Enhanced Inpainting Model Revitalizes Historical Paintings with Vision Transformer

1st Xinran Duan
College of Artificial Intelligence
Beijing Normal University
Beijing, China
202011081033@mail.bnu.edu.cn

2nd Chaoyong Jiang
College of Artificial Intelligence
Beijing Normal University
Beijing, China
jcydwx@mail.bnu.edu.cn

3rd Yachun Fan*
College of Artificial Intelligence
Beijing Normal University
Beijing, China
fanyachun@bnu.edu.cn

***Abstract*—This paper presents a deep learning architecture for restoring ancient paintings, which have immense historical and artistic value as they vividly record history from diverse perspectives. Due to the passage of time, many historical works have suffered damage, which requires time-consuming manual restoration by skilled professionals. Our proposed method utilizes a sophisticated edge detection model to extract structure information from the paintings, including texture, painting style, and structure, which are applied for restoration. The effectiveness of the proposed method was validated by training and testing on various ancient painting datasets. This work has significant value in that it can expedite and enhance the accuracy of the restoration process without compromising the original artistic style and intent, thereby better preserving and transmitting our historical culture. We believe that the contribution of this work is meaningful for VR cultural heritage conservation and presentation.**

***Index Terms*—image inpainting, ancient paintings, generative adversarial network, transformer**

I. Introduction

In latest years, virtual reality (VR) technology has become more and more prevalent [1] and has been applied to all aspects of life, such as education [2], [3], medical rehabilitation [4], roaming [5], etc. Among them, VR cultural heritage conservation, i.e., digitizing cultural heritage into the virtual world to show people is a hot research topic in recent years. Based on deep learning, ancient painting restoration can recover missing or blurred murals, surface textures of cultural relics, etc., which is significant for VR digital cultural heritage conservation and display.

Ancient paintings are an essential part of human cultural heritage. They record the history, culture, and art of ancient times and are closely related to the society, culture, and politics of their era. They not only reflect the humanistic environment and social life of their time but also serve as a record of the aesthetic concept and artistic style of their era. By restoring ancient paintings, we can gain a better understanding of the techniques, styles, and characteristics of ancient art. This understanding can help us restore historical truth more accurately and aid in comprehending the development of ancient culture and art.

However, these works of art may suffer damage in various ways during long periods of preservation. For instance, wood or paper may be eroded by moths, leading to damage to the painting surface, impurities such as grease or liquid may adhere to the painting surface, leading to stains and extensive masking damage. Furthermore, the painting surface may show wear, scratches, or scuff marks. Restoring ancient paintings requires filling in the missing areas in damaged paintings with reasonable content while preserving the original style and characteristics without altering the artistic style and intent.

Traditional algorithms for restoring ancient paintings [6]–[9] reconstruct by heuristically searching for similar pixel blocks. While this approach can solve the local restoration problem for small areas, it is insufficient for large-area restoration. This is because large-area restoration requires extracting more texture information and the global structure. In addition, traditional methods have poor reconstruction effects on details, especially on images with complex edge information. With the development of artificial intelligence technology, deep learning methods are increasingly being applied in image restoration.

Although convolutional neural networks (CNNs) [10] and generative adversarial networks (GANs) [11] have proven effective for restoration tasks such as large image defects, there are still some limitations in their direct use to deal with ancient painting restoration.

- The current methods for restoring images do not account for the styles unique to ancient paintings. The diverse artistic styles of paintings from different eras add to the difficulty of restoration work.
- CNN-based methods currently available have limited local perceptual fields and inductive bias. This makes it challenging for these methods to learn semantically consistent textures and gain an overall understanding of the image, resulting in unnatural imperfections and loss of details in the image restoration process.
- As the masked area in ancient paintings does not have any location information, current methods have a tendency to repeat meaningless restoration in large, irregularly masked areas.
- The existing GAN-based methods are slow to train and difficult to control the quality of the generated images, which cannot be adapted to the precise task of ancient

*Corresponding author

painting restoration.

To address the limitations of existing methods, we propose a novel model based on the improved incremental transformer structure [12]. This model repairs the image structure through vision transformer, thereby restoring the image. To improve the model for restoration of paintings with complex structure information, the context-aware tracking strategy (CATS) [13], a pixel-level edge detection model, was combined with a three-stage image restoration model to effectively tackle the task of restoring ancient paintings.

II. Related Work

A. Edge Detection

As a fundamental computer vision task of locating the boundaries of perceptually salient objects in natural images, edge detection has a long history [14] and plays an essential role in solving various problems such as image restoration [15], [16], image segmentation [17], virtual reality occlusion [18], [19], etc. In the early stages, many models [20]–[23] used low-level features of images for edge detection. These methods have excellent performance and require less computing power. Many algorithms can generally obtain clear edge maps, such as the Canny algorithm [20] which has been used until now. However, it is necessary to solve the problem of high-frequency texture suppression, so deep learning methods [24]–[28] are introduced. These edge detection methods learn multi-level edge weighting to obtain the final edge map. Deep learning methods significantly improve the performance of edge detection, as hierarchical deep features generated by large receptive fields can robustly suppress false positives in textured regions [25]–[27]. However, in the case of localization ambiguity, these methods need to be improved with morphological non-maximum suppression [24], [26], [27], [29], [30]. The CATS model focuses on the clarity of the edge map. To obtain a sharper edge map, the CATS model separates the blend of features obtained by convolution.

B. Auxiliary Information for Image Inpainting

Auxiliary information is relatively important for image restoration, such as edge information of image structure [12], [15]. Reference [15] proposed an edge generator that predicts the edge information of missing regions of an image to approximate the structure information, and demonstrated that the structure information of an image can be used as a priori information to effectively improve the restoration effect. To facilitate structure refinement, [31] proposed a multiscale restoration (MST) network with a novel encoder-decoder structure to recover the input image from sketch tensor space. ZeroRA based Incremental Transformer Structure (ZITS) [12] works similarly, but uses the transformer to obtain global information about the image to repair the lines and edges. Our work is inspired by this work. Moreover, the design of this model is more flexible in that it does not need to be retrained to handle new tasks only new structure information needs to be added to the pre-trained model, which is one of the factors that we believe that the model has good practical application.

C. Vision Transformer in Image Inpainting

Thanks to the ability to obtain global information, the transformer has achieved excellent results for many tasks in the fields of natural language processing (NLP) [32], [33] and computer vision (CV) [34], [35]. The use of transformer was first proposed in [34] to handle vision tasks, such as image recognition, and proved that the performance of convolution models can be equal to or even exceed convolution under large-scale training. Reference [35] verified the scalability and generalization of the visual transformer with the design of Masked Autoencoder (MAE). Due to the complexity of squaring, many works have been devoted to reducing its time and space complexity, for example, the axial attention mechanism [36] used in our model is one of them. We use transformer to repair the image structure and guide subsequent modules for image restoration with better results compared to CNN. Since our tasks are similar to [35], the model also inherits the advantages of the MAE unsupervised model, which is convenient for processing downstream tasks. Most of the ancient painting datasets [37], [38] are typically orders of magnitude small, un-preprocessed, and unlabeled.

The use of deep learning methods to process sequential information such as ancient texts is relatively common [39], [40], but few related works have explored ancient painting restoration. Qiaole Dong et al. [12] used an augmented, attention mechanism-based architecture for image restoration tasks, using an attention mechanism (transformer architecture) to obtain global information about the image. In addition, the transformer architecture can easily introduce positional information encoding to provide positional information of the occluded regions, thus improving the effectiveness of the model in restoring images. However, their model chooses the Canny algorithm to extract image edges for image structure restoration in order to compress the computational effort, which is an unmistakable design. Nevertheless, we propose to use CATS model [13] for more accurate structure information in order to strive for excellence on ancient paintings. Such an improvement would be helpful for the specific task of ancient painting restoration. Our model shows promising potential, and we believe it will pave the way for future research in this area.

III. Method

To accomplish the task of restoration caused by large defects in ancient paintings, etc., we propose to fuse the edge detection model and the three-stage image restoration model, as shown in Fig. 1.

The CATS model mitigates edge localization ambiguity with two main designs: tracking loss and a context-aware fusion (Co-Fusion) block. In addition to the weighted cross-entropy loss, tracking loss further introduces a set of boundary tracking functions to distinguish confusing pixels from edges, and a texture suppression function to handle texture regions to smooth them across the board. Under the supervision of

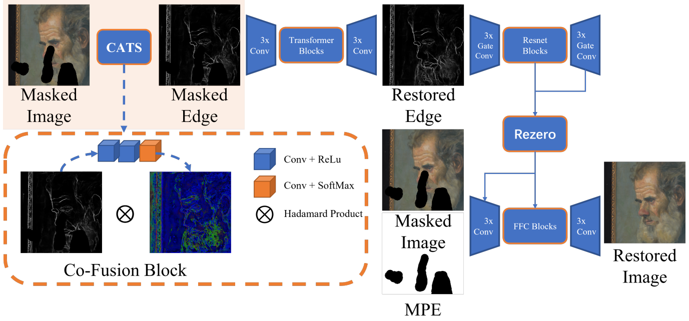


Fig. 1. The overview of our model. Our model emphasizes the use of the CATS architecture (as shown in the top left of the figure) in the preprocessing phase to obtain fine-grained structure information, which consists of three VGG-based [41] edge detectors (HED [26], RCF [27], BDCN [28]) and CoFusion blocks. the first phase of the three-stage reconstruction model uses a transformer-based model to repair the image structure, the second stage encodes the structure with Resnet, and the third stage uses fast Fourier convolution to aid in image restoration using structure information. In particular, the Resnet blocks are from the intermediate layer of [15], and the final stage of the reconstruction model uses Rezore [42] to inject the encoded structure into the third stage to aid in the reconstruction.

the tracking loss, the Co-Fusion block learns to selectively divide edges and non-edges at the pixel level, and this block replaces the weighted edge fusion part of the fusion process by considering the image features in each dimension.

Reference [12] proposed an image restoration method based on incremental transformer structure. The method embeds position encoding and masking information into the transformer encoder and decoder to achieve more accurate image restoration. The restoration model consists of three stages: transformer-based structure reconstruction, structure feature encoder, and Fourier convolutional texture restoration. The model takes as input the original image, masked regions, and edge information of the image, and down-samples it in the first stage to repair low-resolution structures, which are learned by the axial attention mechanism and encoded by the standard attention module, and the recovered structure information can be simply up-sampled to arbitrary resolutions. The second stage encodes the up-sampled sketch space features. Finally, this structure information is added to the module of the last stage in the form of weighted residual connections (Rezore) [42] to repair the texture. It has been shown that adding a learnable parameter before the residuals when only residual concatenation is needed can effectively accelerate the depth model convergence, i.e.

$$x_{i+1} = x_i + \alpha_i F_i(x_i) \qquad (1)$$

where x_i and $F_i(x_i)$ denote the input and output of layer i, respectively, and α_i is the learnable parameter.

In addition, we use masked position encoding (MPE) to represent the distance and direction from the unmasked region to the masked region. We wish to highlight several contributions to this work.

- We combine two high-performance models. This necessarily entails an arithmetic cost but is more conducive to extract the stylistic features of ancient paintings in the unmasked region. The arithmetic power of this model is acceptable from subsequent experimental results.
- To explore the practical application effect of the incremental model. To the best of our knowledge, no research explores and employs the most cutting-edge models in computer vision for practical applications in a certain domain. We believe that this work is enlightening for other researchers' work.
- Tested on SNGFaces [37] and the ancient Chinese painting dataset [38]. And, based on the practical tests we draw some valid conclusions.

A. *Edge Detection*

For this module, we use the context-aware tracing strategy (CATS) model, which obtains clear edge features by tracking loss and context-aware fusion (Co-Fusion) blocks. Obviously, only the unmasked part goes through the edge detection module.

a) The tracing loss: The tracking loss can be expressed as

$$TracingLoss(Y_P, Y_L) = L_{WCE} + \lambda_{BT} L_{BT} + \lambda_{TS} L_{TS} \qquad (2)$$

where Y_P and Y_L denote the edge prediction result and the true edge, respectively, L_{WCE} is the weighted cross entropy, L_{BT} is the boundary tracing function, L_{TS} is the texture suppression function, λ_{BT} and λ_{TS} are hyperparameters for balancing the individual elements in the tracking loss. During model training, L_{WCE} performs coarse edge learning, L_{BT} handles the refinement of edge localization by feature unmixing, and L_{TS} provides a strong overall suppression of texture regions. With L_{BT} and L_{TS}, the tracking loss handles the non-edge points collected according to the surrounding environment with target-specific suppression, achieving clear edge generation with less localization ambiguity than a single weighted cross-entropy. The expressions and necessary descriptions of the weighted cross-entropy, the boundary tracking function for feature solution blending, and the texture suppression function are given below, respectively.

Weighted cross-entropy can effectively supervise the network to learn edge maps, but it is difficult to balance the attention of edge and non-edge data even after adding the hyperparameter λ, which makes it difficult for the model to distinguish pixel regions with similar features to edges or high-frequency regions with continuous smooth changes of pixels. Therefore using the boundary tracking function L_{BT} as

$$L_{BT} = -\sum_{P\in E_L} \log\left(\frac{\sum_{i\in L_P} y_{P_i}}{\sum_{i\in R_P^{\varepsilon}\backslash L_P} y_{P_i} + \sum_{i\in L_P} y_{P_i}}\right) \qquad (3)$$

where S_E is the set of true edge points, R_P^{ε} denotes a patch containing edge fragments whose center is an edge pixel, and the set of edge pixels in R_P^{ε} is denoted as L_P. This loss function will train the model to force the predicted edge pixels in all patches to converge to the true edge pixels, while suppressing the blurred results caused by confusing pixels with similar features. For our work, it definitely provides a better representation of the structure information of the unobscured part of the image. And texture suppression function L_{TS} can be write as

$$L_{TS} = -\sum_{P \in Y_L \setminus E_P} \log(1 - \sum_{i \in R_P^t} \frac{y_{P_i}}{|R_P^t|}) \tag{4}$$

where E_P denotes the edges and their confused pixels used in L_{BT}, and R_P^t denotes the patches centered on non-edge pixels. texture suppression allows to obtain a clearer edge structure avoiding unnecessary information interfering with the subsequent structure recovery. In fact, the functions of L_{TS} and L_{BT} are complementary.

b) Context-aware fusion block: Combining edge information from different dimensions is the key to obtain accurate edge results, and previous works [26], [27], [29] commonly use weighted averaging to deal with edge details of lower dimensions and global information of higher dimensions. We use a context-aware fusion (Co-Fusion) module, which is designed based on a self-attentive mechanism that absorbs information from edges of different dimensions and avoids its limitations. In this module, a simple convolution is used to extract scores from multidimensional heatmaps as weights, and the weight map determines the contribution of each heatmap to the result. The Hadamard product of the heat map and the weight map is the final output after the activation function, as shown in the bottom left of Fig. 1.

B. Image Restoration

Reference [12] uses the Transformer restoration structure and incrementally adds the structure to the subsequent CNN texture restoration network. As mentioned above, for this module we develop the incremental model for image restoration.

a) Transformer-based Structure Restoration: This module repairs the masked image structure at a lower resolution, which can reduce the arithmetic overhead while making full use of the learning capability of the global information of the transformer. In addition to dimensionality reduction, the model uses a combination of axial attention and standard attention to control the overhead of substantial time complexity, as shown in Fig. 2 where relative position encoding (RPE) [43] is used to provide spatial information.

After the encoding is complete the convolution is up-sampled to the same size as the original input. We use binary cross-entropy (BCE) loss to optimize the structure repair module by predicting the complete edge E_P and the true complete edge E_L computing the loss denoted as

$$L = BCE(E_P, E_L) \tag{5}$$

To obtain clear high-resolution edge maps, up-sampling using a learning approach also effectively avoids the vignetting problem generated by interpolation.

b) Fourier CNN Texture Restoration: Reference [44] proposed a resolution-stabilized painting restoration using Fourier convolution. We use convolution to down-sample to a certain size to repair and then up-sample to the original size. The module is a self-encoder with a Fourier convolution layer at its core, which consists of two branches: local conventional convolution and global fast Fourier transform post-convolution. This model has a global perceptual field and local invariance but still requires predicted architectural features to complement the restoration task.

c) Structure Feature Encoder: For the predicted complete structure information, and similarly as before downsampling to small sizes is encoded into the feature space using full convolution (FCN). In contrast to the above process of down-sampling and up-sampling, here we use gated convolution [45] to extract relatively sparse structure features.

d) Loss function: We calculate the loss of the unmasked part using a simple minimum absolute value deviation (L1 loss), i.e.

$$L_{L1} = \overline{M} \otimes |I_L - I_P|_1 \tag{6}$$

where $\overline{M}$ identifies the pixel as belonging to the masked or unmasked region, identified by 0 and 1, respectively, $\otimes$ denotes the Hadamard product, and I_L and I_P denote the true and predicted images, respectively. The adversarial loss includes two parts: generator loss L_G and discriminator loss L_D. In the above model, the last two stages are regarded as the generator G, and we design the discriminator D based on the discriminator in [46]. the generator loss is denoted as

$$L_G = -\mathbb{E}_{I_P}[\log D(I_P)]. \tag{7}$$

The discriminator loss is expressed as

$$\begin{aligned} L_D = &-\mathbb{E}_{I_L}[\log D(I_L)] \\ &-\mathbb{E}_{I_{P,M}}[\log D(I_P) \otimes \overline{M}] \\ &-\mathbb{E}_{I_{P,M}}[(1-\log D(I_P)) \otimes (1-\overline{M})]. \end{aligned} \tag{8}$$

We also used the gradient penalty in [47], as

$$L_{GP} = \mathbb{E}_{I_L}\|\nabla_{I_L} D(I_L)\|^2. \tag{9}$$

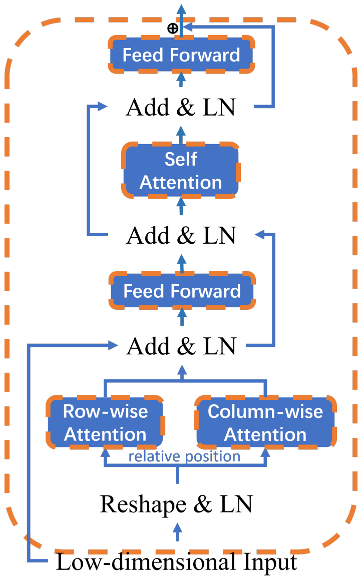


Fig. 2. The transformer block in transformer-based structure reconstruction.

TABLE I
HYPERPARAMETER SETTING

λ_{L_1}	λ_{ADV}	λ_{FM}	λ_{HRF}
10	10	100	30

Therefore, the antagonistic loss is expressed as

$$L_{ADV} = L_D + L_G + \lambda_{GP} L_{GP} \tag{10}$$

where λ_{GP} takes 10^{-3} and we use the feature matching loss L_{FM} mentioned in [48]. The formula is written as

$$L_{FM} = \mathbb{E}_{I_L} \| D(I_L) - D(I_P) \|_1. \tag{11}$$

In addition to this, we also used high receptive field perception loss (HRF loss) [44] which is written as

$$L_{HRF} = \mathbb{E}([\phi_{HRF}(I_L) - \phi_{HRF}(I_P)]^2) \tag{12}$$

where ϕ_{HRF} is a pre-trained network that evaluates the distance between the features extracted from the predicted image and the target image, which we implement using dilated convolution. The final loss function of the above model is

$$Loss = \lambda_{L_1} L_{L_1} + \lambda_{ADV} L_{ADV} + \lambda_{FM} L_{FM} + \lambda_{HRF} L_{HRF} \tag{13}$$

where the hyperparameters are set as Table I.

IV. EXPERIMENT

A. Dataset

For the CATS model, there is no dedicated ancient painting dataset. We use pre-trained model parameters, and this model is only used to assist the restoration model to improve performance in our work, so we will not repeat it here. For the restoration model, it is not enough to just use the existing ancient painting dataset, so we perform incremental supplementary training based on the pre-trained model.

The CATS model was pre-trained using BSDS500 [49], consisting of 200 training images, 100 validation images and 200 test images. Each image in this dataset is annotated by several annotators.

The restoration model was pre-trained using the Places2 [50] dataset and the indoor dataset. Approximately 1800k images from Places2 and 20055 indoor images were used as training sets and tested on 256×256 and 512×512 sizes, respectively. For our task, we also trained and tested the model on the following datasets separately. And thanks to the excellent feature that all parts of ZITS are self-models, we do not need a dedicated image restoration dataset.

SNGFaces [37], a face image dataset, the images are derived from high-resolution scanned images of oil paintings. The dataset contains 621 high-quality PNG images with a resolution of 2048×2048, and 644 high-quality PNG images with a resolution of 1024×1024.

Another dataset is the ancient Chinese painting dataset organized by [38], which contains 2936 ancient bird painting images, 2720 flower ancient painting images and 2610 landscape ancient painting images, a total of 8266 ancient painting images.

TABLE II
FID LOSS OF THE ORIGINAL MODEL AND OUR MODEL

	With Canny	With CATS
Chinese ancient paintings(Bird)	29.67	**28.99**
Chinese ancient paintings(Flower)	54.95	**51.62**
Chinese ancient paintings(Landscape)	52.11	**50.44**
SNGFaces(1024 × 1024)	**101.45**	102.65
Average FID	52.56	**51.09**

TABLE III
FID LOSS OF MST MODEL AND OUR MODEL ON SNGFACES

	MST	Ours
FID	113.45	**102.65**

B. Implementation Details

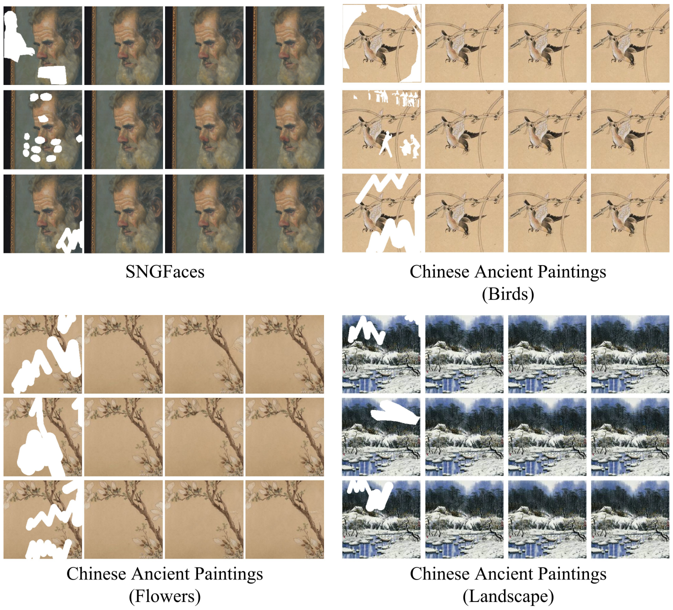
a) Training Design: As a complement to the pre-trained model, it is sufficient to emulate the training settings in [12] to suit our downstream tasks. Specifically, the entire model is implemented in PyTorch, where the module in the first stage uses the Adam optimizer to train 80 epochs and 100 epochs on the SNGFaces and ancient Chinese painting datasets, respectively. The modules from the final two stages use the Adam optimizer to train 60 epochs and 80 epochs on the two datasets. Our number of iterations is much smaller than the pre-training, due to the smaller training volume required by the pre-trained model to handle the downstream task fine-tuning, our computing power limitation, and the increased computing power demand after the addition of the CATS model.

b) Covering Design: We also use the same masking setup in [12], which includes 1000 irregular masks with masking rates from 10% to 50%. This design also facilitates our comparison with the original ZITS model in terms of training and testing.

c) Restoration Results: As can be seen from Fig. 3, the results of the original model and our model restoration are generally similar, with most of the test data showing better details under our model and some data recovered more completely on the original model. The specific data in Table II also confirms such a view.

As can be seen from Table II, our model outperforms the original model in the Chinese ancient painting dataset with more data, while it performs poorly in the SNGFaces dataset relative to the original model. We analyze the main reasons in the main characteristics.

d) More Experiments: To evaluate the effectiveness of the reconstruction based on transformer in this task, we compare it with MST [31]. The result is presented in Fig. 4 and Table III, where the structure information repaired through transformer-based outperforms the CNN-based approach for the auxiliary effect on the reconstruction. A similar ablation experiment was conducted in [12], and while our result aligns with theirs, the validation on a different dataset enhances the generalization of this conclusion.

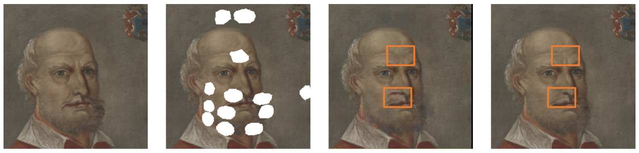

SNGFaces

Chinese Ancient Paintings (Birds)

Chinese Ancient Paintings (Flowers)

Chinese Ancient Paintings (Landscape)

Fig. 3. Some of the results. The four sets of images are some examples of test results for each dataset, from left to right, Masked Images, Original Images, Predicted Image with Canny, and Predicted Images with CATS. Predicted Image with Canny means Predicted Image with CATS indicates our model using CATS model.


Original Image

Masked Image

Result from MST

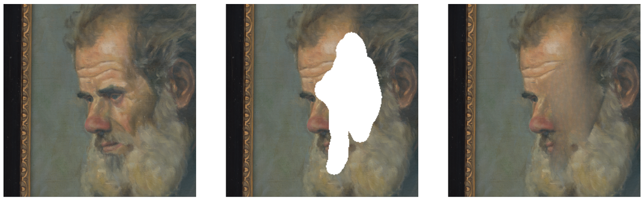
Our Result

Fig. 4. An example comparing the results of MST and our model, the differences in the reconstruction have been ticked off with boxes.

C. Main Features

The ZITS model exhibits a notable characteristic whereby if the entities in an image are fully covered, the resulting inpainted image will fail to restore those entities, as demonstrated in Fig. 5.

One critical factor is the inability of the first stage in the ZITS model to acquire structur information from the obscured regions. In essence, this is akin to extracting a portion of an image and attempting to replace it with generated pixels, a complex task that necessitates more advanced algorithms and larger datasets.


Original Image

Masked Image

Predicted Image

Fig. 5. As the left eye portion of the original image is obscured, our model cannot accurately restore this region. The image is 1024 × 1024 in size.

As previously mentioned, our model's overall performance is slightly superior to the original ZITS model in testing, and the effect may be slightly inferior in the test of some data. We attribute this to two primary factors: computing power and data set limitations, insufficient number of training iterations, even fine-tuning based on the pre-trained model requires more training; compared with the original model, our optimization mainly focuses on the process of structure restoration, while the image structure of ancient paintings is very diverse, most of which have unusual features, which increases the difficulty of our task.

In conclusion, the results of the restoration need to be interpreted while considering the differences in the datasets. Our analysis indicates that the differences in the SNGFaces dataset and the Chinese ancient painting dataset, such as the small amount of data, lack of fixed features between pictures, and larger image size, can make it challenging for the improved model components to learn effective features. However, the pre-trained parameters of the original model match its components, enabling it to achieve relatively efficient performance with only a small amount of data and diverse features.

V. Conclusion

In this paper, we propose a model combining CATS and ZITS techniques for a practical antique painting restoration task. Our proposed model employs the CATS model to detect edge structure information, replacing the Canny algorithm in the original model. We tested it on the oil painting dataset SNGFaces and the Chinese ancient painting dataset and yielded valid conclusions, which are presented below.

Compared to the original model, our proposed model outperforms on larger data sets, showcasing the effectiveness of replacing the simple Canny algorithm. However, it should be noted that our model also requires increased training volume and computing power. Furthermore, in our experiments, we observed some performance fluctuations as the number of fine-tuning steps increased, although the final performance ultimately surpassed that of the original model without bottleneck issues. While our model has demonstrated significant improvements, it is important to acknowledge its limitations regarding the original model design and computational requirements. Therefore, we recommend researchers with appropriate specialized needs to use different models depending on the data.

We obtained the performance improvement by improving only a part of ZITS and fine-tuning it with specialized data. We believe the work is valuable for the conservation and presentation of digital cultural heritage such as VR murals, VR artifact surface textures, and VR museums. We hope this perspective will inspire future work.

References

[1] Tianren Luo, Zhenxuan He, Chenyang Cai, Teng Han, Zhigeng Pan, and Feng Tian. Exploring sensory conflict effect due to upright redirection while using vr in reclining lying positions. *UIST 2022 - Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 10 2022.

[2] Zhigeng Pan, Tianren Luo, Mingmin Zhang, Ning Cai, Yongheng Li, Jinda Miao, Zheng Li, Zhipeng Pan, Yuze Shen, and Jijian Lu. Magicchem: a mr system based on needs theory for chemical experiments. *Virtual Reality*, 26:279–294, 3 2022.

[3] Jijian Lu, Tianren Luo, Mingmin Zhang, Yuze Shen, Peng Zhao, Ning Cai, Xiaozhe Yang, Zhigeng Pan, and Max Stephens. Examining the impact of vr and mr on future teachers' creativity performance and influencing factors by scene expansion in instruction designs. *Virtual Reality*, 26:1615–1636, 12 2022.

[4] Tianren Luo, Ning Cai, Zheng Li, Zhigeng Pan, and Qingshu Yuan. Vr-dlr: A serious game of somatosensory driving applied to limb rehabilitation training. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12523 LNCS:51–64, 2020.

[5] Tianren Luo, Kezheng Chen, Mengjun Liu, Kang Sun, Jili Xu, and Zhigeng Pan. Design and implementation of interactive vr campus roaming system. *Proceedings - 8th International Conference on Virtual Reality and Visualization, ICVRV 2018*, pages 122–123, 7 2018.

[6] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[7] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. *Proceedings of the IEEE International Conference on Computer Vision*, 1:305–312, 2003.

[8] A. Criminisi, P. Pérez, and K. Toyama. Object removal by exemplar-based inpainting. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2, 2003.

[9] James Hays and Alexei A. Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (TOG)*, 26, 7 2007.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63:139–144, 6 2014.

[12] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:11348–11358, 3 2022.

[13] Linxi Huan, Nan Xue, Xianwei Zheng, Wei He, Jianya Gong, and Gui Song Xia. Unmixing convolutional features for crisp edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:6602–6609, 11 2020.

[14] 1937-Roberts Lawrence G. Machine perception of three-dimensional solids. 1963.

[15] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, pages 3265–3274, 10 2019.

[16] Jie Yang, Zhiquan Qi, and Yong Shi. Learning to incorporate structure knowledge for image inpainting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:12605–12612, 4 2020.

[17] Hongzhi Wang and John Oliensis. Generalizing edge detection to contour detection for image segmentation. *Computer Vision and Image Understanding*, 114:731–744, 7 2010.

[18] Tianren Luo, Zehao Liu, Zhigeng Pan, and Mingmin Zhang. A virtual-real occlusion method based on gpu acceleration for mr. *26th IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2019 - Proceedings*, pages 1068–1069, 3 2019.

[19] Tianren Luo, Mingmin Zhang, Zhigeng Pan, Zheng Li, Ning Cai, Jinda Miao, Youbin Chen, and Mingxi Xu. Dream-experiment: A mr user interface with natural multi-channel interaction for virtual experiments. *IEEE Transactions on Visualization and Computer Graphics*, 26:3524–3534, 12 2020.

[20] JOHN CANNY. A computational approach to edge detection. *Readings in Computer Vision*, pages 184–203, 1 1987.

[21] J. Kittler. On the accuracy of the sobel edge detector. *Image and Vision Computing*, 1:37–42, 2 1983.

[22] David R. Martin, Charless C. Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:530–549, 5 2004.

[23] Ren Xiaofeng and Liefeng Bo. Discriminatively trained sparse code gradients for contour detection. *Advances in Neural Information Processing Systems*, 25, 2012.

[24] Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Zhijiang Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015:3982–3991, 10 2015.

[25] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015:4380–4389, 12 2014.

[26] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. *International Journal of Computer Vision*, 125:3–18, 4 2015.

[27] Yun Liu, Ming Ming Cheng, Xiaowei Hu, Jia Wang Bian, Le Zhang, Xiang Bai, and Jinhui Tang. Richer convolutional features for edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1939–1946, 8 2019.

[28] Jianzhong He, Shiliang Zhang, Ming Yang, Yanhu Shan, and Tiejun Huang. Bi-directional cascade network for perceptual edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:100–113, 2 2019.

[29] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. 11 2016.

[30] Yaroslav Ganin and Victor Lempitsky. N4-fields: Neural network nearest neighbor fields for image transforms. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9004:536–551, 2015.

[31] Chenjie Cao and Yanwei Fu. Learning a sketch tensor space for image inpainting of man-made scenes. *Proceedings of the IEEE International Conference on Computer Vision*, pages 14489–14498, 3 2021.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-December:5999–6009, 6 2017.

[33] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, 10 2018.

[34] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 10 2020.

[35] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:15979–15988, 11 2021.

[36] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. 12 2019.

[37] Github - kondela/sngfaces-dataset: Dataset containing high quality images of oil portrait paintings made on canvas.

[38] Tingting Qiao, Weijing Zhang, Miao Zhang, Zixuan Ma, and Duanqing Xu. Ancient painting to natural image: A new solution for painting processing. *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, pages 521–530, 3 2019.

[39] Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. Restoring and attributing ancient texts using deep neural networks. *Nature 2022 603:7900*, 603:280–283, 3 2022.

[40] Hafeez Anwar, Saeed Anwar, Sebastian Zambanini, Fatih Porikli, and Campus Pakistan. Deep ancient roman republican coin classification via feature fusion and attention. 2020.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 9 2014.

[42] Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. *37th Conference on Uncertainty in Artificial Intelligence, UAI 2021*, pages 1352–1361, 3 2020.

[43] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[44] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, pages 3172–3182, 9 2021.

[45] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-form image inpainting with gated convolution. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:4470–4479, 6 2018.

[46] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:5967–5976, 11 2016.

[47] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *Advances in Neural Information Processing Systems*, 2017-December:5768–5778, 3 2017.

[48] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. 2018.

[49] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:898–916, 2011.

[50] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1452–1464, 6 2018.