IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 13, 2020, accepted September 29, 2020, date of publication October 7, 2020, date of current version October 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029084

User-Guided Chinese Painting Completion–A Generative Adversarial Network Approach

JIETING XUE, JINGTAO GUO, AND YI LIU

Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiao Tong University, Beijing 100044, China

Corresponding author: Yi Liu (yiliu@bjtu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (No. 61300072, 31771475).

ABSTRACT Image completion models based on deep neural networks have been a research hot spot in computer vision. However, most of the previous methods focus on natural images, such as faces and landscapes. In this paper, we propose a novel image completion model for a special set of artificial ancient Chinese paintings to address this limitation. Specifically, we integrate three complements: the Wasserstein Generative Adversarial Networks (WGAN), Perceptual loss, and Mean Squared Error (MSE) to train the model robustly. We propose a unique generator which can not only pay more attention to complete the details of ancient Chinese paintings but also can provide the synthesized lines to help artists to analyze paintings conveniently. Additionally, we also allow a user to supply a structure hint to guide our model to complete Chinese paintings according to his/her preference. Extensive experiments firmly demonstrate the effectiveness of our approach to complete ancient Chinese paintings and remove abnormal color blocks from them.

INDEX TERMS Deep learning, Generative adversarial network, Image completion.

I. INTRODUCTION

Chinese painting is an essential core of Chinese culture. Compared with European paintings, Chinese paintings pay more attention to profound connotations, although they are not gorgeous and multicolored. Over thousands of years, there are many world-renowned ancient Chinese paintings, such as the Along the River During the Qingming Festival (Fig. 1a), Spring Morning in the Han Palace (Fig. 1d), etc. Unfortunately, many paintings have been damaged during the long history. Traditional physical repair methods may cause secondary damages, and it is difficult for us to fix it correctly. Although existing image processing software, such as Photoshop, may be helpful in this perspective, these methods cannot fix missing contents of paintings intelligently as artists, and the labor costs are usually expensive. Thereby, it is necessary to develop an image completion approach to fix these paintings based on a deep understanding of semantics. With the development of smart convolution neural networks, we can automatically restore the missing part of the paintings. On the one hand, this new approach reduces the repair cost significantly. On the other hand, it also maintains enough flexibility and applicability to allow users to fix the painting following his/her talent.

The most related work of this research is image completion. Its purpose is to restore the lost area of a damaged image in a semantically reasonable and visually realistic manner.

Traditional image completion models usually divide into two categories. One is based on texture synthesis [1]. Although these methods can achieve excellent results in detail, capturing the global structure of an image is difficult for them. The other is based on external databases [2]. Methods of this type assume that areas surrounded by similar contents could be copied and pasted into the missing region. However, there is no guarantee that an exact match can be found in the database. If not, a serious error may occur in completed images, making the final results unsatisfactory. Besides, this kind of method is often time-consuming and requires extensive computing resources. In short, traditional image completion methods cannot understand the semantics of an image, which makes the rationality and fidelity of the completed results often fall below expectations.

Recently, deep neural networks have been proven to be capable of understanding abstract semantics of images [3]. As a result, image completion models developed on deep neural networks have gradually become a standard. For

The associate editor coordinating the review of this manuscript and approving it for publication was Qiangqiang Yuan.

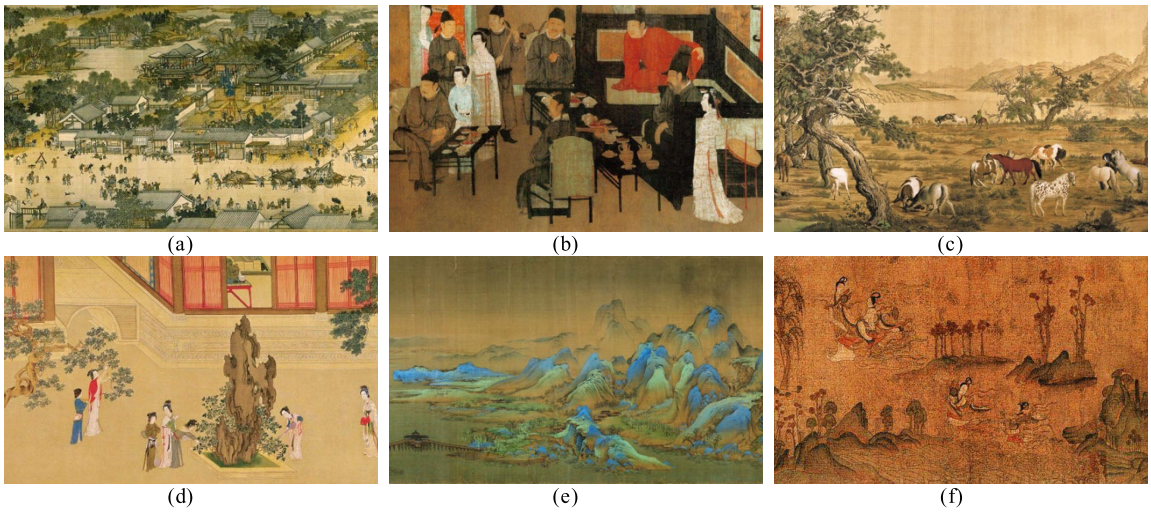


FIGURE 1. (a) Part of the Along the River During the Qingming Festival (original size: 159010*4339); (b) Part of the Night Revels of Han Xizai (original size: 84481*3472) (c) Part of the One Hundred Horses (original size: 11161*91403) (d) Part of the Spring Morning in the Han Palace (original size: 76307*3307) (e) Part of the A Thousand Li of Rivers and Mountains (original size: 152089*6083) (f) Part of the Rhapsody on Goddess of Luo (original size: 76307*3307).

example, Yu *et al.* [4] utilize the semantics which learned from large-scale datasets to fill content into non-stationary images with an end-to-end approach. Li *et al.* [5] proposed a deep face completion model, which combines reconstruction loss, adversarial loss, and semantic parsing loss to optimize their model jointly. Like these methods, most current image completion models focus on natural images rather than artificial images, such as ancient Chinese paintings. The lines and textures of paintings are typically more complex, and they are quite different from natural images. That means previous image completion approaches may not work equally well for artificial images, especially for ancient Chinese paintings. Besides, most ancient painting completion models rarely notice the lines of paintings. Lines are an essential tool for modeling paintings and represent the unique artistic language of painters. The study of lines has been one of the research hotspots for ancient painting researchers. We believe that the application of lines is extensive, and fixing lines is as important as ancient paintings. Therefore, there is still a big gap between these models and actual demand.

To address these problems, we propose a user-guided supervised model based on Wasserstein Generative Adversarial Networks (WGAN) to complete artificial ancient Chinese paintings with random irregular holes. Similar to the traditional Generative Adversarial Networks (GAN) [6], our model also contains a generator and a discriminator network. We propose a novel generator to make the synthesized paintings realistic in content and lines. To our knowledge, it is one of the first deep learning approaches to complete artificial images with state-of-art results.

In summary, this work makes the following contributions:

- We propose a novel deep learning model based on WGAN to complete artificial ancient Chinese paintings with irregular holes. Besides, our model also allows a user to fix a painting following his/her imagination. This is one of the first approaches to complete artificial images with satisfactory results.
- We introduce a unique generator which can not only complete the details of ancient Chinese paintings realistically but also can provide the synthesized lines to help the artists to analyze paintings conveniently.
- Extensive experiments demonstrate that our approach also can be used in removing abnormal color blocks from ancient Chinese paintings, which is also quite useful in many scenarios.

The rest of this paper is organized as follows. Section 2 reviews several critical works related to our research. Section 3 presents the network structure and objective functions of our model. Extensive experiments and results are documented in Section 4. Finally, we conclude this paper in Section 5.

II. RELATED WORK

In this section, we briefly review previous works on Generative Adversarial Networks and image completion methods.

A. GENERATIVE ADVERSARIAL NETWORKS (GAN)

Recently, GAN [6] models have achieved great success in various computer vision tasks, including image translation [7], [8], shadow removal [9], and realistic super-resolution [10], [11], *etc*. The *vanilla* GAN framework consists of a generator and a discriminator. The generator learns to synthesize samples as close as possible to real ones. The discriminator learns to classify real and fake samples accurately. As the discriminator continues to improve, the performance of the generator is also increasing. Thus, through this adversarial training, GAN could produce realistic perceptual samples. However, the *vanilla* GAN loss function is based on the Jensen-Shannon (JS) divergence, which may lead to a severe

gradient disappearance and unstable problems. Arjovsky and Bottou [12] employ Wasserstein distance to redefine the loss function and regularize discriminator by weight clipping to make the training process more stable. In this work, we build our model by employing a more efficient WGAN.

B. IMAGE COMPLETION

Image completion has been a hot research spot in computer vision. Patch-based methods such as [13]–[15], sample blocks from the non-hole region of an image or other images, and seamlessly paste the most relevant blocks to the missing area. These methods are limited by a fact that the best textures for implanting are often different from the sampled blocks, and the structures of images are not modeled. If such methods are used in completing ancient Chinese paintings with excessive lines, results may be highly monotonous and inconsistent. Partial differential equation diffusion approaches [16] complete images by gradually diffusing known pixels around the missing area. These approaches may introduce pseudo-structure problems easily and are only suitable for completing narrow holes, which have insufficient generality and are inappropriate in our task.

Recently, deep neural networks have made significant breakthroughs in various tasks of computer vision. Iizuka *et al.* [17] proposed a fully-convolutional image completion network with a global and a local context discriminator. The global discriminator evaluates the full completed image's coherence, and the local discriminator focuses on enhancing the local consistency between the completed region and its surroundings. Zhao *et al.* [18] proposed an end-to-end network that uses a different image to guide the model to synthesis new content for the missing regions. Portenier *et al.* [19] introduced an end-to-end image editing system based on sketches. Yu *et al.* [20] proposed the gated convolution, which provides a learnable and dynamic feature selection mechanism for the pixel of each channel of all layers.

III. METHODOLOGY

In this part, we first describe the general framework and the objective functions of our model. Then, we introduce an irregular mask construction method and the structure hint.

A. NETWORK

Most of the current image completion models [17]–[21] only focus on natural images, such as faces and landscapes, but minimal ones address artificial images. In this work, we design a novel image completion model for artificial ancient Chinese paintings based on the WGAN framework. As other methods based on GAN, our model also consists of a generator and a discriminator. The generator learns to complete an ancient Chinese painting with random irregular holes as realistically as possible. The discriminator learns to distinguish the authenticity of the image. The entire architecture is presented in Fig. 2.

1) GENERATOR

Image completion networks are usually based on a popular *UNet − like* [22] architecture such as [7], [21]. The generator based on this architecture usually hires more than five down-sample layers to extract deep feature maps. Then, it stacks multiple deconvolution layers to generate repaired results. This strategy is very time-consuming and requires a large number of computing resources. On the other hand, the details of images are seriously lost due to the multiple down-sampling layers, which can be fatal for multi-lined artificial ancient Chinese paintings. Besides, previous experiments and researches [23] demonstrate that the up-sampling network based on deconvolutions may cause a severe checkerboard-effect problem, which will greatly impact the quality of results.

Therefore, we propose a novel architecture for the specific Chinese ancient painting completion task. First, we abandon the popular multilayer down-sampling structure, but use three down-sampling layers and followed by four dilated-convolution layers to expand the model receptive field in the encoder. Dilated-convolution utilizes decentralized kernels, which allow a larger input area to be used in the calculation without using more parameters and computing power. Second, we employ the pixel-shuffle [24] with convolutions to replace the conventional deconvolution layers in the decoder. Pixel-shuffle expands the channel of the feature map through convolution and then enlarges the image through periodic shuffling. Third, we apply the skip-connection [30] between encoder and decoder to keep low-level information consistency. Skip-connection conveys shallow features into deep layers, which usually contain many low-level details. Finally, we introduce two decoders in the last up-sampling layer. The decoder of lines outputs a completed picture of lines and the decoder of content outputs a completed painting. Since the generator needs to optimize the line decoder and content decoder simultaneously, this undoubtedly strengthens the encoder's attention to the details of paintings and promotes completed results more realistic and natural in boundaries. On the other hand, it can also support artists to edit and analyze ancient paintings conveniently. Our model largely solved the problems of using GANs to complete ancient Chinese paintings. The specific network architecture is shown in Fig. 3.

2) DISCRIMINATOR

The goal of a discriminator is to distinguish whether an image is real or not. In order to complete random irregular holes correctly, we employ the Patch GAN [7], [20] model as our discriminator. Compared with the *vanilla* discriminator, Patch GAN can detect the image's authenticity in a larger receptive field, and make the training process more robust. The specific network architecture is illustrated in Fig. 4.

B. OBJECTIVE FUNCTION

Previous methods [17], [25] mostly rely on pixel reconstruction loss to complete the contents of the missing area. This

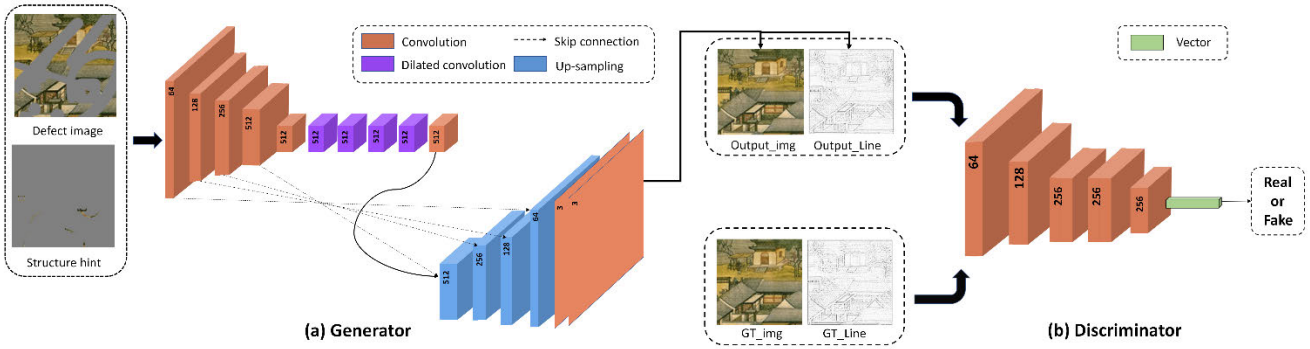


FIGURE 2. Overview of our model, consisting of two blocks: a generator G and a discriminator D. (a) G takes a structure hint and a broken image as input and generates the repaired painting and lines. We synthesize broken images by doing an inner product on raw images and masks. Then we use the depth-wise concatenation between broken images and structure hints to make them as a whole. (b) D takes in both an image and a corresponding sketch as input. It learns to distinguish between real and fake image pairs.

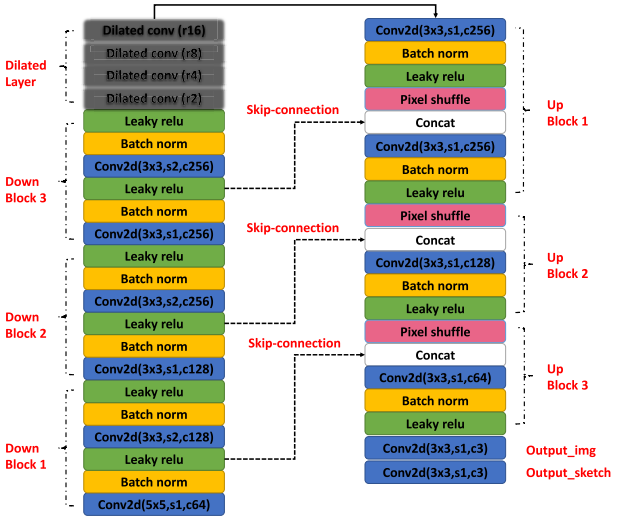


FIGURE 3. The architecture of our generator. It consists of three down-sampling blocks, four dilated-convolution layers, and three up-sampling blocks. We use skip-connection between the down-sampling blocks and the up-sampling blocks.

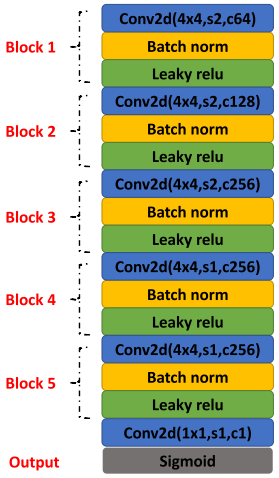


FIGURE 4. The architecture of our discriminator.

loss pays more attention to pixels consistency rather than semantic consistency. It may cause semantic errors in results. To avoid this limitation, we integrate three loss functions: adversarial loss, perceptual loss, and the Mean Squared Error (MSE) loss to train our model jointly. In the following texts, we denote G as the generator, D as the discriminator, M as the random defect mask, and S as the structure hint, which simulates the user's input. During the training phase, S is randomly sampled from the missing area of the ground truth. When we test the model, we default S does not contain any information.

We use MSE loss between completed results and corresponding ground truths to constrain G to learn low-level pixel information thoroughly. We obtain the ground truth of lines by two steps. First, we use smoothing filtering on a raw image to alleviate the effect of noise. Then we use edge filtering to extract the lines of an image. More formally, we use $\odot$ to represent the pixel multiplication. The MSE loss can be described as (1).

$$\mathcal{L}_{MSE}(G) = \|M \odot (G(x, M, S) - x)\|_2 \tag{1}$$

By minimizing the MSE loss, G generates a result with smooth textures but perceptual unsatisfying. To solve this limitation, we hire the Perceptual loss [26] in our model, which is closer to perceptual similarity. Specifically, we adopt the first four layers in the pre-trained Inception-V4 [31] network as a feature extraction model. By minimizing the L1 loss between real features and generated features, repaired results are consistent with ground truth in deep semantics. We use Θ_k to represent the feature map in the Inception-V4 network. The perceptual loss can be formulated as (2).

$$\mathcal{L}_p(G) = \sum_q \|\Theta_k(G(x, M, S)) - \Theta_k(x)\|_1 \tag{2}$$

In addition to the MSE loss and the Perceptual loss described so far, we also use an adversarial loss to support

results more clear. In this paper, we use the Wasserstein GAN loss as our adversarial loss function. We denominate the $D(x)$ as a probability distribution given by D. The generator tries to minimize this objective. In contrast, the discriminator tries to maximize this objective. It can be expressed as (3).

$$\min_G \min_D \mathcal{L}_{GAN}(G, D) = \mathbb{E}_{x\sim pdata(x)}[D(x)] - \mathbb{E}_{z\sim p_z(z)}[D(G(z))] \quad (3)$$

Finally, the overall loss function is defined as (4). We formulate the λ_{perc} as the weight of the perceptual loss and λ_{gan} as the weight of the adversarial loss.

$$\mathcal{L} = \mathcal{L}_{MSE} + \lambda_{perc}\mathcal{L}_p + \lambda_{gan}\mathcal{L}_{GAN} \quad (4)$$

C. IRREGULAR DEFECT MASK GENERATION

Previous studies generate defects by randomly deleting rectangular areas in an image. We believe it is insufficient in actual needs. Therefore, we propose a method that can generate a defect mask with arbitrary shapes. First, we set up a shape set A, including points, lines, rectangles, circles, and ellipses. Then we randomly sample 4 to 7 shapes from A and draw them with black color on a pure white image as large as the ground truth. It is worth noting that the thickness samples from 10 to 30 pixels randomly. Finally, the pixels in the mask will be normalized to 0-1. We get defect images by making the inner product between binary masks and raw images.

D. STRUCTURE HINT GENERATION

Our model allows a user to supply a structure hint to guide our model to complete Chinese paintings according to his/her preference. In the training phase, the structure hint is simply a small fraction of pixels which sampled from the raw image randomly. The sampled locations are selected by a binary mask, which is obtained by the method described in III.C. It is worth noting that we modify the thickness of the sampling range to 0-3 pixels in this binary mask, which means the structure hint we input to the model may be empty. We use this strategy to alleviate our model's excessive dependence on structural hints. Even if we do not provide any structural hints, our model can still produce realistic results. Besides, we only use the structural hint during training. When we test the model, unless a user offers his/her preference through the structure hint we default the structure hint is empty.

IV. EXPERIENCE AND RESULTS

In this section, we first describe the dataset and training details. Then we express the comparative experimental results. We also introduce the user-guided experiments and other applications of our model. In the end, we verify the effectiveness of our model through an ablation study.

A. DATASET

We select six of China's most famous ancient paintings as our basic dataset, including *Along the River During the Qingming Festival (Ming dynasty), Spring Morning in the Han Palace, The Night Revels of Han Xizai, A Thousand of Rivers and Mountains, One Hundred Horses and Rhapsody on Goddess of Luo*. We also use the *Hanging Scroll of Fairyland of Peach-Blossom Spring* as our supplementary testing dataset, which does not participate in training. We crop these ancient paintings into 256×256 image blocks with a sliding distance of 128 pixels. To avoid overlap between the training and testing set, we randomly locate certain areas on the six ancient paintings as our testing regions, and the rest are used for training. The total area of these test regions accounts for about 20% of the whole image. We finally obtained 171,987/52,122 image blocks as the training/testing sets, of which 12716 images are only for testing.

TABLE 1. The detail of the ancient painting dataset.

Dataset	Original size	Test set	Train set
Along the River During the Qingming Festival	159010*4339	6,601	26,817
One Hundred Horses	11161*91403	11,067	50,251
Spring Morning in the Han Palace	102680*5229	6,120	25,119
A Thousand Li of Rivers and Mountains	152089*6083	9,876	44,128
The Night Revels of Han Xizai	84481*3472	2,891	14,243
Rhapsody on Goddess of Luo	76307*3307	2,851	11,429
Hanging Scroll of Fairyland of Peach-Blossom Spring	24114*9064	12,716	0
Total	-	52,122	171,987

B. TRAINING ALGORITHM

We randomly select a small batch of image x from the training sets and use $x \odot mask$ to simulate the defect image, where $\odot$ represents the pixel multiplication, and the *mask* is a binary 3D array obtained by the method in III.C to identify whether each pixel in the image is lost or not. Through this processing, we get a small batch of broken image z. The z and the structure hint are input into the generator after depth-wise concatenation. Optimization of our model consists of jointly minimizing and maximizing conflicting objectives, which may cause the training process to be unstable. To avoid this problem, we train the model in three stages. First, we initialize the generator by optimizing the MSE loss and Perception loss for 20,000 iterations. Then, we fix the generator and train the discriminator with the adversarial loss for 2000 iterations. Finally, the generator and the discriminator are trained alternately, and all loss functions are used jointly to optimize the model for 100,000 iterations. Because one would not like to see any changes in the non-hole regions of a painting, we use $G_{out} \odot mask + z \odot (1-mask)$ as completion results, G_{out} is the output of the generator.

C. TRAINING DETAILS

We use Google's TensorFlow framework [27] to implement our algorithm. During the training phase, we set

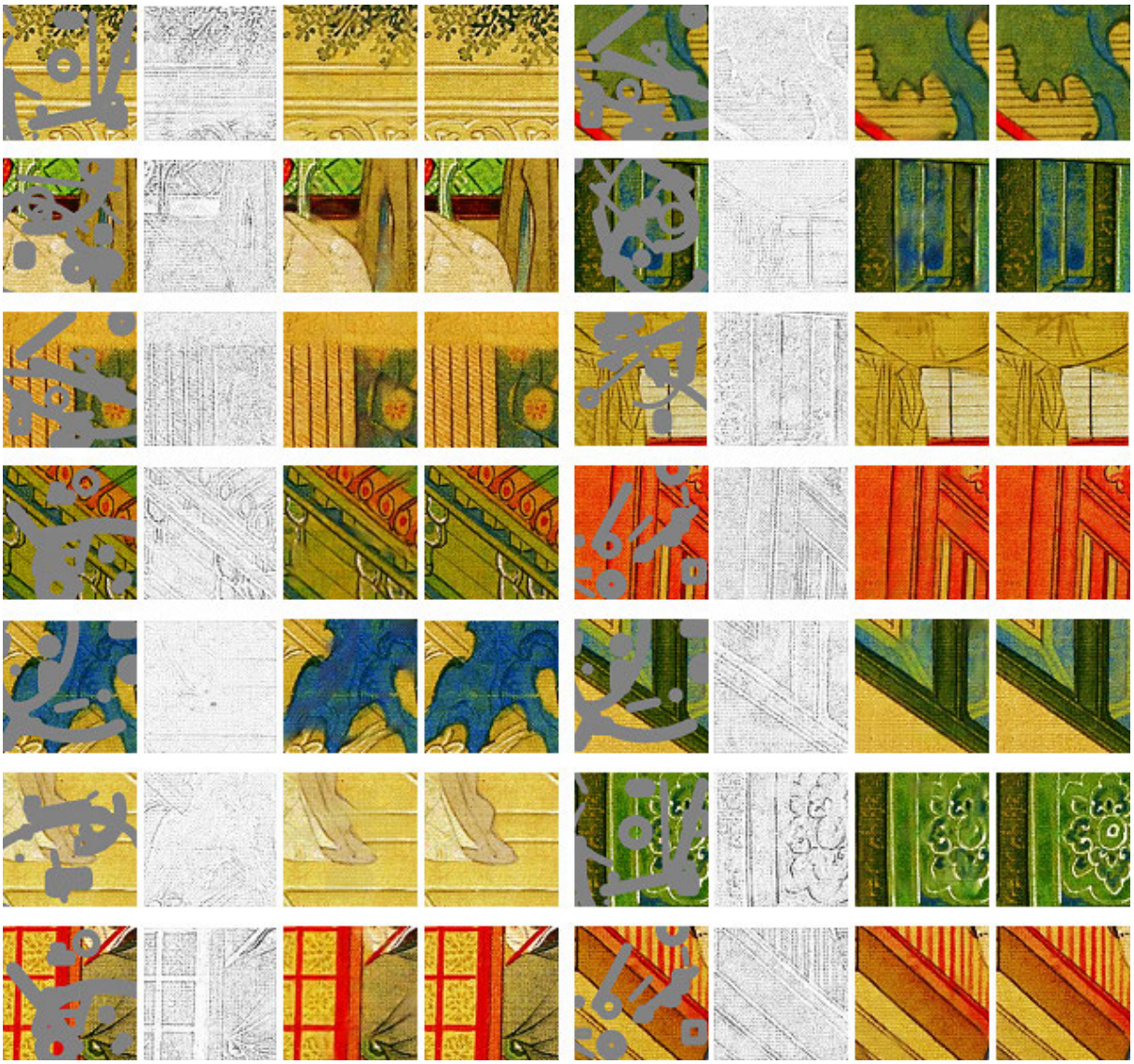


FIGURE 5. Part of image completion results on the ancient painting dataset. From left to right, four images as a group, indicating: the masked image, the completed picture of lines, the completed image, and the ground truth.

$\lambda_{gan} = 0.001$ and $\lambda_{perc} = 0.01$. We trained our model with mini-batch stochastic gradient descent with batch size 16. To speed up training, we used Adam optimizer with the learning rate of 0.0001. Additionally, except for the last layer of the generator and discriminator, we use the Leaky Rectified Linear Unit (LeakyReLU) with a hyperparameter of 0.2 [28]. The decay factor of the Batch Normalization Layer is 0.9. We set the same experimental settings in the training and testing phase.

D. RESULTS

Our model is compared with two recent image completion models, PConv [21] and GL [17]. All models have trained 100,000 iterations on the same ancient painting dataset mentioned above. For a fair comparison, except for the user-guided experiment, the structure hint does not contain any information in other experiments. In order to make the evaluation only focus on the broken area, we use $G_{out} \odot mask + x \odot (1 - mask)$ on the results of all models, where G_{out} is the output of a model and x is the raw image.

1) QUALITATIVE RESULTS

Fig. 5 shows some samples generated by our model on the test set. We find that even with large or complex semantic structures in occluded regions, our model can perfectly produce semantically consistent and visually realistic results. To further verify the effectiveness of our model, we also test it on the supplementary testing dataset, which is not involved in the training. As shown in Fig. 6, our model has excellent performance even on the data that have never been seen before. This indicates our model can learn the deep semantic information of the artificial image successfully.

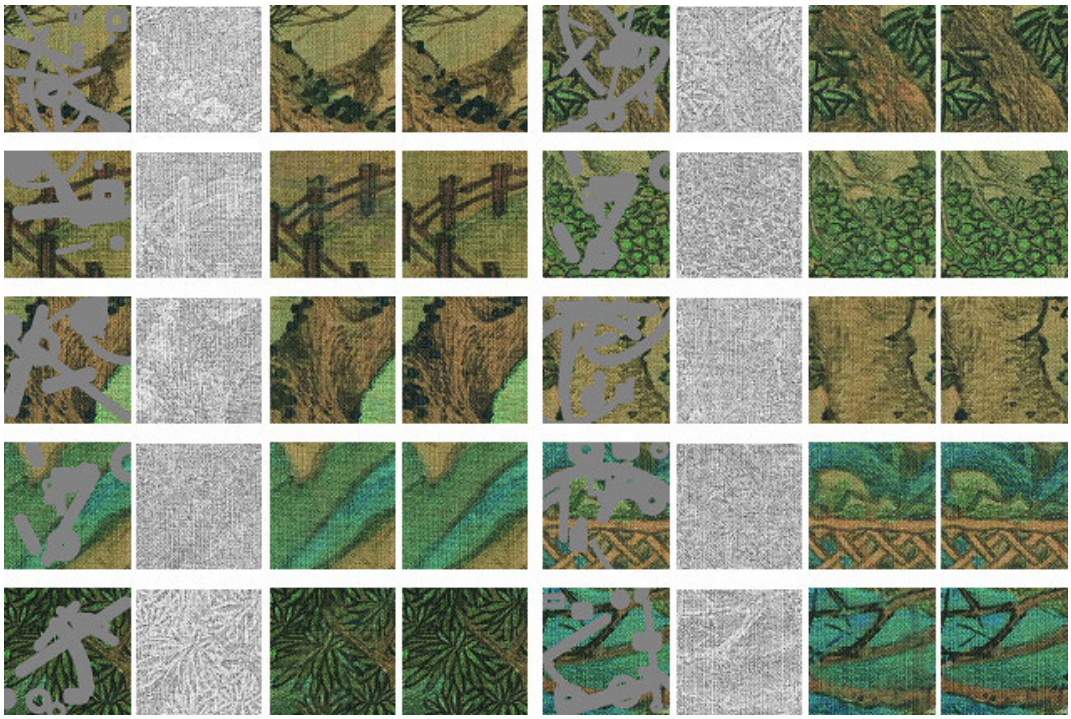


FIGURE 6. Part of image completion results on the supplementary dataset, which does not participate in training. From left to right, four images as a group, indicating: the masked image, our completed sketch, our completed image, and the ground truth.

FIGURE 7. Comparisons of the results on the ancient painting dataset. From left to right, five images as a group, indicating: the masked image, the result of [17], the result of [21], our result, and the ground truth.

Fig. 7 shows the visual effect of completing ancient Chinese paintings using three approaches. GT represents the ground truth. It is worth noting that since the local discriminator of GL only works for regular holes, which is not the case for our problem, only the global discriminator in the algorithm is used. The intact Pconv algorithm is used without any modification. We can observe from Fig. 7 that GL will create many boundary artifacts and blurred textures. Although Pconv's result is good, our approach achieved the best results in texture details and edge details, which confirms that the completion of ancient paintings cannot be easily done by conventional image completing algorithms. Besides, we

TABLE 2. Quantitative results over the test results on the Ancient painting dataset.

Methods	L1 (%)	SSIM	PSNR
GL	9.89	0.617	17.23
PConv	7.31	0.641	19.19
Ours	**7.28**	**0.643**	**19.23**

perform the same post-processing on the results of all models, which means all models are compared based on fair and consistent standards. Although the post-processing contributes to enhancing the visual effect of paintings, the repaired lost areas of other contrast methods still have the problem of blurred and twisted edge structures and serious loss of details. On the contrary, our model solves this problem well because it pays more attention to the edge information of an image. Therefore, our algorithm indeed has an edge on this specific task.

2) QUANTITATIVE RESULTS

Following previous image completion works [17], [21], we use three different metrics to evaluate the quality of the synthesized images. The first is the L1 loss, which directly measures the overall error in pixel values. The lower the L1 score, the more similar the completed result and ground truth. The second is the Structural Similarity Index (SSIM), which evaluates the similarity between the two pictures from the brightness, contrast, and structure. The higher the SSIM value, the more similar the completed result is to the ground truth. The third is Peak Signal to Noise Ratio (PSNR), which is based on the difference between the corresponding pixels. The higher the PSNR score, the more realistic the results are. Table 2 shows the comparison results. Our method has the best performance among all the three metrics.

3) USER-GUIDED RESULTS

In the previous section, we synthesized visually realistic and semantically reasonable pixels for the missing regions of ancient paintings. In order to enhance the versatility and flexibility of our model, we also allow a user to offer a structure hint based on his/her preference to guide the painting completion process. Our model is able to leverage on this guidance information to make the synthesized image have the desired structure-property. We invited five artists to provide structural hints with personal preferences, not ones sampled from the original painting. The results are shown in Fig. 8, where it is obvious that the synthesized image possesses the structure hint offered by the user. That means our model allows users to modify the structure hint to guide the completed results with personal preferences.

4) OTHER APPLICATION RESULTS

We find that our approach can also be used for removing abnormal color blocks from ancient Chinese paintings. It is useful since there are inevitable watermarks and stains on the ancient paintings, which should be removed in the digitization process. In this experiment, we still use the same data set and network architecture as the completion model, but we preprocess the image differently. We directly construct stains of random shape and color on the raw image, with a thickness of 5-10 pixels. Additionally, we do not use the structure hint and the mask because they are unnecessary in this task. As shown in Fig. 9, some random strokes and patterns are added to the input images. Our approach can remove them effectively to obtain images nearly identical to the ground truth.

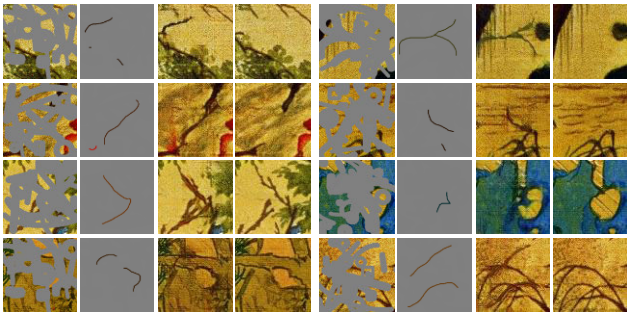


FIGURE 8. User-guided ancient painting completion results. From left to right: the masked image, the user's structure hint, our result, and the original image.

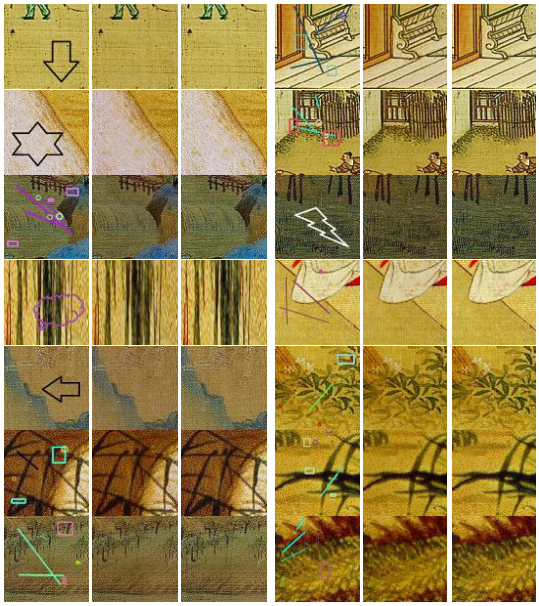


FIGURE 9. Part of the abnormal color removal results. From left to right, each of the three images is a group, indicating: the abnormal color image, our result, and the ground truth.

E. ABLATION STUDY

Here we analyze the effectiveness of two decoders. We remove the second decoder which works for completing lines and compare it with the original model. All models are trained on the same dataset with the same set of parameters. Fig. 10 visualizes part of the detailed comparison results. From sample images, we can see the results generated by

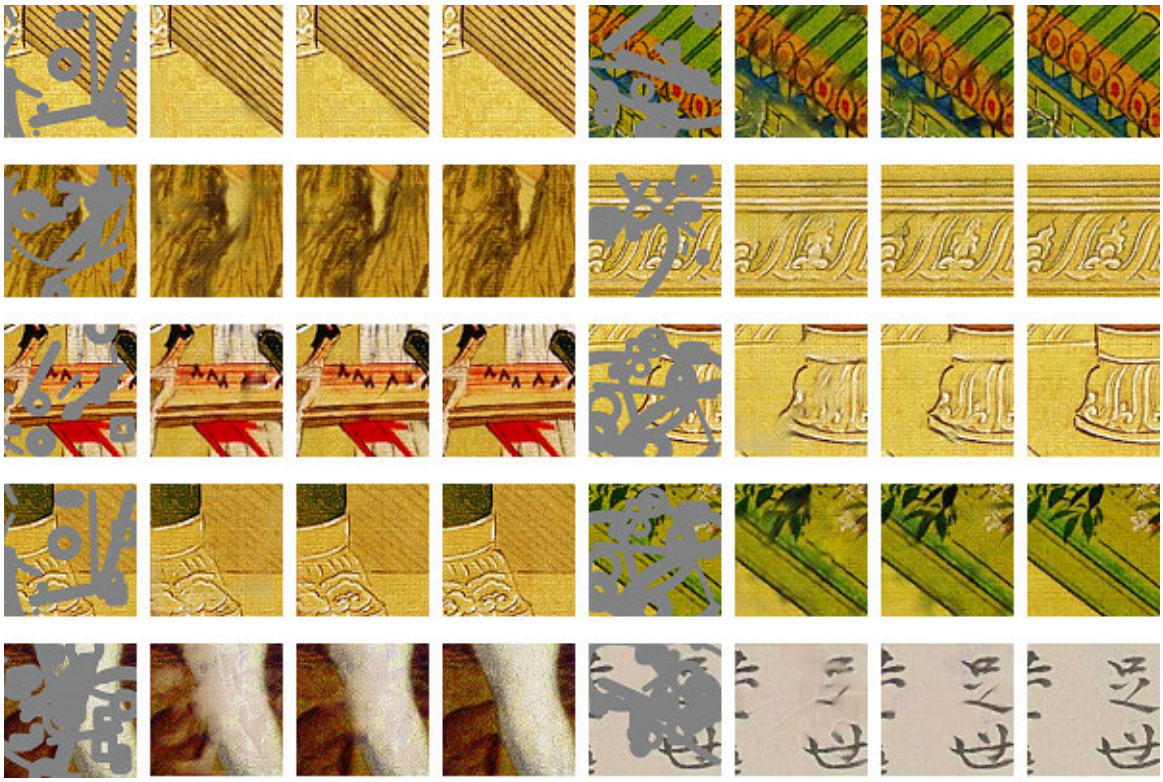


FIGURE 10. Analyzing the effects of two decoders. From left to right, four images as a group, indicating: the masked image, the completed image of the "w/o Lines" model, the completed image of the "Full" model, and the Ground Truth. "Full" is the original model described in this work. "w/o Lines" indicates the model without the line decoder.

the model with two decoders are more realistic and natural in boundaries and details. This indicates the decoder of lines has significant force on the quality of images and it encourages the generator to pay attention to details.

V. CONCLUSION

This paper proposes a novel and unique deep learning model for completing the high-resolution images of ancient Chinese paintings. Both qualitative and quantitative experiments have demonstrated the effectiveness of our approach and its clear advantage over competing methods. Additionally, it not only can be used to help a user to generate his/her desired completion results by providing a simple structure but can provide users with sufficient choices. It also works very well in removing the abnormal color blocks on the paintings. We believe this is one of the early attempts to design a deep learning synthesizer for a special type of image, which may stimulate researchers to pursue more possibilities in this direction.

ACKNOWLEDGMENT

The authors acknowledge support from the National Natural Science Foundation of China (No. 61300072, 31771475).

REFERENCES

[1] Z. Qiang, L. He, and D. Xu, "Exemplar-based pixel by pixel inpainting based on patch shift," in *Proc. CCF Chin. Conf. Comput. Vis.* Singapore: Springer, 2017, pp. 370–382.

[2] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph. (ToG)*, vol. 28, no. 3, p. 24, 2009.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[4] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[5] Y. Li, S. Liu, J. Yang, M. H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3911–3919.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[9] B. Ding, C. Long, L. Zhang, and C. Xiao, "ARGAN: Attentive recurrent generative adversarial network for shadow detection and removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10213–10222.

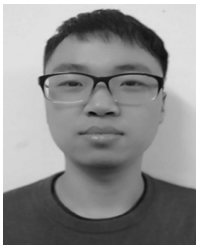
[10] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[11] N. C. Rakotonirina and A. Rasoanaivo, "ESRGAN+: Further improving enhanced super-resolution generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 63–79.

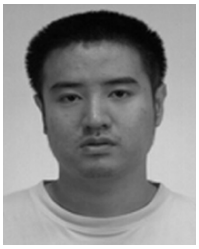
[12] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," 2017, *arXiv:1701.04862*. [Online]. Available: http://arxiv.org/abs/1701.04862

[13] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn. SIGGRAPH*, 2001, pp. 341–346.
[14] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Image completion using planar structure guidance," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–10, Jul. 2014, doi: 10.1145/2601097.2601205.
[15] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum, "Image completion with structure propagation," in *Proc. ACM SIGGRAPH Papers SIGGRAPH*, 2005, pp. 861–868.
[16] T. F. Chan and J. Shen, "Nontexture inpainting by curvature-driven diffusions," *J. Vis. Commun. Image Represent.*, vol. 12, no. 4, pp. 436–449, Dec. 2001.
[17] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017.
[18] Y. Zhao, B. Price, S. Cohen, and D. Gurari, "Guided image inpainting: Replacing an image region by pulling content from another image," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1514–1523.
[19] T. Portenier, Q. Hu, A. Szabo, S. A. Bigdeli, P. Favaro, and M. Zwicker, "Faceshop: Deep sketch-based face image editing," *ACM Trans. Graph. (TOG)*, vol. 37, no. 4, p. 99, 2018.
[20] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4471–4480.
[21] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 85–100.
[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 234–241.
[23] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, Oct. 2016.
[24] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[25] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.
[26] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.
[27] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: http://arxiv.org/abs/1603.04467
[28] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1, 2013, p. 3.
[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.

JIETING XUE received a B.E. degree in software engineering from Northwestern University in 2018. She is currently pursuing an M.S. degree with the school of computer science and technology, Beijing Jiao Tong University. Her research interests include deep learning and computer vision.



JINGTAO GUO received the B.S. degree in information management and information system and M.S. degree in computer science and technology from Beijing Jiao Tong University, in 2014 and 2017, respectively. He is currently a Ph.D. candidate at Beijing Jiao Tong University. His interests include machine learning and computer vision.



YI LIU received B.S. and Ph.D. degrees from Peking University in 2004 and 2009, respectively. His current research interests include reasoning and uncertainty modeling in system biology, machine learning, information retrieval and 3D geometric processing.