Coarse-to-Fine Generative Model for Oracle Bone Inscriptions Inpainting

Shibin Wang *[1,2], Wenjie Guo[1], Yubo Xu[1,2], Dong Liu[1,2], and Xueshan Li[1,2]

[1]School of Computer and Information Engineering, Henan Normal University (Henan), China.
[2]Oracle Bone Intelligent Computing Laboratory, Henan Normal University (Henan), China.

Abstract

Due to ancient origin, there are many incomplete characters in the unearthed Oracle Bone Inscriptions(OBI), which brings the great challenges to recognition and research. In recent years, image inpainting techniques have made remarkable progress. However, these models are unable to adapt to the unique font shape and complex text background of OBI. To meet these aforementioned challenges, we propose a two-stage method for restoring damaged OBI using Generative Adversarial Networks (GAN), which incorporates a dual discriminator structure to capture both global and local image information. In order to accurately restore the image structure and details, the spatial attention mechanism and a novel loss function are proposed. By feeding clear copies of existing OBI and various types of masks into the network, it learns to generate content for the missing regions. Experimental results demonstrate the effectiveness of our proposed method in completing OBI compared to several state-of-the-art techniques.

1 Introduction

Since the earliest discovery of Oracle Bone Inscriptions(OBI), over 5,000 distinct character forms have been identified, which have significantly advanced our comprehension of many characters' meanings. These deciphered OBIs provide invaluable historical information crucial for understanding various aspects of ancient Chinese politics, society, religion, and more.

Recognizing and interpreting are important topics in the field of OBI research. Due to the lack of physical objects, the images of rubbings in the recorded books are the main carriers of research. However, some OBIs have suffered varying degrees of residual erosion and damage on their surface, resulting in a large number of incomplete fonts in the inscriptions and rubbings seen today. With the rapid development of image generation technology, many image restoration problems difficult to solve in traditional methods have found new research avenues. The comprehensive application of artificial intelligence and other technologies has become a new research direction in the restoration of OBIs.

Zeng et al. (2019) proposed the Pyramid Context Encoder Network (PEN). It is based on the U-Net structure and encodes and decodes contextual semantics to ensure visual and semantic consistency. Li et al. (2020) developed the Recurrent Feature Reasoning (RFR) network, featuring a plug-and-play RFR module and a Knowledge Consistent Attention (KCA) module. They infer the hole boundaries and capture the distant feature information. Wu et al. (2021) introduced a two-stage (coarse-to-fine) model. It combines a Local Binary Pattern (LBP) Waller et al. (2013) network and incorporates a new spatial attention mechanism. These methods have enhanced image processing. However, they only grasp limited connections between textures and edges. They fail to fully comprehend image semantics and complex structures. Additionally, they overlook the interplay between global and local features. Given the complexities behind incomplete fonts and unique font features, existing image restoration models struggle to effectively complete OBI image inpainting tasks.

To meet these challenges, we propose a two-stage (coarse-to-fine) font inpainting network. Our network incorporates a dual discriminator structure to capture both global and local image information. Specifically, we employ a global discriminator to focus on the spatial correlation between damaged and undamaged regions. The local discriminator concentrates on the local patch information. To effectively understand the intrinsic features of the image, we introduce a novel loss function to accurately restore the structure and details. Through extensive comparisons, our framework demonstrates

*Corresponding author: wangshibin@htu.edu.cn

Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), pages 107–114

state-of-the-art performance in OBI image inpainting tasks.

2 Method

2.1 Network Architecture

The network is a two-stage deep generative model. Both stages consist of encoder-decoder pipeline and follow an adversarial model Goodfellow et al. (2014). The network architecture is shown as Figure. 1. The damaged image consists of the missing regions filled with white pixels, represented as I_{in}. L_{in} denotes the LBP Waller et al. (2013) structural information extracted from the damaged oracle I_{in} in the grayscale channel. M represents a binary mask, where 1 indicates the missing regions and 0 indicates the known regions.

In the first stage, the generator G_1 includes seven feature extraction blocks and feature restoration modules. Each feature extraction block consists of LeakyReLU Xu et al. (2015), a convolutional layer, and InstanceNorm2d Ulyanov et al. (2016). The decoder generates the content of the missing region through seven feature restoration modules, which consist of ReLU Nair and Hinton (2010), transposed convolution, and InstanceNorm2d Ulyanov et al. (2016). Finally, G_1 and D_1 generate the completed LBP structural information L_{out} and L_o.

In the second stage, an additional spatial attention layer is added to the fifth layer of the encoder. This layer builds the correlations not only within the known region but also among the missing regions.

Due to a single discriminator judging the image authenticity solely from a global perspective and being unable to handle the details, artifacts and structural inconsistencies may arise in the restoration results. The dual discriminator, on the other hand, judges the image from both global and local perspectives. They compete with each other to learn more effective weights.

2.2 Dual Discriminator

The structure of Dual PatchGAN Isola et al. (2017) Discriminator (DP) is as shown in Figure. 2. The left branch is a global discriminator that focuses on the spatial correlation between damaged and undamaged regions. Its input consists of an image and a mask, and output is a 3D feature. Each CSL block consists of a 5×5 convolution, SpectralNorm Miyato et al. (2018) and LeakyReLU with α= 0.2. In the first two CSL blocks, the number of convolutional output channels is 64 and 128, while in the others it is 256. The right branch is a local discriminator with five 4×4 convolutions, which focuses on the local patch. The first four layers use the LeakyReLU with α= 0.2, the Sigmoid for the last layer and the BatchNorm2d for normalization in the middle three layers. The local discriminator can be formulated as:

$$\begin{aligned}\tau_{\text{adv2}} = &\min_{G_2}\max_{D_2}\mathbb{E}_{I_g}\left[\log D_2(I_{\text{g}})\right] \\ &+\mathbb{E}_{I_{in}}\left[\log\left(1-D_2\left(G_2\left(I_{\text{in}}, M\right)\right)\right)\right]\end{aligned} \tag{1}$$

Our objective function for the global discriminator can be formulated as:

$$\tau_{\text{adv3}} = -\mathbb{E}_{I_{in}\sim\mathbb{P}_{I_{in}}(I_{in})}\left[D_3\left(G_2\left(I_{\text{in}}\right)\right)\right] \tag{2}$$

$$\begin{aligned}\tau_{\text{D}_3} = &\mathbb{E}_{I_g\sim\mathbb{P}_{data}(I_g)}\left[\text{ReLU}\left(1-D_3\left(I_g\right)\right)\right] \\ &+\mathbb{E}_{I_{in}\sim\mathbb{P}_{I_{in}}(I_{in})}\left[\text{ReLU}\left(1+D_3\left(G_2\left(I_{\text{in}}\right)\right)\right)\right]\end{aligned} \tag{3}$$

where G_2 represents the second stage generator, D_2 and D_3 represent the right and left branches of the dual discriminator, respectively.

2.3 Multi-level Fusion Loss Function

We reduce the difference between the original image and the inpainting image by using a multi-level fusion loss function(MLFLF) to enhance the stability of training.

The reconstruction loss is defined as:

$$L_{\text{r}} = \left\|L_{\text{o}} - L_{\text{g}}\right\|_2 \tag{4}$$

$$L_{\text{o}} = L_{\text{in}} \odot (1-M) + L_{\text{out}} \odot M \tag{5}$$

The adversarial loss Yan et al. (2018) is defined as:

$$\begin{aligned}\tau_{\text{adv1}} = &\min_{G_1}\max_{D_1}\mathbb{E}_{L_g}\left[\log D_1(L_{\text{g}})\right] \\ &+\mathbb{E}_{L_{in}}\left[\log\left(1-D_1\left(G_1\left(L_{\text{in}}, M\right)\right)\right)\right]\end{aligned} \tag{6}$$

The pixel-level reconstruction loss is responsible for directly comparing each pixel of the generated image with the target image:

$$\begin{aligned}&\text{L}_{\text{valid}} \\ &= \frac{1}{\text{Sum}(1-M)}\left\|(L_{\text{out}} - L_g)\odot(1-M)\right\|_1\end{aligned} \tag{7}$$

$$\text{L}_{\text{hole}} = \frac{1}{\text{Sum}(M)}\left\|(L_{\text{out}} - L_g)\odot M\right\|_1 \tag{8}$$

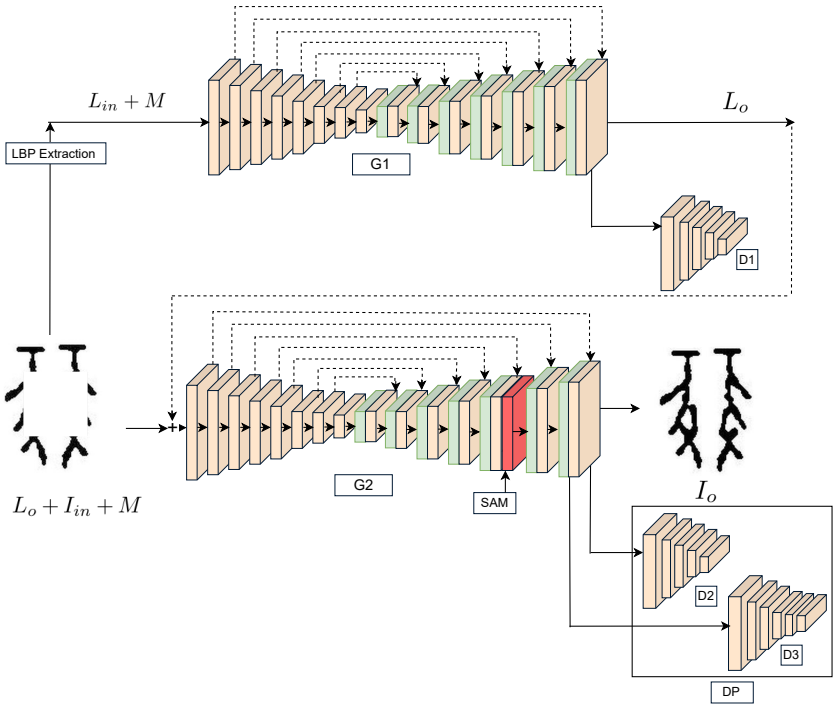

Figure 1: The network architecture of our proposed method.

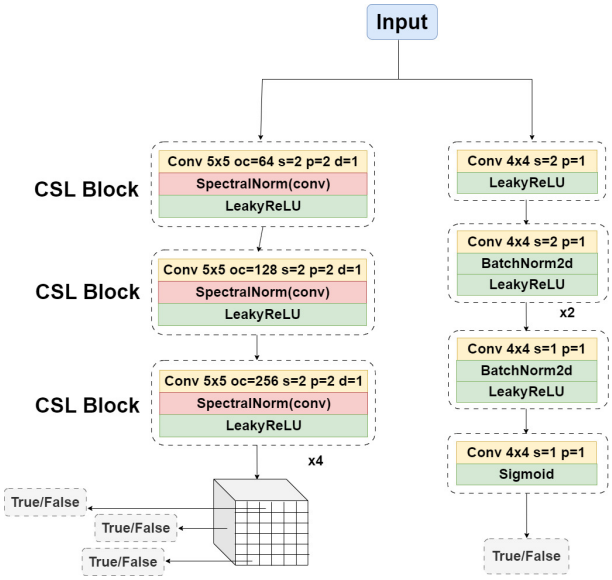

Figure 2: Proposed DP to introduce efficient local and global consistencies.

The Total Variation (TV) Liu et al. (2018) loss reduces noise and discontinuities, resulting in a smoother and more continuous appearance:

$$L_{\text{tv}} = \|L_o(i,j+1) - L_o(i,j)\|_1 + \|L_o(i+1,j) - L_o(i,j)\|_1 \quad (9)$$

The multi-scale loss compares the differences between ground truth images and mapping results of different scales:

$$\tau_m = \sum_{h \in d} \|\Phi_h(I_o) - \Phi_h(I_g)\|_2 \quad (10)$$

$$I_{\text{o}} = I_{\text{in}} \odot (1-M) + I_{\text{out}} \odot M \quad (11)$$

We apply the perceptual loss Johnson et al. (2016) and style loss Gatys et al. (2016) defined on the VGG-16 Simonyan and Zisserman (2014) (pre-trained on ImageNet Deng et al. (2009)) to enhance the recovery of structural and textual information.

$$I_{\text{per}} = \sum_i \|\Psi_i(I_o) - \Psi_i(I_g)\|_1 \quad (12)$$

$$I_{\text{style}} = \sum_i \|\delta_i(I_o) - \delta_i(I_g)\|_1 \quad (13)$$

where Ψ_i is the feature map of $i-th$ layer in ImageNet-pretrained VGG-16 network, $\delta_i(\cdot) = \Psi_i(\cdot)\Psi_i(\cdot)^T$ is from (Buades et al., 2005).

3 Experiments

3.1 Dateset

We select 2000 OBI images for training and 100 for testing from the oracle bone images produced by the Key Laboratory of Oracle Information Processing of the Ministry of Education in Henan Province. To better validate the results of the experiment, we use the masks to simulate the broken regions of

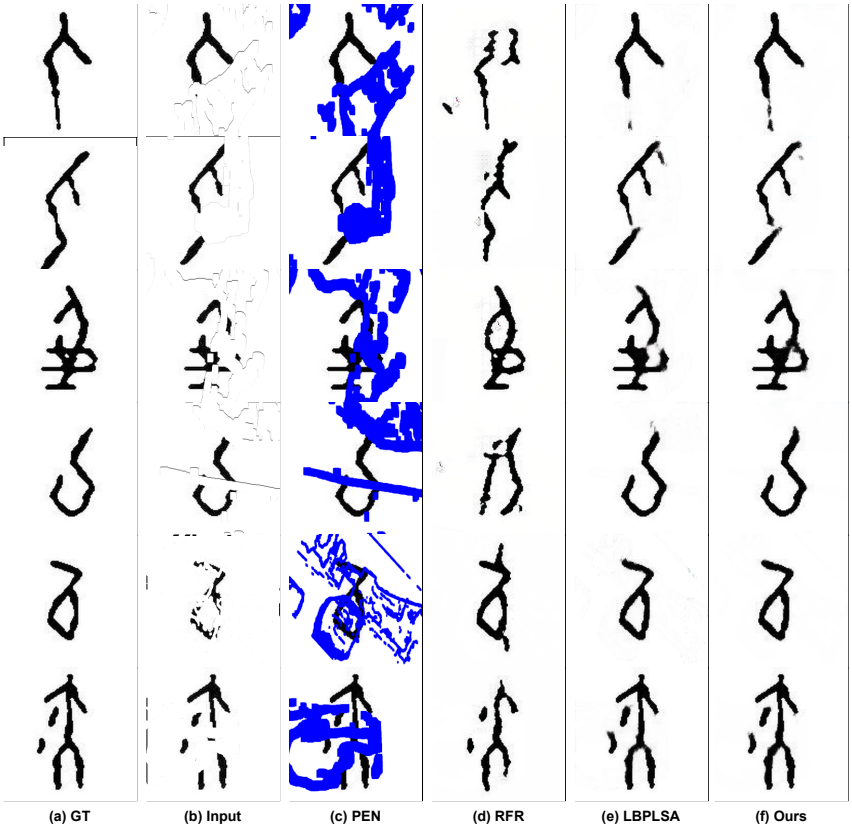

Figure 3: Comparison of qualitative results between the proposed method and other approaches on the irregular mask. Our proposed method generates more effective structural and texture information.

OBIs. These masks are divided into irregular and regular types. The irregular masks are obtained from the NVIDIA dataset Liu et al. (2018), while the regular masks are square masks of fixed size (25% of the total image pixels) placed in the center of the image.

3.2 Qualitative Comparisons

In this section, we conduct the experimental comparisons with other image restoration models.

For the restoration of OBI with irregular masks, the visualization results are shown in Figure. 3. The input image (b) shows the damaged OBI images. (c) demonstrates the results of using the PEN network Zeng et al. (2019) with a mode collapse. (d) using the RFR network Li et al. (2020) fails to accomplish the complementation task effectively. Note in particular the comparison between (e) the LBPLSA network Quan et al. (2022) and (f) our network. Our network evidently produces more realistic completion results from the smoother strokes in the first row of Figure. 3. And fewer or no artifacts appear at the end of the strokes in the rest of the lines. In contrast, the LBPLSA network exhibits severe artifacting and discontinuities in strokes. It fails to adequately complete the objectives. The presence of artifacts indicates that the network did not accurately understand the missing content in the image. As a result, it fills in unrealistic textures and structures.

We also explore the classic center mask completion scenario in image inpainting. Given that most of the OBI content lies in the center, it is challenging for the network to infer the main content of the characters from just one stroke at the boundary. The generated results are depicted in Figure. 4. We can see that the generated results of the PEN network (c) collapse again and the RFR network (d) fails to meet the target requirement. Focus on the comparison between LBPLSA (e) and our method (f) again, LBPLSA generates the images with more artifacts and doesn't effectively learn the semantic information of the OBIs. For instance, in the second row of Figure. 4, the strokes generated by the LBPLSA are opposite to the ground truth. More artifacts are present in rows 5 and 6. Under the same experimental configuration, our network achieves results that are closer to the ground truth.

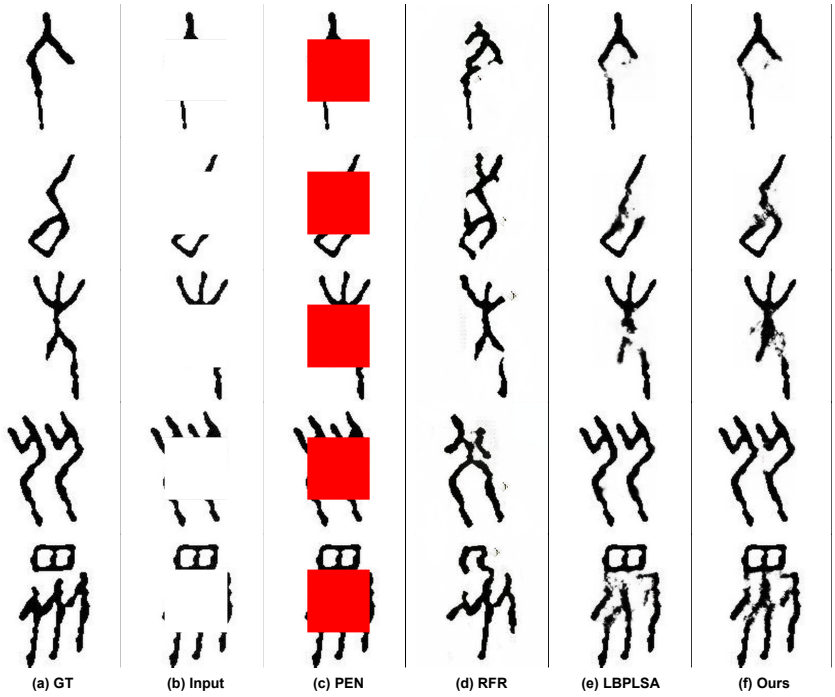

Figure 4: Comparison of qualitative results between the proposed method and other approaches on the rectangle mask. Our proposed method generates more effective structural and texture information.

	method	PEN	RFR	LBPLSA	LG	Ours
PSNR+	Irregular	8.67	14.67	25.67	27.01	**29.61**
	rectangle	9.08	14.30	26.73	22.46	**33.27**
SSIM+	Irregular	0.6337	0.8507	0.9719	0.9781	**0.9826**
	rectangle	0.7800	0.8397	0.9497	0.9449	**0.9623**
L1-	Irregular	0.1628	0.0567	0.0091	0.0066	**0.0058**
	rectangle	0.1425	0.0598	0.0162	0.0197	**0.0143**

Table 1: Comparison between the proposed method and state-of-the-art methods on the oracle dataset (+ indicates higher is better, - indicates lower is better).

3.3 Quantitative Comparisons

In terms of evaluation metrics, we follow the structural similarity index (SSIM) Wang et al. (2004) and peak signal-to-noise ratio (PSNR) as outlined in references Ren et al. (2019). The evaluation results are presented in Table 1.

Compared with other methods, the scores of each indicator in our model have been improved. The DP structure can effectively capture both the global and local image information. Additionally, the loss function component introduced MLFLF optimizes semantic plausibility and structural consistency. The integration of DP structure and MLFLF component produces the images with reduced pixel-level differences and leads to significant improvements across SSIM, PSNR, and L1 distance metrics, which indicates the high accuracy and effectiveness in image inpainting tasks.

3.4 Ablation Studies

The ablation studies are conducted under mask rates ranging from 20% to 30%. We evaluate the effectiveness of our proposed method by contrasting three different experimental settings, including the LBPLSA method, the SN method only with DP component and the complete method. The generated results are depicted in Figure. 5. Part (a) represents the ground truth OBIs. The input images with various degrees of damage are generated by masks, shown in (b). The completion results of LBPLSA (c), SN (d), and ours (e) are sequentially displayed.

Compared with the LBPLSA method, the SN method shows some improvement with the introduction of the DP structure. The incorporation

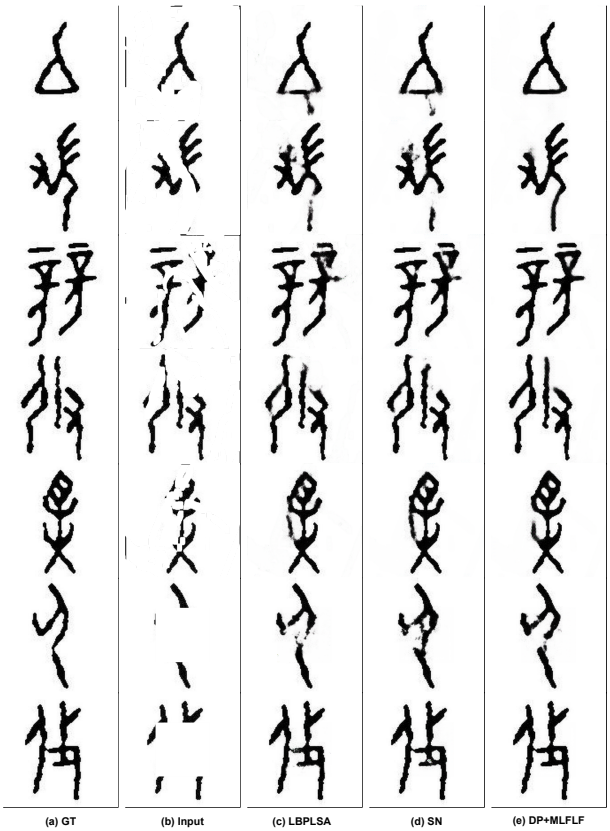

Figure 5: Qualitative results comparison of ablation study.

	method	NO	DP	DP+MLFLF
PSNR+	Irregular	25.67	28.66	**29.61**
	rectangle	26.73	26.88	**33.27**
SSIM+	Irregular	0.9719	0.9798	**0.9826**
	rectangle	0.9497	0.9566	**0.9623**
L1-	Irregular	0.0091	0.0068	**0.0058**
	rectangle	0.0162	0.0159	**0.0143**

Table 2: Quantitative results of ablation study on oracle dataset. (+ indicates higher is better, - indicates lower is better).

of the DP structure enables the model to better capture both global and local image information, which improves the restoration results to a certain extent. However, the SN method lacks the further optimization from the MLFLF component. It still has certain limitations and fails to fully exploit the intrinsic features of the images.

Furthermore, our complete method achieves further improvements across all metrics within the experimental scope. By leveraging the dual advantages of DP and MLFLF, our method can more accurately restore the structure and details of the images. This makes the restoration results closer to the original images. Compared to the methods only with DP, the addition of the MLFLF component further enhances the clarity and quality of the restored images. This leads to better performance across metrics, such as SSIM, PSNR, and L1 distance, as demonstrated in the ablation study metrics presented in Table 2.

Through the ablation studies, we validate the crucial roles of DP and MLFLF in image restoration tasks. The DP structure enhances the model's understanding of images, while the MLFLF module further optimizes detail and texture restoration. This showcases significant advantages across all metrics. These experimental results validate our method's effectiveness. They emphasize the importance of leveraging the dual advantages of DP and MLFLF in image inpainting tasks.

4 Conclusion

We propose the two-stage (coarse-to-fine) network for efficient OBI image inpainting. This new framework consists of an enhanced LBP network and integrated DP and MLFLF components. Specifically, we design a novel dual discriminator network. The first stage LBP learning network adopts a U-Net architecture, aimed at accurately predicting structural information in missing regions. This guides the second image inpainting network in better filling missing pixels. In the second stage image generation network, we employ dual discriminators to complete the masked regions. Compared to several state-of-the-art methods, experimental results demonstrate the effectiveness of DP and MLFLF components in the proposed method in completing OBI image inpainting tasks.

In the future, we plan to further develop our network to achieve more powerful functions, such as increasing the speed, realizing editing functions, and improving the efficiency of paleographers. Our goal is to solve the problem of more complex noise or higher mask coverage. We believe that our two-stage (coarse-to-fine) generation model can be extended to very high-resolution coloring applications by improving the first-stage generation results.

Acknowledgments This study was funded by the scientific and technological project in Henan Province in 2022 (Grant No. 222102210187), the Key Research Project for Higher Education Institutions in Henan Province (Grant No. 24A520018).

References

Antoni Buades, Bartomeu Coll, and J-M Morel. 2005. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 60–65. Ieee.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer.

Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. 2020. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7760–7768.

Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Weize Quan, Ruisong Zhang, Yong Zhang, Zhifeng Li, Jue Wang, and Dong-Ming Yan. 2022. Image inpainting with local and global refinement. *IEEE Transactions on Image Processing*, 31:2405–2420.

Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. 2019. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 181–190.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Ben M Waller, Mark S Nixon, and John N Carter. 2013. Image reconstruction from local binary patterns. In *2013 International Conference on Signal-Image Technology & Internet-Based Systems*, pages 118–123. IEEE.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Haiwei Wu, Jiantao Zhou, and Yuanman Li. 2021. Deep generative model for image inpainting with local binary pattern learning and spatial attention. *IEEE Transactions on Multimedia*, 24:4016–4027.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. 2018. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European conference on computer vision (ECCV)*, pages 1–17.

Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. 2019. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1486–1494.