End-to-End Partial Convolutions Neural Networks for Dunhuang Grottoes Wall-painting Restoration

Tianxiu Yu
Dunhuang Academy
Jiuquan, Gansu, China
dhagraph@163.com

Cong Lin
Jinan University
Zhuhai, Guangdong, China
conglin@jnu.edu.cn

Shijie Zhang
Tianjin Medical University
Tianjin, China
shijie.zhang@tmu.edu.cn

Shaodi You *
corresponding author, CSIRO
Canberra, Australia
youshaodi@gmail.com

Xiaohong Ding
Dunhuang Academy
Jiuquan, Gansu, China
dxh7155@qq.com

Jian Wu
Dunhuang Academy
Jiuquan, Gansu, China
dunhuang_wujian@126.com

Jiawan Zhang
Tianjin University
Tianjin, China
jwzhang@tju.edu.cn

Abstract

In this paper, we focus on training a deep neural network to in-paint and restore the historical painting of Dunhuang Grottoes. Dunhuang Grottoes is more than 1000 years old and the wall-painting on the grottoes has suffered from various deterioration. The ground truth does not exist either. Furthermore, learning the style of the artists is not straight forward because the wall-paintings are created by thousands of artists over more 400-500 years. As the very first attempt to solve this problem, we propose an end-to-end image restoration model for Dunhuang wall-painting. The end-to-end image restoration model employ U-net with partially convoluational layers to construct, which is capable in restoring non-rigid deteriorated content given a loss content mask and a wall-painting image. To learn the various artists style from real data, the training set and validation set are collected by using a zooming-in-like and random cropping approach on the digital RGB images photographed on the healthy Grotto-painting. We also synthesize the deteriorated paintings from real data. To ensure the synthetic content in the masked region is consistent to the ground truths in term of texture, colors, artistic style and free of unnecessary noises, the loss function is in a hybrid form that comprises transition variation loss, content loss and style loss. Our contributions are of three folds: 1) proposed using partial convolutional U-net in restoring wall-paintings; 2) the method is tested in restoring highly non-rigid and irregular deteriorated regions; 3) two types of masks are designed for simulating deteriorations and experimental results are satisfactory.

1. Introduction

Image restoration is a fundamental technology in image processing and computer vision. It is also very important for modern archaeologist and historian that there is strong motivation to utilize computer vision and image processing technology to automatically in-paint and restore historical documents and paintings.

Unlike traditional technology, in recent years, deep neural network enables the computer to learn from existing content and style. It has been proved to be more reliable than traditional rule based technologies. Specifically, to restore the loss regions, one needs to learn the well-defined style and the similar texture of the target image; then synthesize the labeled lost non-rigid regions with new texture in consistent local context. One of the most common methods for art-work restoration is to scan or photocopy the wall-painting into digital image, then to employ digital image processing techniques to restore the lost parts. Nonetheless, wall-painting restoration is an ill-post problem, owing that most of the color and texture of region to be restore are lost in the history and the true knowledge no longer exists. This requires that the image restoration method is capable in inpainting loss region with synthetic contents which in

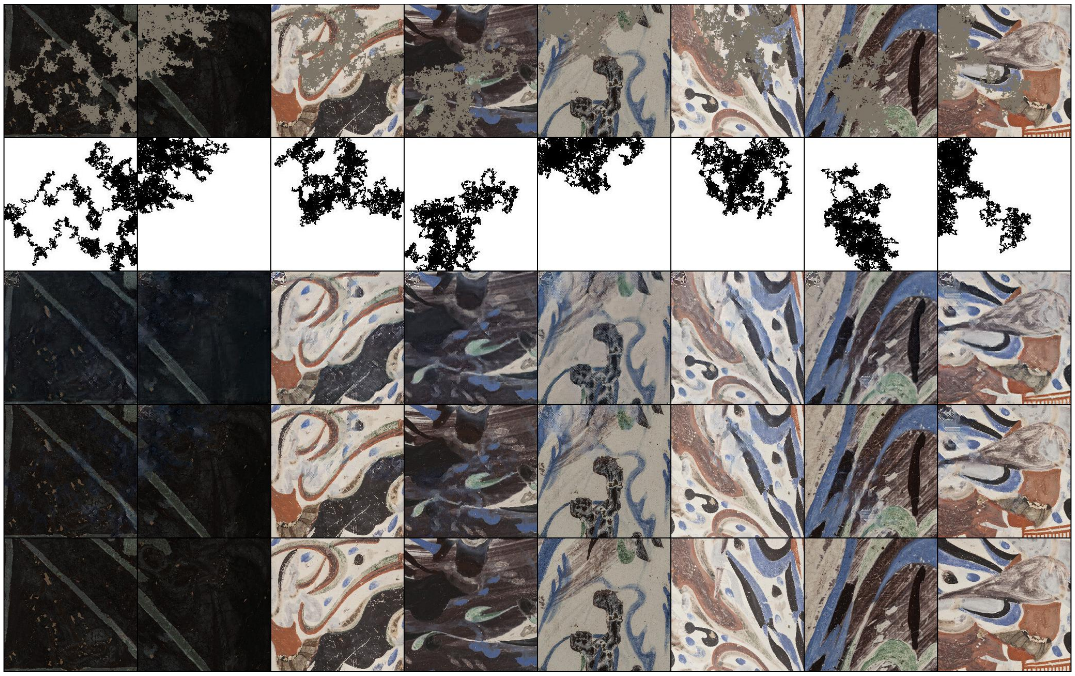

Figure 1. Some restoration results. Each column represents the related data of a image sample. From row 1 to row 5: deteriorated input image, input mask, output predicted image, merged final result, groundtruth image

consistency of peripheral regions in term of artistic style, semantic perception and textural distribution. Unlike natural image inpainting, which is aiming to generate reasonable content. Inpainting of artwork relying on generating the content more similar to the artist's style. It is, however, not that straight forward as natural image inpainting, because the number of existing artworks of particular artist is limited. Furthermore, for historical painting, there might be not existing a groundtruth for training the deep neural network.

In this paper, we focus on training a deep neural network to in-paint and restore the historical paintings of Dunhuang Grottoes. Dunhuang Grottoes is more than 1000 years old and the wall-painting on the grottoes has suffered from various types of deterioration. The ground truth does not exist either. Furthermore, learning the style of the artist is not straight forward because the wall-paintings are created by thousands of artists over a time span of more than 400 years. To this end, we proposed using end-to-end partial convolutional neural network to build up a novel digital image restoration algorithm for Dunhuang Grottoes. Particularly, partial convolutional uses a mask, which indicates irregular deteriorated regions, as input conditions. Since deterioration on Grotto-painting occurs in various forms with different shapes, the input with flexible representation of deterioration and the output of a single RGB image fit wall-painting restoration task properly. To learn the various styles from multiple artists. The training set and validation set are collected by using a zooming-in-like and random cropping approach on the digital RGB images photographed on the healthy Grotto-paintings. The Grotto-painting photographs captured using DSLR cameras from Grotto sites are too large for any practical neural networks. Simply re-scaling the photographs will lead to loss of fine details; and the network built on these re-scaled photos is probably not able to restore regional deterioration. Thus, we randomly generate small rectangular bounding boxes with proper high-width ratios to crop regional patches and save as training or validation samples. There are two advantages of this approach; on one hand, local fine details and rich textures are preserved in the cropped samples; on the other, it generates numerous samples for feeding a larger neural network. Our loss function is in a hybrid form that comprises transition variation (TV) loss, content loss and style loss. The content loss and style loss, proposed by Johnson et al.[7], have been successfully applied on image style transfer and super-resolution. The content loss measures the textural difference between the predicted and the target; while style loss measure difference of the colors and artistic style. The TV loss is used in suppressing unnecessary noise generated in texture synthesis. In the image restoration, we use these loss functions to ensure the synthetic content in the masked region is consistent to the ground truths in term of texture, colors, artistic style and free of unnecessary noises. The optimizer uses value of these loss combination to tune network parameters and narrows distances between predicted images and ground truths.

In summary:

1) We applied the end-to-end networks with partial convolutional layers onto a wall-painting dataset in different styles and presented a state-of-the-art performance.

2) We create a reasonable data collection from the Dunhuang Grottoes Paintings which enables image inpainting through data driven approach.

3) We provide proper analysis and comparison on the results using two types of masks, which are designed for simulating deteriorations.

2. Related Works

Image restoration has been one of the focuses in computer vision community for decades; many researchers have proposed fruitful methods in the literature. Nonetheless, due to challenges from the rich information in visual signal and the almost infinite possible ways of inpainting the loss content, and partly because of efficiency requirement and lack of groundtruth information, the automatic image restoration technique for some specific problems still need further investigation. Methods in this domain could be classified into two categories: conventional image processing based techniques and convolutional networks based methods.

2.1. Conventional methods

Early proposed methods are based on the diffusion techniques. These diffusion based approaches [1][3] [9] fill in the to-be-repaired holes by propagating the semantic information in its peripheral regions. The propagation of semantic information can be isophote direction field [1][3] or relies on statistical illumination or color features [9]. These early proposed methods, though usually be applied in dust removal task in film scanning, can only coarsely in-paint small holes from such as tiny molds in paintings and photo scratches. The more sophisticated patch-based methods, which are capable in restoring larger deteriorated regions, out-performance diffusion-based methods in image inpainting and set the new base-line performance onto a higher level. The first patched-based method [4] proposed a texture synthesis framework to use a novel copy-paste scheme. The copy-paste scheme searches possible patches from images in source dataset and paste the patch into loss region in the target image. Some methods [2] [8] [16][13][5] follows the patch-based framework and further introduced optimization methods to smooth the inconsistency between synthetic and original textures. Particularly, PatchMatch [2] has greatly reduce computational cost and increase the processing speed to the sub-real-time level. Rather than synthetically generating learned textural content, patch-based methods rely on matching local pixels or their texture features; and thus unable to restore textures or objects which are not exactly in source dataset. Another drawback is that, when putting into practice, the source dataset must be used along with patch-based methods.

2.2. CNN based methods

The state-of-the-art image restoration techniques are driven by the convolutional neural networks, which has set new base-lines in many signal processing areas in recent years. Early methods [11][17] focus on inpainting a regular rectangular blank patch in the center of a target image. The Context Encoder [11] was the first proposed to use an asymmetric end-to-end convolutional neural network (CNN), in which input an 128×128 image with blank patch on one end and output the estimated 64×64 patch on the other end. Context Encoder encodes visual information of non-blank region and maps it onto the groundtruth content in the center by taking the advantages of the powerfully feature embedding capability of the convolutional networks. Yang et al. [17] extended Context Encoder by introducing multi-scale neural patch synthesis approach based on joint optimization of image content and texture constraints. The improvement are reflected on the semantic fine details in the filled content. Song et al. [15] proposed to use an stacking Multi-scale inference and Patch-Swap operation to refine the semantic texture in restored regions. Iizuka et al. [6] and Yu et al. [18] drop the assumption of centered rectangular blank patch and more flexibly assume inpainting regions can be in non-rigid shapes. The regions to be in-painted are given in the form of mask. One of the side-product advantages is that these methods reduces the risk of over-fitting the rectangular shape of blank patches. Iizuka et al. [6] use generative adversarial framework with two discriminators for judging the local synthetic texture and global generated image respectively. Yu et al [18] extended [6] by integrating an attention mechanism. Lately, Liu et al. [10] proposed Partial Convolution, which merges the mask and image gradually in partial convolutional down-sampling in encoder and deconvolutes the encoded low-level features onto the global predicted texture. The local in-painted texture are further merged with original undeteriorated texture to form a final restored image.

3. Methodology

To automatically in-paint the deteriorated regions with synthesized content, it is desirable to build up an end-to-end neural network, which output a map containing the synthesized content in the corresponding regions without additional post-processing. The end-to-end neural network takes in the original image and the mask at one end; then output a predicted image as the same size of the original image at the other end. Using the powerful capability of regression and prediction of CNN, the method avoids hand-designed pre-processing optimization for the restored se-

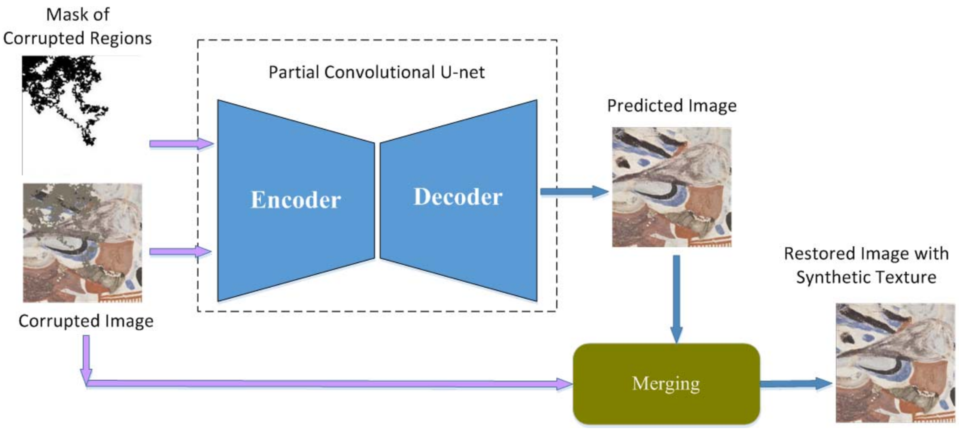

Figure 2. The training procedure of the image restoration method.

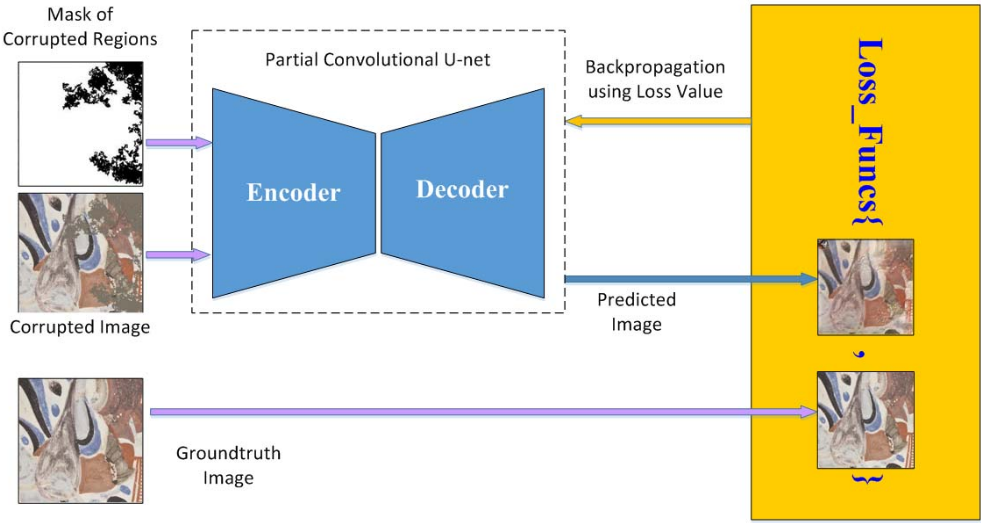

Figure 3. The inference and restoration procedure of the image restoration method.

mantic content. 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 the width and height of the network input is N, 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, it will be re-scaled to be square so that the ratio is consistent with the network input. The mask M, which is of size N, labels whether pixels belongs to deteriorated regions or intact regions. The labeled deteriorated regions are generally in irregular non-rigid shapes. The output image contains textural and stylistic information of both labeled deteriorated region and the unlabeled region. The end-to-end networks 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. We will introduce the network architecture, the loss functions and the implementation details following subsections.

3.1. Network Architecture

To generate the synthesized content for deteriorated regions of the Dunhuang Grottoes, we take the advantages of the U-like end-to-end network with skip connections and partial convolutions(PConv)[10]. The network architecture, a variant of encoder-decoder configurations, is shown in figure 2.

The encoder and decoder are not mirrored in symmet-

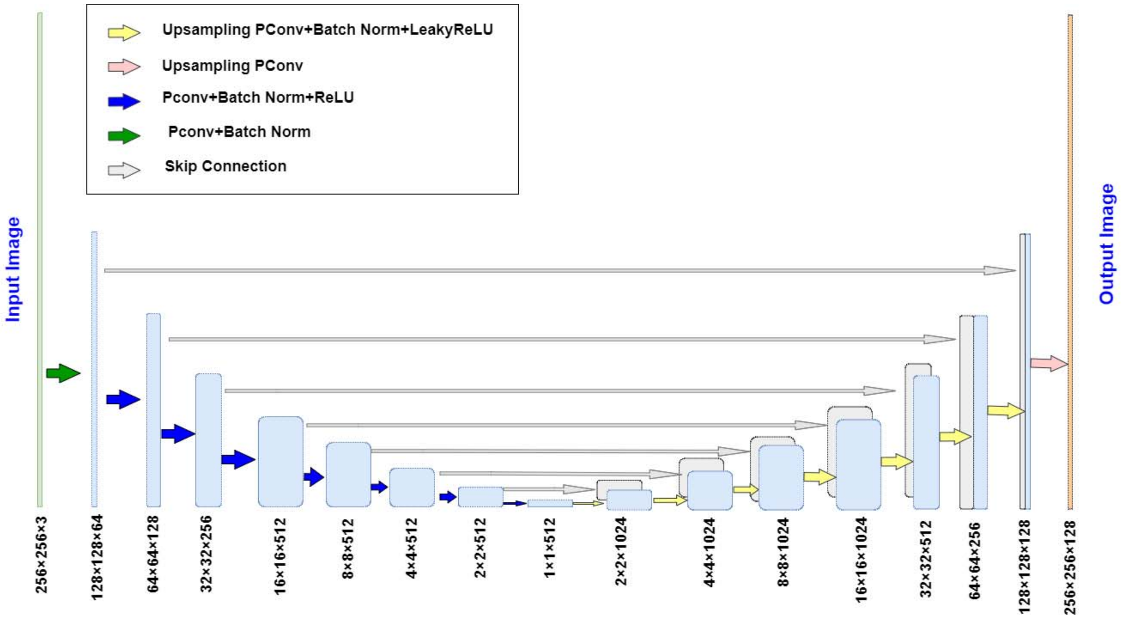

Figure 4. The architecture of the proposed networks.

ric structure like the conventional U-nets [12]. The encoder contains 8 partial convolutional layers and the size of kernel 7, 5, 5, 3, 3, 3, 3 and 3. The channel sizes in the layers are 64, 128, 256, 512, 512, 512, 512, and 512. 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 channel sizes in the decoding partial convolutional layer are 512, 512, 512, 512, 256, 128, 64, and 3 with kernel sizes 7, 5, 5, 3, 3, 3, 3 and 3 respectively. Different from using ReLU in the encoding, the LekyRuLU activation function is used in decoding with parameter alpha = 0.2.

The skip connections directly pass feature maps from ith layer of encoder to (8-i)th layer of decoder. Feature maps from encoder are concatenated to the feature maps from decoder in the axis of channels. In the conventional encoder-decoder networks, when the input went through eight down-sampling layers, the information pass through the bottleneck layer is limited. In the reverse process, which is up-sampling, the decoder may not have enough features to effectively recover most details for an end-to-end image generation task. To this end, using skip connections across networks could overcome this limitation. A skip connection builds up a pipeline for sharing the low-level features from layers in the encoder to the corresponding layers in decoder. Thus, it helps the decoder recover more details in the output.

The segment-aware based partial convolution [10] is essentially an multiplication-based conditioning method. The mask, which labels pixels in deteriorated foreground regions as 0 and those in intact background as 1, serves as a conditional inverse attention for the networks. In other words, the intact background regions are to be learned in the convolutional layer while the deteriorated foreground regions are to be ignored. The mathematical formula of partial convolution operation is given as follows:

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

where X' and M' are input feature maps and mask in the receptive field; $\circ$ is the element-wise multiplication; W_T and b is the weight and bias of a filter; 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:

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

Each partial convolutional block downsamples the feature maps and the mask. When mask reach to the bottleneck of the U-like networks, the value in the mask will be all 1s, which means all the masking information are fused into the embedding low-level features. During the down-sampling and decaying process, decayed mask and partial convolution not only smooth feature maps but also fill the vacant regions in the subsequent feature maps, in which is of all zeros in the first layer.

3.2. Loss

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:

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

where $L_{content}$, L_{style} and L_{TV} are the content loss, style loss [7] and total variation (TV) loss; and $\lambda_{content}$, λ_{style} and λ_{TV} are the balancing coefficients for the corresponding loss respectively. For the content loss and style loss, we employ VGG-16 ImageNet pre-trained networks [14] as the loss network ϕ to extract the deep feature maps. The pre-trained loss network ϕ has already learned to encode semantic and perceptual information using ImageNet dataset. Thus, the network is no longer trained or updated in the training stage. We use the one pass of feed forward network ϕ to obtain activation maps of a given image I from first four convolutional blocks; then the activation maps are reshaped into deep features in the form of 1-D vectors, denoted as $\{\phi(I)_i\}; i = 1, 2, 3, 4$.

Content Loss. Taking advantages of the deep features, the content distance of i^{th} layer-wise features is the Euclidean distance of the corresponding two vectors. The content loss in our task, taking accounts of distances from each pair of images (I_{gt}, I_{out}) and (I_{gt}, I_r), is defined as summation of content distances of all levels, mathematically shown as follows:

$$\begin{aligned} L_{content} = \sum_{i}^{1,2,3,4} \frac{1}{C_i N_i N_i} [& |\phi(I_{gt})_i - \phi(I_{out})_i|_2 + \\ & |\phi(I_{gt})_i - \phi(I_r)_i|_2] \end{aligned} \qquad (4)$$

where C_i and N_i are the number of channels and the length of a square feature map output from i^{th} layer. The content loss allows us to measure the differences of compoenential content and overall spatial structure between a pair of images while style loss measures the differences of stylistic characteristics, like colors, textures, common patterns.

Style Loss. The style loss is computed using style features, which is obtained by further computing autocorrelation (Gram matrix) [7] of the deep feature from the pre-trained network. Let $G_i^{\phi}(I)$ be the Gram matrix of i^{th} layer-wise deep features of a given image $\{\phi(I)_i\}; i = 1, 2, 3, 4$; the elements of $C_i \times C_i$ Gram matrix of $\phi_i(I)$ is computed as follows:

$$G_i^{\phi}(I)_{c,c'} = \frac{1}{C_i N_i N_i} \sum_{h=1}^{N} \sum_{w=1}^{N} \phi_i(I)_{w,h,c} \phi_i(I)_{w,h,c'} \qquad (5)$$

Similar to content loss, the style loss takes account of the differences both in (I_{gt}, I_{out}) and (I_{gt}, I_r):

$$\begin{aligned} L_{style} = \sum_{i}^{1,2,3,4} [& |G_{i\ gt}^{\phi}(I) - G_{i\ out}^{\phi}(I)|_2^F + \\ & |G_{i\ gt}^{\phi}(I) - G_i^{\phi}(I_r)|_2^F] \end{aligned} \qquad (6)$$

where $| \cdot |_2^F$ is the the squared Frobenius norm. Using distances content and style feature from pre-trained network rather than pixel-wise distance between two images, the network would avoid learning hard matching of pixels and focus on generalizing the perceptual visual information. Thus, by reducing the content and style loss of the output/restored image and the groundtruth image, the end-to-end network outputs are gradually optimized to narrow the perceptual gaps between ground true and the restored images.

Total Variation Loss. To encourage the spatial smoothness in the restored region P, the total variation (TV) regularizer is adopted as a loss term. The TV loss of region P given the restored image I_r is as follows:

$$\begin{aligned} L_{TV}(I_r|P) = \sum_{(i,j)\in P} [& |I_r^{(}i,j) - I_r^{(}i+1,j)|_1 + \\ & |I_r^{(}i,j+1) - I_r^{(}i,j)|_1] \end{aligned} \qquad (7)$$

3.3. Training and Implementation

We employed two stages to train the partial convolutional end-to-end network: 1) the first stage is to pretrain a partial convolutional network with diverse low-level feature-extracting capability; 2) the second stage will finetune the pre-trained model to fit in our grottoes restoration task.

In the first stage, the partial convolutional network is trained on the Place2 dataset [19] to obtain a pre-trained model. The pre-training allows the network to generalize its low-level filters on a diverse dataset so that it could extracted various deep features for the latter stage. The Place2 dataset contains A 10 million image, in which cover numerous different kinds of texture. As the Place2 dataset is large enough to contains 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. During pre-training, we use Adam as optimizer and set the learn rate to 2e-4. The size of mini batch is 16. The weighting coefficients $\lambda_{content}$, λ_{style} and λ_{TV} for corresponding loss functions are set to be 0.05, 1000, and 0.1 respectively. The loss value is back-propagated through all parameters in the network.

In the second training stage, the network is fine-tuned by fixing some weights in low level filters. More specifically, the parameters of batch normalization layer in the encoder of the network are frozen and no longer to be updated. Data augmentation techniques are used in pre-processing the input images in order to avoid over-fitting the style and dynamically generate more training samples. The augmentations include 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 parameter settings of fine-tuning are similar to the pre-train stage except that the learn rate decreases to 5e-05.

The proposed method is implemented in PyTorch 0.4. The input and output sizes of the end-to-end networks is 256x256. Images of 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@3.7GHz, 16GB RAM, Ubuntu 16.04 OS, nvidia GTX Titan Xp with 11GB memory. During pre-training and fine-tuning, it takes around 0.9s to process each iteration, which includes forward inference and back-propagation of network updating. It takes 20000 iterations and around 10 hours to pre-train a model; and more than 9000 iterations and 5 hours to fine-tune an effective image restoration model for our wall-paintings dataset.

4. Experimental Results

In the experiments, the implemented method is tested on image samples cropped from raw FeiTian/FlyingSky data which is delicately photographed at the sites and provided by Dunhuang academy. After generating FeiTian/FlyingSky dataset, we conducted two experiments, which consist of: 1) we compare the results of the method perform on two different types of deterioration masks; 2) Comparison on two different types of deterioration masks.

4.1. Experimental Settings and Dataset

To generate large scale cropped image sample from FeiTian/FlyingSky imageset, randomized cropping, rotation and re-scaling are used on the raw images. The cropped 10000 images are splitted into train set and validation set with rough proportion of 4:1. Although some image samples may contain deteriorated texture caused by natural aging, the general image quality are good enough to carry sufficient visual knowledge of artistical content. During training, the image, masked image and the mask form a triplet input sample; while in testing only masked image and the mask are needed.

4.2. Comparison on two different types of deterioration masks

Two types of masks, which are called dusk-like mask and jelly-like mask, are generated to fully test the performance of texture synthesis. In the domain of grottoes restoration in e-Heritage protection, the dusk-like masks simulate deterioration by molds or salting erosion, while the jelly-like masks simulate physical damages or sabotages. The generation process of the dusk-like masks follows these steps: Step 1) Initialize a square blank image with all value set as 1. This blank image serves as a canvas for drawing mask. The size of initial mask is 256x256. Step 2) Randomly pick a start point P_0 on the blank image, and set the pixel value to 0. Step 3) Iteratively perform random walk from P_i to P_{i+1}. Once a pixel is traversed, its value will be set to 0. Note that a pixel is allowed to be walked on more than 1 time. The default number of walk steps is 10,000. The latter type of jelly-like mask, based on the dusk-like masks, is further processed by removing small noises and reserve the irregular block-like regions using open-close functions and image erosion. Fig. 4 and 4 provide the results using dusk-like mask and jelly-like mask respectively. In these results, the style and color are mostly the same as those in the groundtruth. Even if the textural content may not be exactly the same as the original content, those inpainted texture is fine enough to fool human perceptions and the regional context are mostly consistent. For those masked by the dusk-like masks, the synthetic texture in the single-pixel-width random line are so well-blended with region context that sometimes even unnoticeable. We also noticed that, for large block of masked region, the details may not be to fully recovered, which reflected on the relative failure in inpainting some fine details.

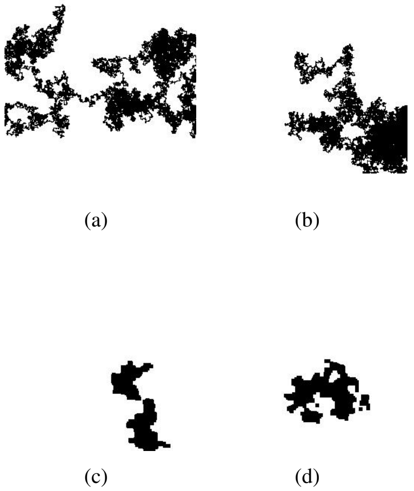

Figure 5. Two types of masks: a) and b) are dusk-like masks; c) and d) are jelly-like masks.

5. Conclusions

In conclusion, we focused on training a deep end-to-end neural network model to in-paint deteriorated regions and restore the historical painting of Dunhuang Grottoes. The

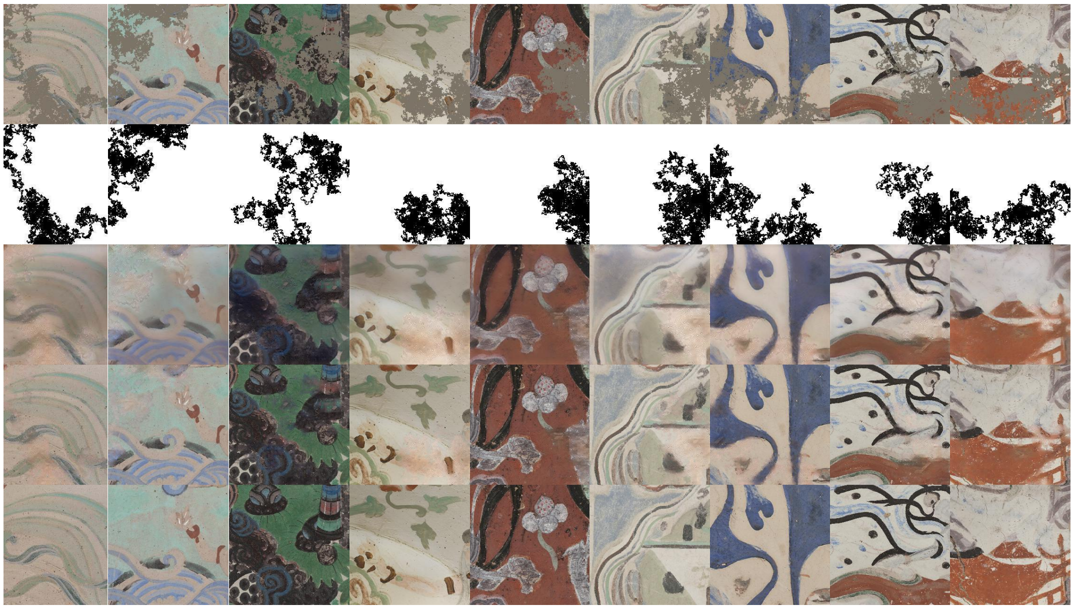

Figure 6. More testing results using dusk-like masks. Each column represents the related data of a image sample. From row 1 to row 5: deteriorated input image, input mask, output predicted image, merged final result, groundtruth image

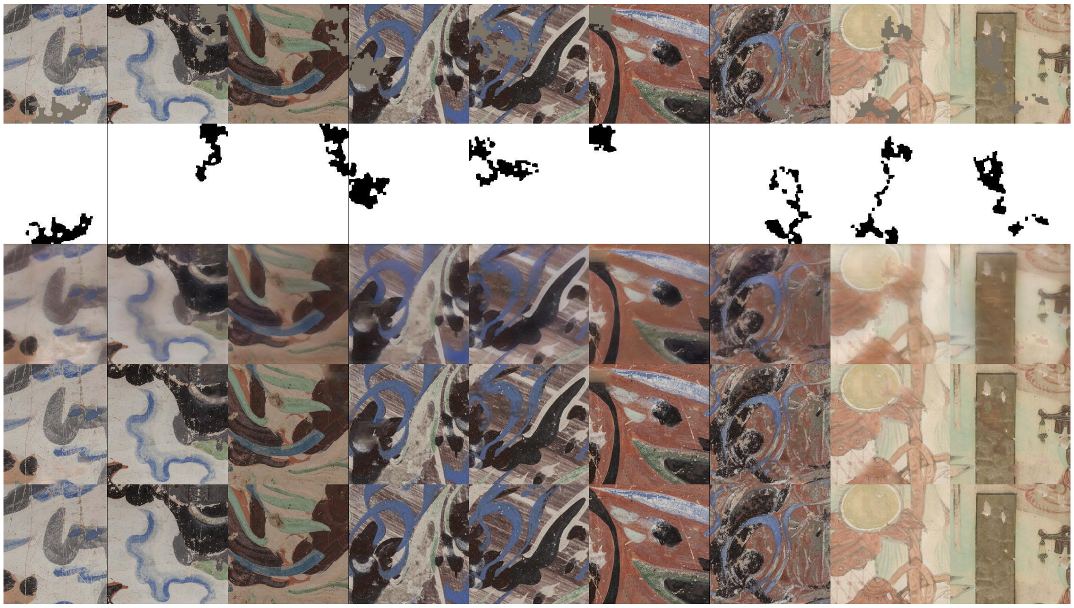

Figure 7. More testing results using jelly-like masks. Each column represents the related data of a image sample. From row 1 to row 5: deteriorated input image, input mask, output predicted image, merged final result, groundtruth image

end-to-end image restoration model employs U-net with partially convoluational layers to reconstruct lost content, which is capable in restoring non-rigid deteriorated content given a lost content mask and a wall-painting image. In order to reduce the difference between the synthetic content in the masked region and the groundtruths in term of texture, colors, artistic style and free of unnecessary noises, a hybrid loss function is used in optimization. Implemented model is fully tested in restoring highly non-rigid and irregular deteriorated regions, using two types of masks designed for simulating deterioration. The experimental results are satisfactory and have shown the method is capable in restoring the loss content properly.

References

[1] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on Image Processing*, 10(8):1200–1211, Aug 2001.

[2] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24:1–24:11, July 2009.

[3] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 417–424, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[4] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1033–1038 vol.2, Sep. 1999.

[5] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf. Image completion using planar structure guidance. *ACM Trans. Graph.*, 33(4):129:1–129:10, July 2014.

[6] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4):107:1–107:14, July 2017.

[7] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing.

[8] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra. Texture optimization for example-based synthesis. *ACM Trans. Graph.*, 24(3):795–802, July 2005.

[9] Levin, Zomet, and Weiss. Learning how to inpaint from global image statistics. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 305–312 vol.1, Oct 2003.

[10] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 89–105, Cham, 2018. Springer International Publishing.

[11] D. Pathak, P. Krhenbhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, June 2016.

[12] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[13] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[15] Y. Song, C. Yang, Z. Lin, X. Liu, Q. Huang, H. Li, and C.-C. J. Kuo. Contextual-based image inpainting: Infer, match, and translate. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 3–18, Cham, 2018. Springer International Publishing.

[16] Y. Wexler, E. Shechtman, and M. Irani. Space-time completion of video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):463–476, March 2007.

[17] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4076–4084, July 2017.

[18] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, June 2018.

[19] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, June 2018.