Bring Ancient Murals Back to Life

Xingeng Zhu(✉), Ying Yu(✉), Xiaochao Deng, and Linxia Yang

School of Information Science and Engineering, Yunnan University, Kunming 650500, China
229933369@qq.com, yuying.mail@163.com

Abstract. Digital inpainting of murals has always been a challenging problem. The damage forms in real murals are complex, such as cracks, flaking, and fading. There are many difficulties in applying deep learning technology to mural inpainting. First, data sets are often difficult to obtain. Second, the network based on supervised learning is unfit to be applied to the real multiple mural damages, which makes the network unpromotable. Third, the output of deep neural network is the combination of the unmasked area in the label image and the corresponding masked area in the generated image, so there is no change in the unmasked area. Murals often fade or change color after a hundred years or more, which leads to the lack of aesthetic feeling in the repaired images. We propose a mural inpainting model based on the translation method with three domains, including a SVD block and a dense spatial attention with mask block. Specifically, the model trains two Variational Auto-Encoders to respectively map the real mural images and the clean mural images to two deep spaces, the mapping network learns the transformation between the two deep spaces by paired data. This transformation can well extend to real mural images. Experiments show that the performance of our model is better than the comparative methods, and the visual quality is improved.

Keywords: Digital inpainting · Mural inpainting · SVD · Dense spatial attention

1 Introduction

As an artistic entity, Chinese ancient mural paintings have rich historical and scientific values. However, due to natural weathering and destruction by human factors, ancient murals show considerable signs of deterioration. Digital inpainting of damaged murals can avoid the irreversible defects of manual inpainting and improve efficiency. Therefore, the digital restoration of ancient murals has great practical significance for preserving cultural relics.

In digital inpainting solutions for murals, these methods can be divided into two categories: traditional methods and learnable methods. Traditional methods are always based on diffusion-based methods [2] or patch-based methods [1].

M. Tanveer et al. (Eds.): ICONIP 2022, LNCS 13624, pp. 231–242, 2023.
https://doi.org/10.1007/978-3-031-30108-7_20

Traditional methods are prone to matching errors, blurring, and structural disorder when applied to murals inpainting. Data-driven deep neural network models bring more possibilities. Ren et al. [9] proposed using generalized regression neural networks for the digital inpainting of Dunhuang murals. Cao et al. [3] proposed an enhanced consistency generative adversarial network, which mainly solves the inconsistency between global and local repaired images. Meanwhile, attentional mechanisms are widely used for image inpainting. Yang et al. [10] proposed dilated multi-scale channel attention to perceive image information at different scales. He et al. [5] proposed a residual attention fusion block that enhances the utilization of practical information in the broken image and reduces the interference of redundant information. The inpainting models based on supervised learning only focus on the damaged parts of murals and cannot solve the global color problem. For this reason, we propose a mural inpainting model based on the translation method with three domains [7]. In addition, embedding a SVD block and a dense spatial attention with mask block to the model. The two branches improve the ability of the model to restore murals. Specifically, the main contributions of this paper are as follows:

- For the first time, we apply weak supervised learning to the mural inpainting task, which can restore the missing parts and perfect the overall appearance of murals.
- We design a dense spatial attention with mask block embedded in the mapping net, which further enhances the network's ability to capture the long-distance mapping relationship of deep features; The SVD effectively filters the high-frequency information while retaining most of the structure and detail information, further expanding the overlap between image features.
- Experiments show that our model not only has an excellent ability to repair the cracks, spots, and scratches but also improves the visual effect.

2 Proposed Approach

In this section, we first introduce the basic network architecture of the inpainting network. Then we describe the two effective blocks, i.e., the Singular Value Decomposition (SVD) and the Dense Spatial Attention with Mask (DSAWM).

2.1 Principle of Inpainting Network

The model formulates mural inpainting as an image transformation process. The training images consist of three parts: the set of real murals $\mathcal{R}$, the synthetic set of $\mathcal{X}$ where images suffer from artificial degradation, and the corresponding set of clean murals $\mathcal{Y}$ that comprises images without degradation. $r \in \mathcal{R}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ represent the image in the three sets. x and y are paired by data synthesizing, i.e., x is degraded from y. The artificial degradation forms include holes, blurs, scratches, and low resolution.

First, $r \in \mathcal{R}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ are mapped respectively to the three deep spaces by $E_{\mathcal{R}} : \mathcal{R} \rightarrow \mathcal{Z}_{\mathcal{R}}$, $E_{\mathcal{X}} : \mathcal{X} \rightarrow \mathcal{Z}_{\mathcal{X}}$ and $E_{\mathcal{Y}} : \mathcal{Y} \rightarrow \mathcal{Z}_{\mathcal{Y}}$. Since $r \in \mathcal{R}$ and

$x \in \mathcal{X}$ are both corrupted, we align their spatial features into a shared space by mandatory strategies. As shown in Fig. 1, let the overlap of the two deep spaces (the part between black dotted lines) as large as possible, so there is $\mathcal{Z}_{\mathcal{R}} \approx \mathcal{Z}_{\mathcal{X}}$.

Then we learn the transformation from the spatial features of corrupted murals, $\mathcal{Z}_{\mathcal{X}}$, to the spatial features of clean murals, $\mathcal{Z}_{\mathcal{Y}}$, through the mapping $T_{\mathcal{Z}} : \mathcal{Z}_{\mathcal{X}} \rightarrow \mathcal{Z}_{\mathcal{Y}}$, where $\mathcal{Z}_{\mathcal{Y}}$ can be further reversed to y through generator $G_{\mathcal{Y}} : \mathcal{Z}_{\mathcal{Y}} \rightarrow \mathcal{Y}$. By learning the spatial transformation. real ancient mural r can be restored by sequentially performing the mappings,

$$r_{\mathcal{R}\rightarrow y} = E_{\mathcal{R}}(r) \circ T_{\mathcal{Z}} \circ G_{\mathcal{Y}} \tag{1}$$

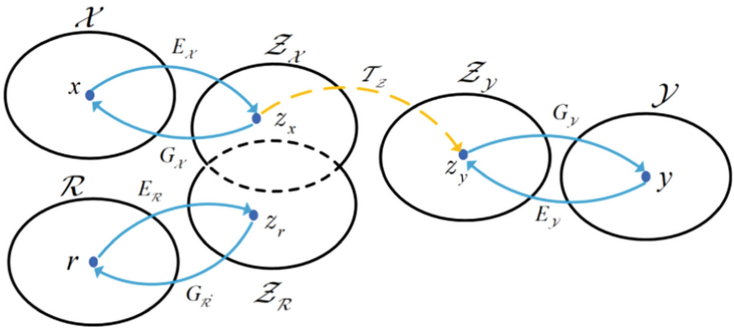


Fig. 1. Illustration of the principle of inpainting network.

We use the network shown in Fig. 2 to implement the inpainting process. The model is trained in two stages. In the first stage, two VAEs are trained to recover the input by unsupervised learning, where VAE_1 takes r and x as input, VAE_2 takes y as input. In the second stage, the mapping network is trained by fixing the weight of the VAEs trained in the first stage, the training input is x, which first enters the encoder of the VAE_1, then passes through the mapping network, and finally is decoded by $G_{\mathcal{Y}}$ of the VAE_2. The loss function of VAE_1 is defined as

$$\min_{E_{\mathcal{R},\mathcal{X}}, G_{\mathcal{R},\mathcal{X}}} \max_{D_{\mathcal{R},\mathcal{X}}} \mathcal{L}_{\text{VAE}_1}(r) + \mathcal{L}_{\text{VAE}_1}(x) + \mathcal{L}_{\text{VAE}_1,\text{GAN}}(r, x) \tag{2}$$

where KL divergence, L_1 distance loss and the least-square loss (LSGAN) [6] are included in $\mathcal{L}_{\text{VAE}_1}(r)$. $\mathcal{L}_{\text{VAE}_1}(x)$ and $\mathcal{L}_{\text{VAE}_1}(r)$ are in the same form and will not be repeated. $\mathcal{L}_{\text{VAE}_1,\text{GAN}}(r, x)$ means training another discriminator $D_{\mathcal{R},\mathcal{X}}$ that differentiates $\mathcal{Z}_{\mathcal{R}}$ and $\mathcal{Z}_{\mathcal{X}}$. The loss function of the mapping network can be expressed as

$$\mathcal{L}_{\mathcal{T}}(x, y) = \lambda_1 \mathcal{L}_{\mathcal{T},\text{L}_1} + \mathcal{L}_{\mathcal{T},\text{GAN}} + \lambda_2 \mathcal{L}_{\text{FM}} \tag{3}$$

where $\mathcal{L}_{\mathcal{T},\text{L}_1}$, $\mathcal{L}_{\mathcal{T},\text{GAN}}$ and $\mathcal{L}_{\text{FM}}$ represent L_1 distance loss, the least-square loss (LSGAN) [6] and feature matching loss, respectively.

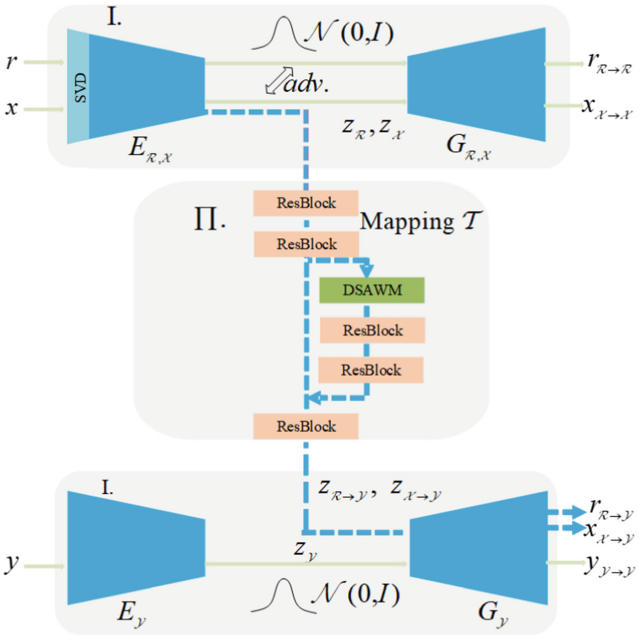


Fig. 2. Architecture of our restoration network

2.2 Singular Value Decomposition (SVD)

Visually, as shown in Fig. 1, the larger the overlap between spatial features $\mathcal{Z}_{\mathcal{R}}$ and $\mathcal{Z}_{\mathcal{X}}$(the part between black dotted lines), the better. To achieve the goal, we propose to add the SVD to the encoder of VAE_1. SVD is the generalization of eigenvalue decomposition on any matrix. Let a matrix $A \in M_{m\times n}$ with rank r, then define the SVD of the matrix as

$$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^{\mathrm{T}}_{n\times n} = U_{m\times m}\begin{pmatrix} D_{r\times r} & O \\ O & O \end{pmatrix}_{m\times n} V^{\mathrm{T}}_{n\times n} \tag{4}$$

where, $U_{m\times m} = A \times A^{\mathrm{T}}$, $V_{n\times n} = A^{\mathrm{T}} \times A$, Σ is a matrix of $m \times n$, $D_{r\times r} = \begin{pmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_r} \end{pmatrix}_{r\times r,}$ $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_{\min(m,n)} = \lambda_r > 0$ is the non-zero eigenvalue of $A^{\mathrm{T}} \times A$, so the SVD can be written as follows:

$$A_{m\times n} = U_{m\times m}\Sigma V^{\mathrm{T}}_{n\times n} = (u_1, \ldots, u_m)\begin{pmatrix}\sqrt{\lambda_1} & & \\ & \sqrt{\lambda_2} & \\ & & \ddots\end{pmatrix}\begin{pmatrix}v_1^{\mathrm{T}} \\ \vdots \\ v_n^{\mathrm{T}}\end{pmatrix} = \sqrt{\lambda_1}u_1 \ldots v_1^{\mathrm{T}} + \sqrt{\lambda_2}u_2 v_2^{\mathrm{T}} + \cdots \tag{5}$$

at this point, each eigenvector v_i in V is called the right singular vector of A, each eigenvector u_i in U is called the left singular vector of A. The singular values are ordered from largest to smallest, so the top N larger singular values and its corresponding singular vectors can approximate the matrix.

(a) (b) (c) (d)

Fig. 3. Effect of SVD: (a) Ground Truth (b) Top 10 singular value composite image, (c) Top 40 singular value composite image, (d) Top 150 singular value composite image.

Using SVD, the gap between the spatial features at high frequencies reduces to a certain extent, and the overlap between the two spatial features further expands. Figure 3 shows the effect of the SVD.

2.3 Dense Spatial Attention with Mask

For the image inpainting task, the ordinary spatial attention mechanism is not applicable because the information in the masked region propagates from the adjacent regions, which results in inaccurate attention scores after normalization. Considering this problem, we propose the Spatial Attention with Mask (SAWM), as shown in Fig. 4. The difference between SAWM and ordinary spatial attention is that the mask is added to the feature map before normalization, ensuring that the damaged areas do not affect the attention score, but the output is missing information. To this end, we propose to solve the problem by multi-level fusion of Dense-net. The proposed mechanism is called Dense Spatial Attention with Mask (DSAWM).

Figure 5 shows the DSAWM. Given the input and its mask, three convolution kernels of 1×1, 3×3, and 5×5 are used to extract multi-scale features, the input image and the multi-scale feature maps are densely connected. This process with x as input can be expressed as

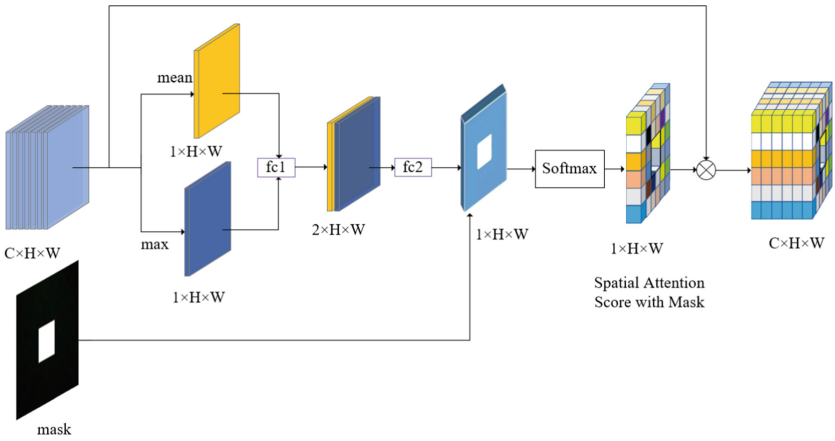


Fig. 4. Spatial Attention with Mask (SAWM).

$$y_1, y_2, y_3 = Conv_{1\times1}(x), Conv_{3\times3}(x), Conv_{5\times5}(x) \tag{6}$$

$$h_1 = \mathrm{SA}(x) \times fc_3[cat(x, y_1) \times mask] \tag{7}$$

$$h_2 = \mathrm{SAWM}(h_1) \times fc_4[cat(x, y_1, y_2)] \tag{8}$$

$$h_3 = \mathrm{SA}(h_2) \times fc_5[cat(x, y_1, y_2, y_3)] \tag{9}$$

$$y = \mathrm{SA}(h_3) \times (x) \tag{10}$$

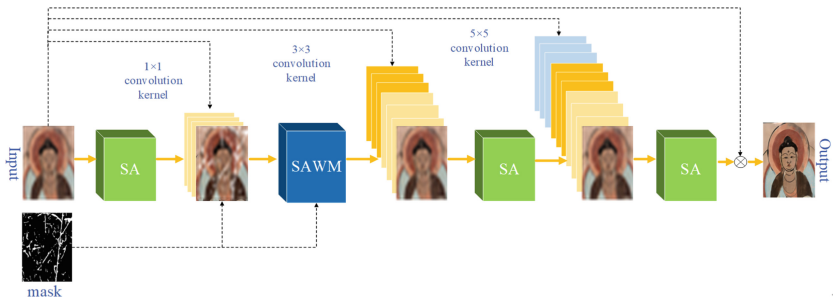


Fig. 5. Dense Spatial Attention with Mask (DSAWM).

where, cat is channel concatenating, fc_3, fc_4, and fc_5 indicate that the number of channels will be $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{4}$ of the number of input data channels by the convolution kernel 1×1 respectively. The mask is added to SAWM for calculating the attention score; SA calculates the ordinary spatial attention score. The last ordinary spatial attention score times the input image to get the final output.

In this way, the information in the unmasked region of the image is used multiple times, and the advantage of dense connectivity to preserve information is also incorporated.

3 Experiments

We select 1535 mural images to form $\mathcal{Y}$, which contains 300 modern murals. $\mathcal{R}$ includes 1632 real damaged murals. The image size is adjusted to 256×256 pixels. The learning rate is set to 0.0002, and batchsize is set to 15. We use the Pytorch framework to train and test the model. The experimental platform equipment configuration: Intel Core i7-6850K 3.60GHz CPU, NVIDIA GeForce GTX 1080Ti GPU.

3.1 Artificial Destruction Murals Repair Comparison

Table 1. Comparison of PSNR and SSIM of inpainting results.

	Criminisi		RN		DS-net		Our	
Image	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM	PSNR/dB	SSIM
1	36.43	0.9824	33.48	0.9769	38.33	0.9848	37.21	0.9885
2	27.37	0.9367	21.84	0.8941	34.16	0.9157	34.77	0.9683
3	25.48	0.8661	24.72	0.8543	28.68	0.8647	31.46	0.9364
4	26.35	0.9950	37.15	0.9726	34.84	0.9792	36.41	0.9872
5	24.46	0.8476	19.86	0.8375	24.02	0.8676	24.12	0.8699
6	23.17	0.8285	18.39	0.8267	24.51	0.88419	24.78	0.8493
7	22.02	0.8272	19.97	0.8267	23.79	0.8363	24.10	0.8614
8	23.68	0.8332	19.36	0.8394	25.03	0.8418	25.28	0.8558

Eight murals are selected for inpainting experiments with artificially added masks, random masks are added to the ones numbered 1–4, and the last four murals are masked with center holes. We compare our model to three state-of-the-art models in experiments: Criminisi [4], RN [11] and DS-net [8]. RN and DS-net use the same way of training VAE_2 in our model. We used peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) to evaluate the image inpainting quality. Table 1 shows the results of the quantitative analysis.

3.2 Experiment in Inpainting Real Damaged Murals

To further verify the effectiveness of our model in restoring the real murals, eight damaged murals are selected for the inpainting experiments. The experimental results are shown in Fig. 6.

Fig. 6. Qualitative comparisons on real damaged murals: (a) Damaged murals, (b) input, (c) Criminisi, (d) RN, (e) DS-net, (f) Ours.

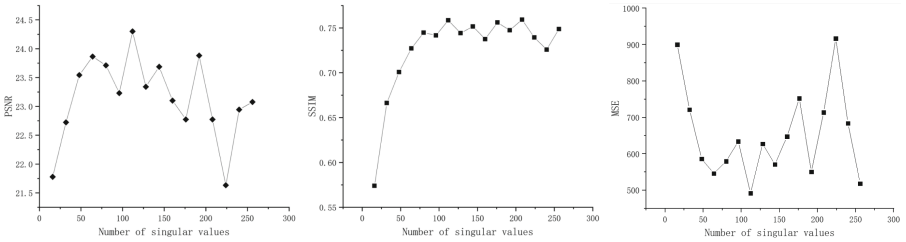
The 1st image has long cracks in the background. It can be observed that the Criminisi algorithm suffers from disruption of the line and fuzzy extension of texture information. RN repair result has blurry area. DS-net has almost no repair ability for the masked area. Note that our repair result is better, which is visually indistinguishable from the inpainting marks. Moreover, the structural consistency is better than the other 3 comparative algorithms. The 2nd and 3rd mural images have some facial cracks. It can be seen that the 3 comparative algorithms have evident artifacts and texture blur in their repair results. Our model achieves better coordination in line fitting, and the contrast of the repaired image is enhanced. For the 4th mural image, some mildew areas can be observed in the lower part. Both the Criminisi and DS-net have varying degrees of structural disorder, and the RN has large fuzzy block effect in its repair result. All these 3 comparative algorithms have obvious repair traces. By comparison, our method can produce better continuity and more reasonable restoration for the deteriorated areas. In the 5th and 6th mural images, there exist lots of scratches. Criminisi, DS-net and our model have visible repair effect, whereas RN performs poorly. Although the Criminisi and DS-net yield noticeable restoration for the scratches, they produced some inpainting errors and residual artifacts. For instance, in the 5th mural image, the restoration result of the Criminisi has a matching error near the corner of the bodhisattva's eye. The 7th mural image has some color falling-off in the hair bun. The Criminisi fails to repair these color falling-offs in this test. RN and DS-net cannot restore color-consistent areas with the surrounding region. By comparison, the repaired areas of our proposed model are visually satisfactory and semantically reasonable. The 8th mural image looks somewhat blurry and has a color inconsistent area in the bodhisattva's face. In this test, The Criminisi can restore this color inconsistency, whereas RN and DS-net cannot produce satisfactory results. Our proposed model can restore the mural image successfully. Moreover, the overall appearance of our result is considerably clearer than the other 3 approaches.

3.3 Ablation Study

To verify the utility of the SVD and the DSAWM, the original translation method with three domains [7] is used as the baseline model, "B" denotes it, "S" denotes the model after adding the SVD to the baseline model, "D" denotes the model after adding the DSAWM to the baseline model, "F" denotes our model. Figure 7 shows the variation of the quantitative index with respect to the different number of singularities in "F".

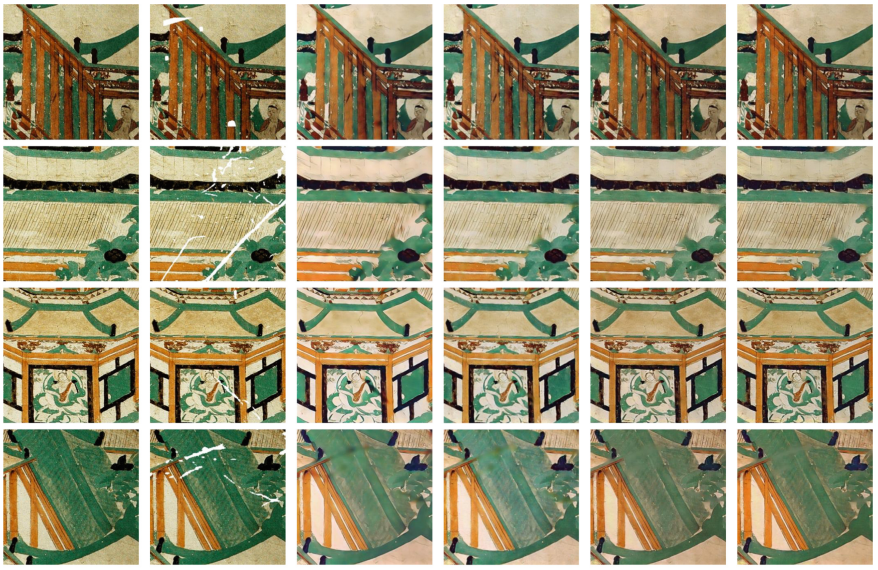
In Fig. 7(a) and Fig. 7(c), the optimal values of PSNR and mean square error (MSE) are received when the number of synthesized singularities is 112; the suboptimal value of SSIM is obtained at the same number 112 in Fig. 7(b). Therefore, the selected number of singularities in all experiments is 112.

The purpose of adding the SVD is to expand the overlap of $\mathcal{Z}_{\mathcal{R}}$ and $\mathcal{Z}_{\mathcal{X}}$, so that the network has better generalization performance. The images of building murals not included in the training dataset are selected for visual analysis. Figure 8 and Fig. 9 show qualitative and quantitative evaluation of the ablation



(a) (b) (c)

Fig. 7. (a): PSNR varies with different numbers of singular values. (b): SSIM varies with different numbers of singular values. (c): MSE varies with different numbers of singular values.



(a) (b) (c) (d) (e) (f)

Fig. 8. Qualitative comparisons of the inpainting networks. (a) Ground Truth. (b) input. (c) the inpainting results of "B". (d) the inpainting results of "S". (e) the inpainting results of "D". (f) the inpainting results of "F".

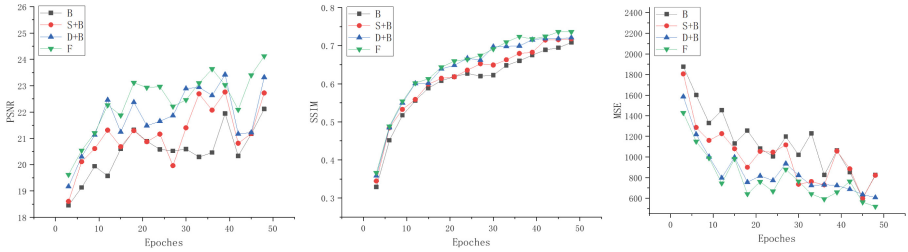


Fig. 9. (a): Compare the test PSNR values of the inpainting networks. (b): Compare the test SSIM values of the inpainting networks. (c): Compare the test MSE values of the inpainting networks.

experiment results, respectively. The comparison between Fig. 8(c) and Fig. 8(d) shows that the restoration result is relatively clearer after adding the SVD. From the comparison between Fig. 8(c) and Fig. 8(e), we can see that the output result of adding the DSAWM module has a better ability to capture colour information, structure information and detail information. The output of "F" combines the above two advantages. From Fig. 9(a) and Fig. 9(c), it can be seen that adding the SVD and the DSAWM can improve the PSNR value and reduce the MSE value, which indicates that the above blocks improve the restoration quality at the pixel level and perception level.

4 Conclusion

This paper proposed a novel DSAWM block and added the SVD to the inpainting model. The DSAWM enhances the ability of the network to capture the long-distance mapping relationships of deep spatial features, the SVD makes the images decomposed and then reorganized, effectively filtering high-frequency information while retaining most of the structure and detail information, further expanding the overlap of image features. Experiments show that our model based on weak supervised learning not only has good restoration ability for cracks, spots and scratches but also has some improvement in visual effects. However, there are still problems of blurred restoration in large damaged areas, which will be studied from the perspectives of obtaining high-quality data sets, reasonable image enhancement algorithm and optimized network.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Grant No. 62166048, Grant No. 61263048) and by the Applied Basic Research Project of Yunnan Province (Grant No. 2018FB102).

References

1. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
2. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th annual Conference on Computer Graphics and Interactive Techniques, pp. 417–424 (2000)
3. Cao, J., Zhang, Z., Zhao, A., Cui, H., Zhang, Q.: Application of enhanced consistent generative adversarial network in mural repairing. J. Comput.-Aided Des. Comput. Graph. **32**(8), 1315–1323
4. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. **13**(9), 1200–1212 (2004)
5. He, P., Yu, Y., Xu, C., Yang, H.: RAIDU-Net: image inpainting via residual attention fusion and gated information distillation. In: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (eds.) ICONIP 2021. LNCS, vol. 13108, pp. 141–151. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-92185-9_12
6. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802 (2017)
7. Wan, Z., et al.: Bringing old photos back to life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2747–2757 (2020)
8. Wang, N., Zhang, Y., Zhang, L.: Dynamic selection network for image inpainting. IEEE Trans. Image Process. **30**, 1784–1798 (2021)
9. Xiaokang, R., Peilin, C.: Murals inpainting based on generalized regression neural network. Comput. Eng. Sci. **39**(10), 1884–1889 (2017)
10. Yang, H., Yu, Y.: Res2U-Net: image inpainting via multi-scale backbone and channel attention. In: Yang, H., Pasupa, K., Leung, A.C.-S., Kwok, J.T., Chan, J.H., King, I. (eds.) ICONIP 2020. LNCS, vol. 12532, pp. 498–508. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63830-6_42
11. Yu, T., et al.: Region normalization for image inpainting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12733–12740 (2020)