the five types of elements, the images generated by virtual restoration and style transfer technologies can be widely used in the exhibition and dissemination of cultural heritage IP.

In terms of the construction of library of elements, virtual restoration and style transfer, the AI technology reported in this paper will reach the ability level that only professional person has, thereby accelerating element search, improving the efficiency of virtual restoration and improving the ability of transfer between different styles. In future work, using the AI technology mentioned in this article to solve specific problems will also be challenged. For the historical property retrieval technology, although it solves the problem of positioning of elements in the entire mural, but for more difficult and more complex problems than Apsaras and non-Buddhist figures, the accuracy needs to be improved. For the virtual restoration technology, the ability needs to be improved for the larger broken area. The style and color generated by the style transfer technology need to be more similar to the style of Dunhuang murals. And the integration ability of AI technology in the e-dunhuang project still needs to be improved.

Table 4 Deception rate and user study (in terms of visual quality) of different methods

	Cycle-GAN	CUT	DualAST	GATYS	DFP	AdaIN	WCT	Avatar-Net
Deception Rate	0.602	0.592	0.580	0.202	0.182	0.062	0.025	0.045
Visual Quality (%)	0.310	0.303	0.176	0.060	0.058	0.051	0.010	0.032

The higher the better . The best scores are reported in bold

In future work, we will further enhance the existing three tasks through multi-modal learning, where visual information is combined with historical documentary information to further improve the performance. Moreover, we consider semantic segmentation for segmenting numerous complex objects from the huge mural painting.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11263-022-01665-x>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anichini, F., et al. (2020). Developing the archaide application: A digital workflow for identifying, organising and sharing archaeological pottery using automated image recognition. *Internet Archaeology*. <https://doi.org/10.11141/ia.52.7>.
- Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., & Verdera, J. (2001). Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on Image Processing*, 10(8), 1200–1211. <https://doi.org/10.1109/83.935036>.
- Banno, A., Masuda, T., Oishi, T., & Ikeuchi, K. (2008). Flying laser range sensor for large-scale site-modeling and its applications in Bayon digital archival project. *International Journal of Computer Vision*, 78(2), 207–222. <https://doi.org/10.1007/s11263-007-0104-6>.
- Barnes, C., Shechtman, E., Finkelstein, A., & Goldman, D. B. (2009). Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, 28(3), 24:1–24:11. <https://doi.org/10.1145/1531326.1531333>.
- Bertalmio, M., Sapiro, G., Caselles, V., & Ballester, C. (2000). Image inpainting. In: *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '00 (pp. 417–424). New York, NY, USA: ACM Press/Addison-Wesley Publishing Co. <https://doi.org/10.1145/344779.344972>.
- Bok, Y., Jeong, Y., Choi, D. G., & Kweon, I. S. (2011). Capturing village-level heritages with a hand-held camera-laser fusion sensor. *International Journal of Computer Vision*, 94(1), 36–53. <https://doi.org/10.1007/s11263-010-0397-8>.
- Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6), 679–698. <https://doi.org/10.1109/TPAMI.1986.4767851>.
- Chen, H., Zhao, L., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., & Lu, D. (2021). Dualast: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 872–881).
- Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4), 743–761.
- DunhuangAcademy: e-dunhuang by dunhuang academy (2015–2021). <https://www.e-dunhuang.com>.
- Efros, A. A., & Leung, T. K. (1999). Texture synthesis by non-parametric sampling. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 2, 1033–1038. <https://doi.org/10.1109/ICCV.1999.790383>.
- Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2), 303–338.
- Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2414–2423).
- Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3354–3361). IEEE.
- Girshick, R. (2015). Fast r-cnn. In: *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 1440–1448). <https://doi.org/10.1109/ICCV.2015.169>.
- Haliassos, A., Barmoutis, P., Stathaki, T., Quirke, S., & Constantinides, A. (2020). Classification and Detection of Symbols in Ancient Papyri (pp. 121–140). Springer.
- He, K., Zhang, X., Ren, S., & Sun, J. (jun 2016). Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE Computer Society: Los Alamitos, CA, USA. <https://doi.org/10.1109/CVPR.2016.90>. <https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90>.
- Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., & Salesin, D. H. (2001). Image analogies. In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (pp. 327–340).
- Hou, M., Zhou, P., Lv, S., Hu, Y., Zhao, X., Wu, W., et al. (2018). Virtual restoration of stains on ancient paintings with maximum noise fraction transformation based on the hyperspectral imaging. *Journal of Cultural Heritage*, 34, 136–144. <https://doi.org/10.1016/j.culher.2018.04.004>.
- Huang, J. B., Kang, S. B., Ahuja, N., & Kopf, J. (2014). Image completion using planar structure guidance. *ACM Transactions on Graphics*, 33(4), 129:1–129:10.

- Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1501–1510).
- Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4), 107:1–107:14. <https://doi.org/10.1145/3072959.3073659>.
- Ikeuchi, K. (2013). e-heritage, cyber archaeology, and cloud museum. In: *2013 International Conference on Culture and Computing* (pp. 1–7). <https://doi.org/10.1109/CultureComputing.2013.77>.
- Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, tkianai, Hogan, A., lorenzomamma, yxNONG, AlexWang1900, Diaconu, L., Marc, wanghaoyang0106, ml5ah, Doug, Ingham, F., Frederik, Guilhen, Hatovix, Poznanski, J., Fang, J., Yu, L., changyu98, Wang, M., Gupta, N., Akhtar, O., PetrDvoracek, & Rai, P. (2020). ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements. <https://doi.org/10.5281/zenodo.4154370>.
- Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision - ECCV 2016* (pp. 694–711). Springer.
- Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., & Winnemoeller, H. (2013). Recognizing image style. arXiv preprint [arXiv:1311.3715](https://arxiv.org/abs/1311.3715).
- Kwatra, V., Essa, I., Bobick, A., & Kwatra, N. (2005). Texture optimization for example-based synthesis. *ACM Transactions on Graphics*, 24(3), 795–802. <https://doi.org/10.1145/1073204.1073263>.
- Levin, Zomet, & Weiss. (2003). Learning how to inpaint from global image statistics. *Proceedings Ninth IEEE International Conference on Computer Vision*, 1, 305–312. <https://doi.org/10.1109/ICCV.2003.1238360>.
- Li, B., Liu, Y., & Wang, X. (2019). Gradient harmonized single-stage detector. In *AAAI Conference on Artificial Intelligence*.
- Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M. H. (2017). Universal style transfer via feature transforms. arXiv preprint [arXiv:1705.08086](https://arxiv.org/abs/1705.08086).
- Lin, T., Goyal, P., Girshick, R., He, K., & Dollár, P. (Oct 2017). Focal loss for dense object detection. In: *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2999–3007). <https://doi.org/10.1109/ICCV.2017.324>.
- Lin, T., Goyal, P., Girshick, R., He, K., & Dollár, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), 318–327. <https://doi.org/10.1109/TPAMI.2018.2858826>.
- Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In: *European conference on computer vision* (pp. 740–755). Springer.
- Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision - ECCV 2018* (pp. 89–105). Springer.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision - ECCV 2016* (pp. 21–37). Springer.
- Lu, M., Zhao, H., Yao, A., Chen, Y., Xu, F., & Zhang, L. (2019). A closed-form solution to universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5952–5961).
- Nazeri, K., Ng, E., Joseph, T., Qureshi, F., & Ebrahimi, M. (Oct 2019a). Edgeconnect: Structure guided image inpainting using edge prediction. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.
- Nazeri, K., Ng, E., Joseph, T., Qureshi, F., & Ebrahimi, M. (Oct 2019b). Edgeconnect: Structure guided image inpainting using edge prediction. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.
- Pan, G., Sun, D., Zhan, R., & Zhang, J. (2018). Mural sketch generation via style-aware convolutional neural network. In *Proceedings of Computer Graphics International 2018* (pp. 239–245). CGI 2018, Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3208159.3208160>.
- Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., & Lin, D. (2019). Libra r-cnn: Towards balanced learning for object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 821–830). <https://doi.org/10.1109/CVPR.2019.00091>.
- Park, T., Efros, A. A., Zhang, R., & Zhu, J. Y. (2020). Contrastive learning for unpaired image-to-image translation. In A. Vedaldi, H. Bischof, T. Brox, & J. M. Frahm (Eds.), *Computer Vision - ECCV 2020* (pp. 319–345). Springer.
- Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., & Efros, A. A. (June 2016). Context encoders: Feature learning by inpainting. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2536–2544). <https://doi.org/10.1109/CVPR.2016.278>.
- Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1), 49–70.
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).
- Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. LNCS, (vol. 9351, pp. 234–241). Springer. <http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a>, (available on [arXiv:1505.04597](https://arxiv.org/abs/1505.04597) [cs.CV]).
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.
- Sanakoyeu, A., Kotovenko, D., Lang, S., & Ommer, B. (2018). A style-aware content loss for real-time hd style transfer. In *proceedings of the European conference on computer vision (ECCV)*. (pp. 698–714).
- Sheng, L., Lin, Z., Shao, J., & Wang, X. (2018). Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8242–8250).
- Simakov, D., Caspi, Y., Shechtman, E., & Irani, M. (June 2008). Summarizing visual data using bidirectional similarity. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). <https://doi.org/10.1109/CVPR.2008.4587842>.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Computer Science*.
- Sofiuk, K., Petrov, I., Barinova, O., & Konushin, A. (2020). F-brs: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Song, Y., Yang, C., Lin, Z., Liu, X., Huang, Q., Li, H., & Kuo, C. C. J. (2018). Contextual-based image inpainting: Infer, match, and translate. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision - ECCV 2018* (pp. 3–18). Springer.
- Sun, D., Zhang, J., Pan, G., & Zhan, R. (2018). Mural2sketch: A combined line drawing generation method for ancient mural painting.

- In: *2018 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). <https://doi.org/10.1109/ICME.2018.8486504>.
- Wang, H. (2019). The study of dunhuang murals in the digital age-taking the maitreya illustration on the north wall in mogao cave 72 as a case study. *DUNHUANG RESEARCH*, 176(4), 26–37.
- Wang, Z., Zhao, L., Chen, H., Qiu, L., Mo, Q., Lin, S., Xing, W., & Lu, D. (2020). Diversified arbitrary style transfer via deep feature perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7789–7798).
- Wexler, Y., Shechtman, E., & Irani, M. (2007). Space-time completion of video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3), 463–476. <https://doi.org/10.1109/TPAMI.2007.60>.
- Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*. (pp. 1395–1403). <https://doi.org/10.1109/ICCV.2015.164>.
- Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., & Li, H. (2017). High-resolution image inpainting using multi-scale neural patch synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4076–4084). <https://doi.org/10.1109/CVPR.2017.434>.
- Yoo, J., Uh, Y., Chun, S., Kang, B., & Ha, J. W. (2019). Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9036–9045).
- Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (June 2018). Generative image inpainting with contextual attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5505–5514). <https://doi.org/10.1109/CVPR.2018.00577>.
- Yu, T., Lin, C., Zhang, S., You *, S., Ding, X., Wu, J., & Zhang, J. (Oct 2019a). End-to-end partial convolutions neural networks for dunhuang grottoes wall-painting restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.
- Yu, T., Lin, C., Zhang, S., You, S., Ding, X., Wu, J., & Zhang, J. (2019b). End-to-end partial convolutions neural networks for dunhuang grottoes wall-painting restoration. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. (pp. 1447–1455). <https://doi.org/10.1109/ICCVW.2019.00182>.
- Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6), 1452–1464. <https://doi.org/10.1109/TPAMI.2017.2723009>.
- Zhu, C., He, Y., & Savvides, M. (June 2019). Feature selective anchor-free module for single-shot object detection. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 840–849). <https://doi.org/10.1109/CVPR.2019.00093>.
- Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.