# Image Inpainting Network Based on Deep Fusion of Texture and Structure

Huilai Liang[1][a], Xichong Ling[2] and Siyu Xia[1][b]
[1]*School of Automation, Southeast University, Nanjing, China*
[2]*Department of Computing, McGill University, Montreal, Canada*

Keywords: Image Inpainting, GAN, Deep Fusion.

Abstract: With the recent development of deep learning technique, automatic image inpainting has gained wider applications in computer vision and has also become a challenging topic in image processing. Recent methods typically make use of texture features in the images to make the results more realistic. However this can lead to artifacts in the processed images, one of the reasons for this is that the structural features in the image are ignored. To address this problem, we propose an image inpainting method based on deep fusion of texture and structure. Specifically, we design a dual-pyramid encoder-decoder network for preliminary fusion of texture and structure. A layer-by-layer fusion network of texture and structure is applied to further strengthen the fusion of texture and structure feature afterwards. In order to strengthen the consistency of texture and structure, we construct a multi-gated feature merging network to achieve a more realistic inpainting effect. Experiments are conducted on the CelebA and Place2 datasets. Qualitative and quantitative comparison demonstrate that our model outperforms state-of-the-art models.

## 1 INTRODUCTION

Image inpainting, also known as image completion, aims to generate reasonable content to fill in the missing areas, and the inpainted images are expected to be both visually and semantically correct. Image inpainting is not only able to complete missing and damaged areas in the image, but also perform image editing, such as object removal, image modification, and more. In the field of image inpainting, there are two approaches in general, one is patch-based diffusion models that exploit low-level features of images, and the other is generative models of deep convolutional neural networks. The former traditional non-learning method is more effective for completing some repetitive backgrounds, but would face many problems for images with complex scenes. The latter based on neural network is capable of extracting high-level semantic features to enhance the understanding of complex scenes. For real world image inpainting tasks, the scene is often complicated, and the shape of the part to be completed can be rectangular or arbitrary. Each layer of feature map in the inpainting network is composed of invalid pixels inside the missing part and valid pixels outside. Vanilla convolution is not suitable for image inpainting because the same filter is applied to all pixels. In order to adapt to image inpainting tasks, some customized convolution methods are applied, such as the partial convolution strategy of (Liu et al., 2018), which updates the mask according to certain rules; Yu *et al.* (Yu et al., 2019) proposed a gated convolution scheme, automatically updating the masks over the network rather than by rules. In addition to improving the update of the mask, choice of loss function make a different to model performance. Many works have added gram matrix loss in image style transfer, vgg loss, etc. Edgeconnect (Nazeri et al., 2019) pointed out that the structural information of the image is also important for image inpainting, and the structural information of the first stage is used to help the inpainting of the second stage images. In fact, the texture and structure of image are related to each other. In order for the structure and texture to make use of each other to facilitate learning in image inpainting. Guo *et al.* (Guo et al., 2021a) adopt a dual-stream network of texture and structure, and propose a dual-branch discriminator. However, the actual results are still unsatisfactory for large-scale masks.

In this paper, we propose an image generation network for deep fusion of image texture and structure,

[a] https://orcid.org/0000-0001-5714-3422
[b] https://orcid.org/0000-0002-0953-6501



Liang, H., Ling, X. and Xia, S.
Image Inpainting Network Based on Deep Fusion of Texture and Structure.
DOI: 10.5220/0011718100003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 147-155
ISBN: 978-989-758-626-2; ISSN: 2184-4313

which can better fuse structural features and texture features to the full. A dual-pyramid encoder-decoder based on gated convolution is proposed to reconstruct structural and texture information. At the same time, a layer-by-layer fusion network of texture and structure is employed to further strengthen the fusion of texture and structure. In order to better maintain the consistency of texture and structure, we construct a multi-gated feature merging network to achieve a more realistic inpainting effect. We conduct experiments on the CelebA and Place2 datasets, and qualitative and quantitative comparisons demonstrate that our model outperforms state-of-the-art models.

The main contributions are as follows:

- We propose a novel encoder-decoder network based on gated convolution that fuses texture and structural features and reconstructs both features.
- A layer-by-layer fusion network of texture-structure is proposed to enhance the consistency of texture and structure.
- Proposed multi-gated feature merging network improves the restoration of details.

# 2 RELATED WORK

## 2.1 Traditional Image Inpainting

Traditional image inpainting methods mainly use similar backgrounds to fill in missing areas. The two common methods in use are, diffusion-based and patch-based. Diffusion-based methods (Bertalmio et al., 2000; Ballester et al., 2001; Esedoglu and Shen, 2002; Shen and Chan, 2002; Chan and Shen, 2001) mainly diffuse the valid pixels at the boundary of the region into the interior of the region. Note that images with high frequencies would produce defective inpainting results. The patch-based methods (Darabi et al., 2012; Xu and Sun, 2010; Barnes et al., 2009; Criminisi et al., 2004) select the most similar patch from the known area to fill the damaged area and utilize the long-distance information. A drawback is that it is computationally expensive to calculate the similarity between the area to be filled and the available. Barnes *et al.* (Barnes et al., 2009) use the fast nearest neighbor algorithm to match based on continuity, which reduces the similarity calculation expense and improves efficiency. This traditional method is based on the fact that the area to be filled can find the same or similar areas in the background, so it works well on images with high repetition, but it is difficult to perfectly fill complex scenes.

## 2.2 Image Inpainting Based on Deep Learning

The deep learning methods can use not only the shallow feature information of the image, but also the deep semantic information of the image, which promises a strong feature learning ability. Pathak *et al.* (Pathak et al., 2016) builds an encoder-decoder generative adversarial network based on U-Net. Iizuka *et al.* (Iizuka et al., 2017) propose a global and local discriminator network to improve global and local consistency. Yu *et al.* (Yu et al., 2018) propose a coarse-fine two-stage inpainting network. To apply the network to an arbitrary mask, Liu *et al.* (Liu et al., 2018) propose partial convolution to update the mask through certain rules. Yu *et al.* (Yu et al., 2019) use a neural network to automatically update masks. Zhao *et al.* (Zhao et al., 2020) use an unsupervised method for image inpainting, adding KL divergence loss and conditional constraint loss. When the missing range of the image is large, the inpainting result tends to have a large range of artifacts, Li *et al.* (Li et al., 2020) propose a progressive inpainting strategy to alleviate this effect. The internal-external learning method (Wang et al., 2021) is applied to image inpainting, which learns semantic from the outside through training on large datasets, while making full use of the internal statistics of a single test image. Zeng *et al.* (Zeng et al., 2020) propose a guided upsampling method, which can upsample images inpainted at low resolutions to high resolution images, reducing memory usage and improving computational efficiency. Inspired by image generation that can arbitrarily generate images from random noise, Zhao *et al.* (Zhao et al., 2021) complete large missing images by reconciling random noise and conditions.

## 2.3 Image Inpainting with Structural Features

Edgeconnect (Nazeri et al., 2019) first predicts the edge information of the image, and then uses the edge information as a prior to improve the inpainting effect. Structureflow (Ren et al., 2019) includes two steps: structure reconstruction and texture reconstruction. Guo *et al.* (Guo et al., 2021b) propose a texture-structure two-stream network and completes image inpainting through mutual constraints of structure and texture. Peng *et al.* (Peng et al., 2021) deconstruct the structural information and texture information of the image, and perform autoregressive modeling on the structural information, through which the sampling obtains a diverse structure. However, these methods

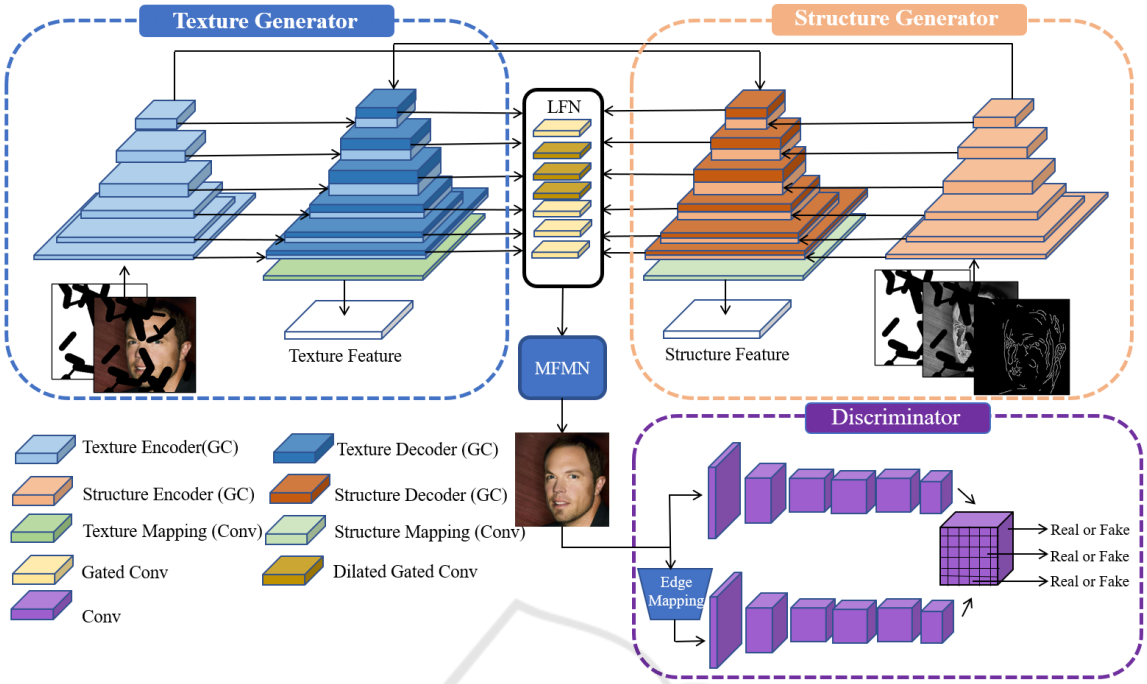


Figure 1: Illustration of our proposed method. The incomplete image and mask are input to the texture generator, and the incomplete grayscale image, edge map, and mask are input to the structure generator. The decoders of the two generators fuse the features with each other and then send them to the LFN module and the MFMN module to get the inpainting result.

do not fully consider the structural information, the utilization of the structural information is only in the reconstruction stage, and the structural information as well as texture information are not fully integrated.

Our research fully considers the structural information in each stage of image inpainting. Through the full fusion of structural information and texture information, texture and structure guide each other, and successfully make the restored image texture consistent and semantically reasonable.

## 3 METHODOLOGY

As shown in the Figure 1, the proposed method consists of four modules: texture-structure double pyramid encoder-decoder, texture-structure layer-by-layer fusion network, multi-gated feature merging network and texture-structure discriminator.

### 3.1 Texture-Structure Double Pyramid Network

Our double pyramid network is similar to U-Net (Ronneberger et al., 2015) network. As shown in the Figure 1, it is divided into two parts: texture generator and structure generator. Both generators are composed of encoder and decoder. The input of texture encoder is incomplete images and masks, meanwhile that the structure encoder is the incomplete grayscale images, edge map and mask. In the latent space, we fuse texture features and structure features at the initial stage. The texture decoder fuses the latent structural features and the features of each layer of the texture encoder to reconstruct the image texture. The structural decoder fuses the latent texture feature and the features of each layer of the structure encoder to reconstruct the structure of the image. The initial fusion of texture-structure enables the reconstruction of texture-structure to guide and promote each other and achieve a better inpainting result. Because the form of the mask may be rectangular or free-form, our structure-texture double pyramid encoder-decoder network adopts gated convolution.

The feature maps of each layer of the texture encoder are represented as $T^L, T^{L-1} \dots T^1$ from deep to shallow. For example, $T^L$ represents the feature map of the L-th layer encoder. The feature maps of each layer of the structure encoder are represented as $S^L, S^{L-1} \dots S^1$ in same order as texture encoder. Their counterparts of decoders are labeled in a similar manner. The feature maps of each layer of the texture decoder network are represented as $\psi^L, \psi^{L-1} \dots \psi^1$, the feature maps of each layer of the structure decoder network are expressed as $\phi^L, \phi^{L-1} \dots \phi^1$. For

each layer of the structure decoder, the feature maps of can be calculated as follows:

$$\begin{aligned} \phi^{L-1} &= f(S^{L-1}, T^L) \\ \phi^{L-2} &= f(S^{L-2}, \phi^{L-1}) \\ &\dots \\ \phi^1 = f(S^1, \phi^2) &= f(S^1, f(S^2, \dots f(S^{L-1}, T^L))) \end{aligned} \tag{1}$$

The feature maps for each layer of the texture decoder are calculated as follows:

$$\begin{aligned} \psi^{L-1} &= f(T^{L-1}, S^L) \\ \psi^{L-2} &= f(T^{L-2}, \psi^{L-1}) \\ &\dots \\ \psi^1 = f(T^1, \psi^2) &= f(T^1, f(T^2, \dots f(T^{L-1}, S^L))) \end{aligned} \tag{2}$$

where f represents gated deconvolution.

## 3.2 Texture-Structure Layer-by-Layer Fusion Network

In the dual-pyramid encoder-decoder network, we found that each layer of the decoder represents the complete information of the different dimensions of the image. Each layer of the texture decoder represents the complete texture information of the incomplete image, and each layer of the structure decoder represents the complete structural information of the incomplete image. Further fusion of texture-structure information can mutually promote the reconstruction of texture and structure, resulting in a better restoration effect. We propose a texture-structure layer-by-layer fusion network.

As shown in the LFN module in Figure 1, the input of the texture-structure layer-by-layer fusion network is the feature maps of each layer of the texture decoder and the structure decoder. The feature maps of each layer from top to bottom of the LFN are represented by $\tau^L, \tau^{L-1}, \dots \tau^1$. They are calculated as follows:

$$\begin{aligned} \tau^{L-1} &= f(\phi^L, \psi^L) \\ \tau^{L-2} &= g(\phi^{L-1}, \psi^{L-1}, \tau^{L-1}) \\ &\dots \\ \tau^1 &= f(\phi^2, \psi^2, \tau^2) \end{aligned} \tag{3}$$

where g denotes dilated gated convolution.

## 3.3 Multi-Gated Feature Merging Network

In order to better maintain the consistency of the structure and texture of the inpainting results, inspired by (Yu et al., 2018), we hope the model can exploit the relationship between the pixels inside and outside the missing part of the images. We enhance on the basis of (Nazeri et al., 2019), adding shortcuts (He et al., 2016) and multi-feature merging, and propose a multi-gated feature merging network.

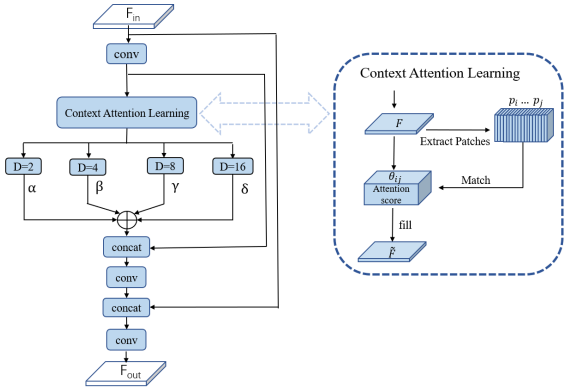


Figure 2: Illustration of the MFMN module, which improves inpainting result through attention mechanism and multi-gated feature merging.

The multi-gated feature merging network structure is shown in the Figure 2. For the contextual attention learning module, we are given a high-level feature map $F$, extracting patches from the feature map and calculate the cosine similarity between the patches: $S_{i,j} = < \frac{p_i}{\|p_i\|_2}, \frac{p_j}{\|p_j\|_2} >$, where $p_i$ and $p_j$ denote the $i^{th}$ and $j^{th}$ patches, after which softmax is applied to obtain the attention score of each patch: $\theta_{i,j} = \frac{exp(S_{i,j})}{\sum_{j=1}^{N} exp(S_{i,j})}$. Then the feature maps are reconstructed according to the attention score: $\tilde{p}_i = \sum_{j=1}^{N} p_j \times \theta_{i,j}$, where $\tilde{p}_i$ represents the $i^{th}$ patch of the reconstructed feature map $\tilde{F}$, and multi-gated feature merging is performed after the reconstructed feature map:

$$\hat{F} = [\alpha\sigma(\tilde{F}) + \beta\sigma(\tilde{F}) + \gamma\sigma(\tilde{F}) + \delta\sigma(\tilde{F})] \oplus F \tag{4}$$

$$F_{out} = Conv(\hat{F} \oplus F_{in}) \tag{5}$$

where $\sigma$ represents the dilated convolution operation, and $\alpha, \beta, \gamma, \delta$ represents the gating factor of different dilated convolutions. The feature maps of different scales are aggregated through gating factor, and the computational efficiency is improved by adding shortcut connections.

## 3.4 Loss Function

Inspired by (Guo et al., 2021a), our discriminator adopts a dual-stream SN-PatchGAN (Yu et al., 2019), which contains a texture discriminator and a structure discriminator with structure mapping, and the output of the dual-stream discriminator is synthesized as the


Figure 3: Comparison of qualitative results on CelebA with existing models. From left to right: Input, PConv (Liu et al., 2018), DeepFillv2 (Yu et al., 2019), CTSDG (Guo et al., 2021a),Crfill (Zeng et al., 2021) , LAMA (Suvorov et al., 2022), Our, Ground Truth.

discriminator of the entire generation network. To avoid the mode collapse problem, spectral normalization (Miyato et al., 2018) is used in the discriminator. The loss function of the entire network includes feature loss, reconstruction loss, VGG loss, style loss, and adversarial loss.

$I_{gt}$ represents the ground truth of the image; $E_{gt}$ represents the ground truth of the structural image; $X_{gt}$ represents the ground truth of the grayscale image and $M$ represents the mask of the incomplete image, where 1 represents valid pixels and 0 represents invalid pixels. So the input of image, structure image, grayscale image of the entire network are represented as $\tilde{I}_{gt} = I_{gt} \odot M$, $\tilde{E}_{gt} = E_{gt} \odot M$, $\tilde{X}_{gt} = X_{gt} \odot M$. The generator of the network is represented by G and consists of three parts: texture-structure double pyramid encoder-decoder network, LFN and MFMN. Denoting the discriminator by D, the entire image inpainting model can be expressed as $I_{pred}, E_{pred} = G(\tilde{I}_{gt}, \tilde{E}_{gt}, \tilde{X}_{gt}, M)$. The final output is therefore $I_{comp} = I_{pred} \odot (1 - M) + \tilde{I}_{gt}$.

Feature loss: In order to make the texture generator and structure generator focus on generating texture $F_{texture}$ and structure $F_{structure}$ respectively, texture mapping $\Phi$ and structure mapping $\Psi$ (convolution stacking) are added to the decoder, and the $l_1$ distance is used to calculate the feature difference:

$$L_{feature} = \| I_{gt} - \Phi(F_{texture}) \|_1 + \| E_{gt} - \Psi(F_{structure}) \|_1 \quad (6)$$

Reconstruction loss: Calculate the reconstruction loss of $I_{pred}$ and $I_{gt}$ with $l_1$ distance:

$$L_{rec} = \mathbb{E} \left[ \| I_{pred} - I_{gt} \|_1 \right] \quad (7)$$

VGG loss: In order to make the image have consistent semantics, we adopt VGG-19 (Simonyan and Zisserman, 2014) to obtain the semantics of $I_{pred}$ and $I_{gt}$, and use the $l_1$ distance to calculate the loss, $\varphi_i$ represents pool1-pool5 of VGG-19:

$$L_{vgg} = \mathbb{E} \left[ \sum_i \| \varphi_i (I_{pred}) - \varphi_i (I_{gt}) \|_1 \right] \quad (8)$$

Style loss: we borrow the loss function of image style transfer and calculate $l_1$ distance of the gram matrix to get the style loss:

$$L_{style} = \mathbb{E} \left[ \| Gram (I_{pred}) - Gram (I_{gt}) \|_1 \right] \quad (9)$$

Adversarial Loss: Adversarial losses are used to guarantee the authenticity of generated textures and struc-

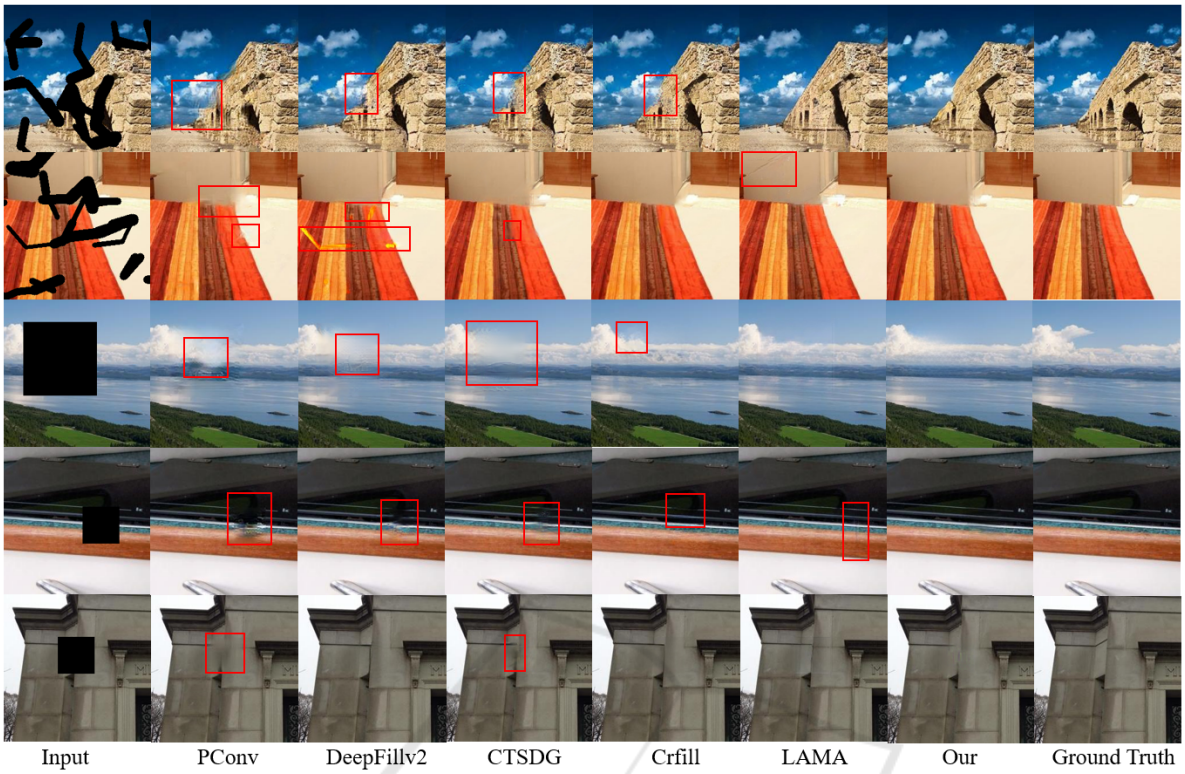


Figure 4: Comparison of qualitative results on Places2 with existing models. From left to right: Input, PConv (Liu et al., 2018), DeepFillv2 (Yu et al., 2019),CTSDG (Guo et al., 2021a),Crfill (Zeng et al., 2021), LAMA (Suvorov et al., 2022), Our, Ground Truth.

tures:

$$L_{adv} = \min_G \max_D \mathbb{E}_{I_{gt},E_{gt}}[\log D(I_{gt},E_{gt})] + \mathbb{E}_{I_{pred},E_{pred}}\left[\log\left(1 - D(I_{pred},E_{pred})\right)\right] \quad (10)$$

The loss function of the entire network is:

$$L = \lambda_1 L_{feature} + \lambda_2 L_{rec} + \lambda_3 L_{vgg} + \lambda_4 L_{style} + \lambda_5 L_{adv} \quad (11)$$

By comparing the experimental results and evaluation indicators, the hyperparameters are set to: $\lambda_1 = 10, \lambda_2 = 50, \lambda_3 = 0.3, \lambda_4 = 200, \lambda_5 = 0.5$.

# 4 EXPERIMENTS

We conducted experiments in Place2 (Zhou et al., 2017) and CelebA (Karras et al., 2017) respectively, and divided the training set, validation set and test set according to the corresponding requirements of the data set. The image size is $256 \times 256$, and the masks are divided into two types, one is rectangular mask including $128 \times 128$, $64 \times 64$, the other is free-form mask and the generation rules of free-form mask are the same as (Yu et al., 2019). The model is trained and tested by pytorch, with GPU NVIDIA 1080TI, learning rate $10^{-4}$, batch-size 4.

## 4.1 Qualitative Comparison

Figure 3 is the result of the CelebA dataset, and Figure 4 is the result of the Places2 dataset. The images used for Figure 3 and Figure 4 are not included in the training set. Bad details in the results have been marked with red boxes. Our method is compared with the representative methods of the current state-of-the-art models. For the face dataset, such as the inpainting results in the fourth row of Figure 3, PConv and DeepFillv2 have serious artifacts for the face, and the nose and eyes are obviously deformed. The effect of CTSDG is slightly better, but from the overall style, the expressions of the characters are unnatural and the structure is not very harmonious. The generated image of Crfill has inconsistent eyes and unnatural mouths. The LAMA has artifacts at the border of the ear.For the landscape dataset, such as the results in the fourth row of Figure 4, texture mosaics appear in PConv and DeepFillv2, white lines in CTSDG and Crfill appear discontinuous, and LAMA has obvious traces at the inpainting boundary. Because our net-

Table 1: Quantitative results over Places2.†Lower is better. ∗Higher is better.

| Mask Type | Rec Mask | | | | | | Free Mask | | |
|---|---|---|---|---|---|---|---|---|---|
| Mask Size | (64,64) | | | (128,128) | | | | | |
| Evaluation | $l_1^{\dagger}(\%)$ | $SSIM^*$ | $PSNR^*$ | $l_1^{\dagger}(\%)$ | $SSIM^*$ | $PSNR^*$ | $l_1^{\dagger}(\%)$ | $SSIM^*$ | $PSNR^*$ |
| DeepFillv2 (Yu et al., 2019) | 0.497 | 0.942 | 33.02 | 2.5 | 0.761 | 24.05 | 1.6 | 0.812 | 26.41 |
| PConv (Liu et al., 2018) | 0.615 | 0.937 | 30.99 | 3.2 | 0.752 | 21.86 | 1.8 | 0.787 | 26.02 |
| CTSDG (Guo et al., 2021a) | 0.473 | 0.941 | 32.99 | 2.5 | 0.762 | 23.90 | 1.5 | 0.810 | 27.09 |
| Crfill (Zeng et al., 2021) | 0.966 | 0.948 | 33.11 | 5.2 | 0.773 | 23.38 | 3.1 | 0.800 | 27.05 |
| LAMA (Suvorov et al., 2022) | 0.480 | 0.946 | 33.24 | 2.4 | 0.772 | 24.34 | 1.8 | 0.768 | 26.65 |
| **Our** | **0.451** | **0.947** | **33.49** | **2.4** | **0.773** | **24.23** | **1.3** | **0.835** | **27.98** |

Table 2: Quantitative results over CelebA.†Lower is better. ∗Higher is better.

| Mask Type | Rec Mask | | | | | | Free Mask | | |
|---|---|---|---|---|---|---|---|---|---|
| Mask Size | (64,64) | | | (128,128) | | | | | |
| Evaluation | $l_1^{\dagger}(\%)$ | $SSIM^*$ | $PSNR^*$ | $l_1^{\dagger}(\%)$ | $SSIM^*$ | $PSNR^*$ | $l_1^{\dagger}(\%)$ | $SSIM^*$ | $PSNR^*$ |
| DeepFillv2 (Yu et al., 2019) | 0.389 | 0.959 | 32.96 | 2.5 | 0.813 | 22.88 | 1.19 | 0.837 | 28.61 |
| PConv (Liu et al., 2018) | 0.419 | 0.954 | 32.72 | 2.2 | 0.806 | 24.05 | 1.18 | 0.835 | 28.67 |
| CTSDG (Guo et al., 2021a) | 0.321 | 0.960 | 33.77 | 1.7 | 0.828 | 25.94 | 0.88 | 0.867 | 30.57 |
| Crfill (Zeng et al., 2021) | 0.741 | 0.957 | 33.44 | 4.0 | 0.812 | 24.53 | 2.30 | 0.838 | 28.79 |
| LAMA (Suvorov et al., 2022) | 0.370 | 0.954 | 34.12 | 1.8 | 0.817 | 25.63 | 1.40 | 0.785 | 28.05 |
| **Our** | **0.292** | **0.963** | **35.61** | **1.5** | **0.839** | **26.89** | **0.83** | **0.873** | **31.12** |

work fully integrates texture and structure features, our results are much better than existing networks in terms of structure consistency and texture clarity. PConv (Liu et al., 2018) is suitable for free mask through mask hard update, and the image inpainting effect of rectangular mask is relatively poor. The same problem occurs with soft mask update of (Yu et al., 2019), but it has a better inpainting results for small rectangular masks. For large rectangular masks, it is prone to producing artifacts. CTSDG (Guo et al., 2021a) considers texture and structure at the same time and it can be applied to both rectangular mask and free-form mask. But when it comes to inpainting details, the results produced by our method is more natural and realistic.Because the structures of faces are relatively similar, our method deeply integrates the structural features of faces, our method is significantly better than Crfill and LAMA on the face dataset. To sum up, our proposed texture-structure deep fusion method significantly outperforms other methods in both detail and structure.

## 4.2 Quantitative Comparison

We use the currently popular image quality evaluation indicators, including $l_1$ error, SSIM (Structural Similarity) and PSNR (Peak Signal to Noise Ratio). We tested two types of masks, rectangular mask and free-form mask on the CelebA and Places2 datasets. The generation method of free-form mask is consistent with (Yu et al., 2019). Rectangular mask tested $128 \times 128$, $64 \times 64$ two sizes. The results of CelebA are shown in Table 2, and the results of Places2 are shown in Table 1. It can be concluded that our method outperforms the existing methods whether it is a face image or a natural image and regardless of the mask shapes.

## 4.3 Ablation Study

In order to verify the effectiveness of adding structural features, LFN module and MFMN module, we use the texture encoder-decoder and discriminator as basenet. For $basenet + structure$, we directly convolve the texture-decoder feature and the structure-decoder feature to get the inpainting result. For $basenet + MFMN$, we remove texture mapping layer of texture decoder, and directly send the texture decoder feature to MFMN to get the inpainting result. For $basenet + structure + LFN$, we remove the MFMN module directly, and add a convolution layer to the output of LFN to get result. Ablation experiments are tested on the place2 dataset. The experiment results are shown in Table 3.

# 5 CONCLUSIONS

In this paper, we use gated convolution to build a structure-texture double-pyramid encoder-decoder network on the basis of U-Net, which realizes the initial fusion of texture and structure. Our proposed layer-by-layer fusion network further fuses the two features. A multi-feature merging network further improves the consistency of texture and structure. For future work, we hope to introduce structural features in the inpainting of high-resolution images.

Table 3: Ablation study on Places2 dataset.†Lower is better. ∗Higher is better.

| Basenet | Structure | LFN | MFMN | $l_1^{\dagger}(\%)$ | $SSIM^*$ | $PSNR^*$ |
|---|---|---|---|---|---|---|
| ✓ | ✓ | | | 1.6 | 0.766 | 26.01 |
| ✓ | | | ✓ | 1.4 | 0.825 | 27.52 |
| ✓ | ✓ | ✓ | | 1.5 | 0.81 | 27.77 |
| ✓ | ✓ | ✓ | ✓ | 1.3 | 0.835 | 27.98 |

# REFERENCES


Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., and Verdera, J. (2001). Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211.

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

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

Chan, T. F. and Shen, J. (2001). Nontexture inpainting by curvature-driven diffusions. *Journal of visual communication and image representation*, 12(4):436–449.

Criminisi, A., Pérez, P., and Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212.

Darabi, S., Shechtman, E., Barnes, C., Goldman, D. B., and Sen, P. (2012). Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on graphics (TOG)*, 31(4):1–10.

Esedoglu, S. and Shen, J. (2002). Digital inpainting based on the mumford–shah–euler image model. *European Journal of Applied Mathematics*, 13(4):353–370.

Guo, X., Yang, H., and Huang, D. (2021a). Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14134–14143.

Guo, X., Yang, H., and Huang, D. (2021b). Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14134–14143.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Iizuka, S., Simo-Serra, E., and Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Li, J., Wang, N., Zhang, L., Du, B., and Tao, D. (2020). Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7760–7768.

Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., and Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Nazeri, K., Ng, E., Joseph, T., Qureshi, F., and Ebrahimi, M. (2019). Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544.

Peng, J., Liu, D., Xu, S., and Li, H. (2021). Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784.

Ren, Y., Yu, X., Zhang, R., Li, T. H., Liu, S., and Li, G. (2019). Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 181–190.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Shen, J. and Chan, T. F. (2002). Mathematical models for local nontexture inpaintings. *SIAM Journal on Applied Mathematics*, 62(3):1019–1043.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., and Lempitsky, V. (2022). Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159.

Wang, T., Ouyang, H., and Chen, Q. (2021). Image inpainting with external-internal learning and monochromic bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5120–5129.

Xu, Z. and Sun, J. (2010). Image inpainting by patch propagation using patch sparsity. *IEEE transactions on image processing*, 19(5):1153–1165.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2019). Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480.

Zeng, Y., Lin, Z., Lu, H., and Patel, V. M. (2021). Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14164–14173.

Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., and Lu, H. (2020). High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *European conference on computer vision*, pages 1–17. Springer.

Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., and Lu, D. (2020). Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5741–5750.

Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E. I., and Xu, Y. (2021). Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.