ISSN 2616-5775 Vol. 5, Issue 3: 70-77, DOI: 10.25236/AJCIS.2022.050310

# Performance Comparison of Irregular Face Inpainting via Deep Learning

**Xinyue Zuo[a], You Yang[b,*], Zixian Hao[c]**

*Institute of Computer and Information Science, Chongqing Normal University, Chongqing, China*
*[a]2020210516095@stu.cqnu.edu.cn, [b]20130958@cqnu.edu.cn, [c]2021110516064@stu.cqnu.edu.cn*
**Corresponding author*

***Abstract:*** *As a specific application of image inpainting, face inpainting is a critical content in the computer vision. It plays an important role in object removal, photo editing and other fields. Deep learning has become the mainstream approach of image inpainting. In specific applications, the corrupted area of face images is usually irregular. For the classical irregular face inpainting approaches based on deep learning, this paper divides it into convolution operator optimization methods and structural information constraint methods, the former includes PConv and GConv and the latter includes EC, PRVS, MED, CTSDG. We fist describe the basic principle of each algorithm and detail about the strengths and limitations. Then we experiment on CelebA-HQ dataset, evaluate and compare the performance quantitatively and qualitatively.*

***Keywords:*** *Image Inpainting, Deep Learning, CNN, GAN*

## 1. Introduction

Face inpainting aims to reconstruct the missing or damaged parts of images according to the known surrounding contents while maintaining the overall consistency. The task originated in ancient times, when artists restored corrupted pieces as much as possible. With the development of artificial intelligence, image inpainting is regarded as one of the important low-level tasks in computer vision areas. Among them, face inpainting has been of great significance in object removal, old photo restoration and other practical applications.

Traditional image inpainting methods can be mainly divided into two categories, i.e., diffusion-based and patch-based. Diffusion-based methods [1-2] gradually diffuse the known pixels around the holes in the corrupted image and synthesize new textures. The larger the corrupted area, the less effective information can be obtained from the center of holes. Diffusion-based methods are mainly used to fill images with less damage, such as scratches in photos. Patch-based methods [3-4] are implemented by finding the most similar patches in the uncorrupted area. For images with relatively large corrupted areas, they can produce better inpainting results, but they lack the perception of semantic information. For example, the corrupted position in the face image is the nose, and there is no corresponding similar patches in the known area, which makes it impossible to form a semantically reasonable result.

Deep learning has made major breakthroughs in image inpainting in the past few years, and CNNs and GANs are the main methods. Convolutional neural network (CNN) is a feedforward neural network that consists of several convolutional layers and pooling layers, and it has excellent performance in image feature learning and expression. GAN [5] is a generative model whose core idea is derived from the Nash equilibrium of game theory. The generator generates new samples by learning the potential distribution of real data samples, and the discriminator determines whether the input data is ground truth or generated data samples. Pathak et al. [6] first brought the idea of encoder-decoder structure and GANs into image inpainting. Experimental data shows that results were both semantic and authentic, filling the deficiency of traditional methods in semantic understanding. This method sets off a hot wave of research on image inpainting based on deep learning. Liu et al. [9] propose replacing vanilla convolutions with partial convolution layers to fill holes of any size, shape and position that caused researchers to explore continuously the inpainting technology of irregular corrupted areas in the field of face image inpainting.

This paper selects six classical irregular face inpainting algorithms based on deep learning proposed in recent years, details the basic principles and improvement strategies of each algorithm, experiments on CelebA-HQ dataset and evaluates the performance quantitatively and qualitatively. By analyzing and

ISSN 2616-5775 Vol. 5, Issue 3: 70-77, DOI: 10.25236/AJCIS.2022.050310

comparing the performance to help researchers select or design algorithms, and promote the application and evolution of irregular face inpainting algorithms.

## 2. Methods based on Deep Learning

Goodfellow et al. proposed generative adversarial network [5] in 2014, which has been widely used in the field of computer vision, such as image generation and image inpainting. GAN contains of two parts. i.e., generator and discriminator. In the training process, the generator generates new samples by learning the potential distribution of ground truth samples to deceive the discriminator. And the task of the discriminator model is to determine whether the given data is ground truth or generated samples. Pathak et al. [6] introduced GAN into the inpainting task and proposed Context Encoder network, which solved the semantic limitations that could not be broken through by traditional inpainting methods. The network structure is shown in Figure 1.

Encoder-decoder structure as the basic framework combined with the constraints of GAN like Context Encoder has become the mainstream method of image inpainting. Liuzuka et al. [7] added a global discriminator while retaining the local discriminator of the Context Encoder to make results more consistent with global semantics, and the method can fill rectangular holes at any position. Since vanilla convolution is difficult to capture the texture information far from holes, Yu et al. [8] introduced the attention mechanism into image inpainting and made full use of the known features around the image during training to improve the details of the inpainting. However, these methods are designed for regular rectangular holes. In practice applications, the area of damage in the image are often irregular. We pick the classic methods of irregular image inpainting in recent years and divide into based on convolution operator optimization and based on structural information constraint. The basic description of algorithms is given in Table 1. The methods based on convolution operator optimization include PConv [9] and GConv [10], and the methods based on structural information constraint include EC [11], PRVS [12], MED [13] and CTSDG [14].

### *2.1. Methods based on convolution operator*

Vanilla convolution feeds all pixels of images to convolution layers. The pixels of a corrupted image are divided known pixels and unknown pixels. By feeding both types of pixels into a vanilla convolutional layer, most of results suffer from blurring and artifacts, and usually rely on costly subsequent processing. Liu et al. [9] realized the inpainting of irregular corrupted images for the first time, considering the pixels in the known area and the unknown area as valid pixels and invalid pixels respectively, and proposed partial convolution layers with automatic mask update function instead of vanilla convolution, convolving only the valid pixels. At the same time, the U-Net structure for image segmentation is also introduced, replacing the vanilla convolution with partial convolution and the ReLU at the decoder with LeakyReLU. The skip connection of U-Net enables the decoder to supplement the features lost in encoder, which helps to produce refined results.

PConv [9] can fill holes of arbitrary shape, size and position in the image. However, the rule-based mask updating mechanism has certain irrationalities, such as the method treating one valid pixel and multiple valid pixels equally when updating the mask.

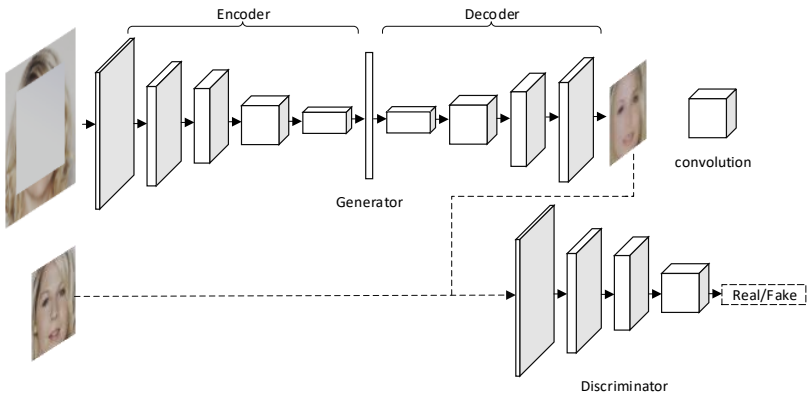


*Figure 1: Context encoder network architecture.*

ISSN 2616-5775 Vol. 5, Issue 3: 70-77, DOI: 10.25236/AJCIS.2022.050310

Yu et al. [10] proposed gated convolution based on Liu and designed a more flexible mask updating mechanism. Gated convolution assigns different weights to different channels at different spatial positions in different layers. Even in the deep network, the mask region still exists. In order to enlarge the receptive field and stabilize the training, the model consists of two stages. The first stage predicts coarse results, and the second stage predicts more refined results based on coarse results. Coarse network training uses reconstruction loss while fine network training uses the combination of reconstruction loss and adversarial loss. The fine network generator consists of two branches, the first branch consists of gated and dilated convolution, and the second branch consists of gated convolution and contextual attention. A faster and more stable spectral-normalized discriminator is also trained to improve the inpainting effect of irregular holes images. The gated convolution coarse-to-fine network model is shown in Figure 2. Gated convolution as a flexible convolution has made great progress in face image inpainting [16].

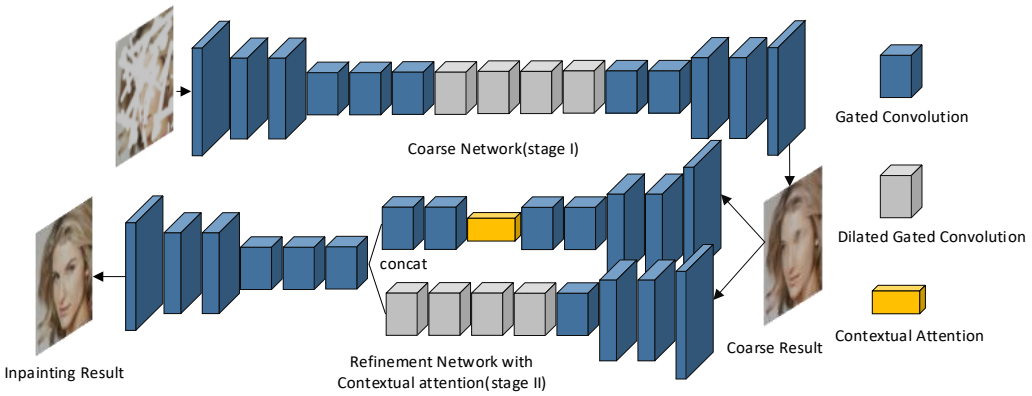


*Figure 2: Framework of the GConv generator.*

GConv [10] achieves dynamic updating of irregular masks, however, the multi-branch structure of this model contains a large number of parameters, which requires more computing resources, and has limitations such as too much smoothing in structural details.

*Table 1: Summary of basic information description.*

| Method | Year | Source | Encoder-Decoder |
|---|---|---|---|
| PConv [9] | 2018 | ECCV | U-Net |
| GConv [10] | 2019 | ICCV | Coarse-Fine network |
| EC [11] | 2019 | ICCVW | Structure-Content network |
| PRVS [12] | 2019 | ICCV | U-Net |
| MED [13] | 2020 | ECCV | U-Net |
| CTSDG [14] | 2021 | ICCV | U-Net(Two stream) |

### *2.2. Methods based on Structural Information Constraints*

Artists restore damaged artworks, usually first completing the outline of the damaged area, and then finely restoring content. Inspired by this, Nazeri et al. [11] used an edge generator to first generate the edge of corrupted images, and then use it as edge priori to guide subsequent inpainting work. The network consists of two stages. The first stage generates edge and the second stage completes corrupted image. Both stages are GAN-based network, and each stage has a generator and a discriminator. The network model is shown in Figure 3. G1, D1 are the generator and discriminator of the edge generator, and G2, D2 are the generator and discriminator in the completion network, respectively. In G1, the broken grey-scale map, edges and mask are input, and the training labels of edges are extracted by the Canny edge detection. In particular, the feature matching loss $L_{FM}$ in D1 is used to compare the activation maps of the different middle layers of the discriminator, to force G1 to produce a representation similar to ground truth images.

Using available edge information as a priori can effectively improve the quality of the results, but the two generators are not the best choice for parameter optimization and can not generate reasonable visual structure for images with large holes, resulting in structural distortion and texture blurring in the structures.

ISSN 2616-5775 Vol. 5, Issue 3: 70-77, DOI: 10.25236/AJCIS.2022.050310

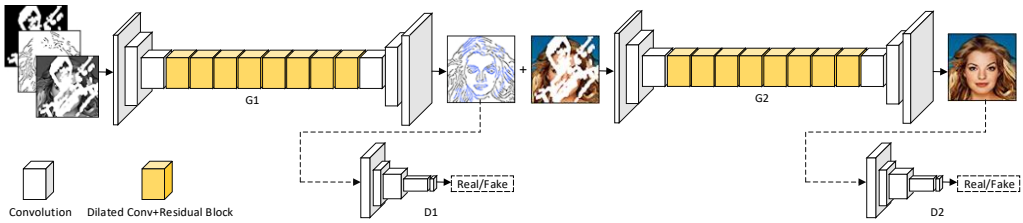


*Figure 3: Framework of EC.*

Li et al. [12] proposed a Visual Structure Reconstruction (VSR) layer to reconstruct visual structure and visual feature of the missing region. Partial edges reconstructed from the VSR are combined with the previous edges and then filled with content to gradually reduce the missing region until the inpainting is completed. VSR layers are stacked on ecoder-decoder. The VSR consists of a structure generator and a feature generator. The structure generator updates some edges of the missing region, which is used to guide the generation of new features. The structure generation process is shown in Figure 4, the image feature $X_{in}$, the edge feature Ein, previous image mask $M_{in}^{Img}$ and the mask for edge $M_{in}^{Edge}$ are input, <·> represents channel concatenation. Updating features and masks by partial convolution, updated feature $X_{pc1}$ is feed into the residual block to produce structure feature $E_{conv}$, After subtracting $M_{in}^{Edge}$ from $M_{pc1}$, multiply element-wise with $E_{conv}$ to form the newly generated structure, combined with $E_{in}$ to output the structure $E_{EG}$. The structure generator only predicts structure near known regions, and the feature generator uses $E_{EG}$ to guide the $X_{in}$ fill the content.

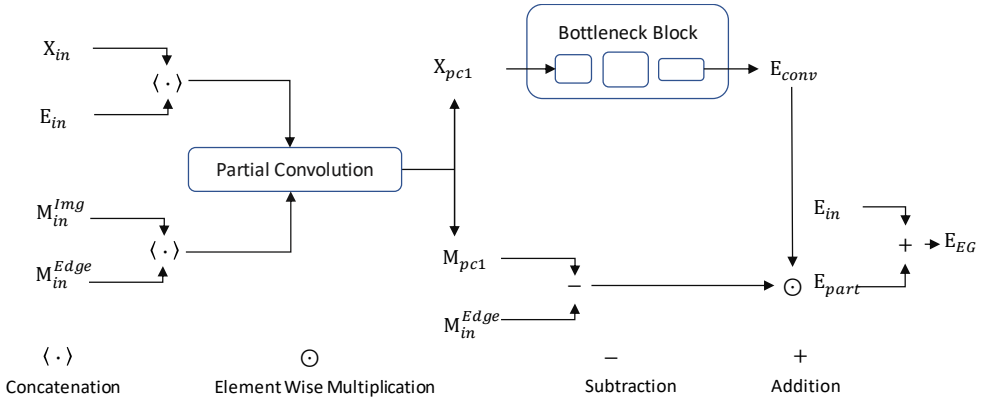


*Figure 4: The generation of structure.*

PRVS [12] has improved considerably in terms of the number of parameters and has somewhat improved the structure details of the results when inpainting images with larger holes, but it lacks consideration of texture detail.

In deep convolutional neural networks, the features extracted from the shallow layer mainly contain low-level texture features, while the features extracted from the deep layer mainly include high-level structural semantic features. Liu et al. [13] performed multi-scale filling of structure and texture features respectively, then fused the features using feature equalization, and finally connected the equalized texture and structural features to the decoder side by skip connection. The encoder contains six convolutional layers, with the features extracted from the first three layers considered as texture features to represent the details of images, and the features extracted from the last three layers considered as structural features to represent the semantics of images. There are also four residual blocks with dilated convolution between the encoder and decoder to increase the perceptual field when encoding features. The two features are used to fill holes by a multi-scale filling module, which mainly uses partial convolution to extract features at different scales using convolution kernels of different sizes, and then uses feature equalization to fuse structural and texture features.

The structure and texture of MED [13] model as well as some objective indicators have been improved. However, the generator where structure and texture are shared does not adequately consider the relationship between structure and texture. It is difficult for texture and structure to convey information to assist each other.

Guo et al. [14] proposed the idea of using structure to constrain the synthesis of texture and texture to guide the reconstruction of structure. The network uses a variant of U-Net that replaces vanilla convolutions with partial convolutions, and the interaction of structure and texture information is

ISSN 2616-5775 Vol. 5, Issue 3: 70-77, DOI: 10.25236/AJCIS.2022.050310

achieved through skip connections. The two-stream network model in the generator is shown in Figure 5, with two U-net in the generator, the first inputting the broken image, mask and outputting the structural features, and the second inputting the broken edge grayscale map, mask and outputting the texture features. The features in the texture encoder are skip-connected to the texture decoder, and the features in the structure encoder are skip-connected to the structure decoder. Therefore, the generated structure features are guided by the texture, and the generated texture features are synthesized within the constraints of the structure, making full use of the relationship between texture and structure and resulting in the results with a reasonable structure and detailed texture. The bi-directional gated feature fusion mechanism (Bi-GFF) is also designed to fuse the perceptual information between the structure and the texture of generated to enhance consistency. The above methods improvement strategies are summarized in Table 2.

*Table 2: Summary of improvement strategies.*

| Method | Strategy | Advantage | Limitation |
|---|---|---|---|
| Pconv [9] | Convolve only valid pixels | Image inpainting of irregular mask firstly | The rule-based mask update mechanism is unreasonable |
| GConv [10] | Flexible update mask | Dynamic update mask | Parameters require more computing resources |
| EC [11] | Add edge generator | Improve structural details | Visual structure is unreasonable when lager holes |
| PRVS [12] | Progressive reconstruction of visual structure | Improved structural details of larger images in corrupted areas | Lack of consideration for texture information |
| MED [13] | Structure and texture multi-scale Fill | Improve texture details | Insufficient consideration of the relationship between structure and texture |
| CTSDG [14] | Structural and texture information mutually guides fill | Improve structure and texture detail | Artifacts occur when corrupted area is large |

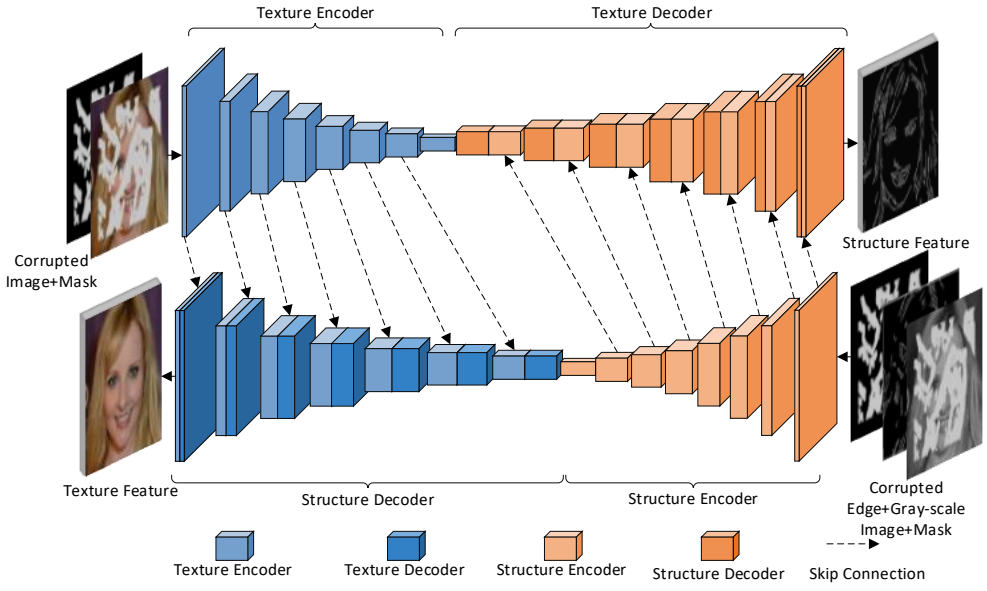


*Figure 5: Two stream networks in CTSD.*

## 3. Experiments

Image inpainting based on deep learning requires massive data to participate in training. The experiments are conducted on CelebA-HQ dataset and NVIDIA irregular mask, using 1000 images for testing. CelebA-HQ is a high-resolution dataset derived from the CelebA dataset, containing 30,000 face images. The NVIDIA irregular mask dataset contains 55,116 training sets and 24,866 test sets. It is

ISSN 2616-5775 Vol. 5, Issue 3: 70-77, DOI: 10.25236/AJCIS.2022.050310

divided into 6 different proportions according to the size of the holes, which randomly located in any position of the image.

EC [11], PRVS [12], MED [13], CTSD [14] are conducted with 8G NVIDIA RTX 3060 ti, Gconv [10] is trained on 2 × NIVIDIA RTX 2080 Super GPU for TensorFlow version reasons, and all images used are resized to 256×256 pixels. PConv [9] official did not give the overall code and did not conduct the experiment.

### *3.1. Qualitative Comparisions*

Six results are randomly selected from five algorithms, and all results are the direct output of trained models without additional subsequent processing. When the mask radio is between 10% and 30%, as shown in the first row to the third row in Figure 6, GConv [10] based on two stages can produce sharp texture details, but the face is distorted and the structure is too smooth due to without considering the structure information. EC [11], PRVS [12], and CTSDG [14] can usually generate relatively reasonable structures, but the details are not perfect. The filled area in MED [13] has blurred and distorted phenomenon, such as artifacts in the mouth of the third row of faces. With the mask area gradually increasing at 30%-40%, the CTSDG [14] in the fourth row more closely approximates the original image and achieves better results in both texture and structure, with PRVS outperforming EC [11] in structural detail but with poorer hair detail. When the mask area is between 40% and 60%, artifacts appear in the GConv [10] in the fifth row and sixth row, and the face is distorted in EC. The structure of the progressive in PRVS [12] generation structure is more reasonable but distorted, the MED [13] has artifacts, and the edges of CTSDG [14] are blurred.


*Figure 6: Qualitative comparison on CelebA-HQ and NVIDIA irregular mask.*

### *3.2. Quantitative Comparisions*

The most commonly used objective evaluation metrics in image inpainting are PSNR, SSIM, MAE, FID, etc., which are compared in this paper in different irregular mask ratios. Table 3 shows the results implemented on CelebA-HQ, where LPIPS [15] is computed on deep features of VGG pre-trained on ImageNet, which is more consistent with human perception than traditional metrics. Overall Objective evaluation shows that the CTSDG [14] algorithm is superior to other algorithms.

ISSN 2616-5775 Vol. 5, Issue 3: 70-77, DOI: 10.25236/AJCIS.2022.050310

*Table 3: Quantitative comparison on CelebA-HQ and NVIDIA irregular mask.*

| Ratio | Method | PSNR↑ | SSIM↑ | MAE↓ | FID↓ | LPIPS↓ |
|---|---|---|---|---|---|---|
| 10-20% | GConv [10] | 30.92 | 0.969 | 0.0106 | 4.87 | 0.057 |
| | EC [11] | 31.23 | 0.971 | 0.0096 | 4.24 | 0.051 |
| | PRVS [12] | 31.81 | 0.975 | 0.0084 | 3.85 | 0.050 |
| | MED [13] | 32.92 | 0.979 | 0.0089 | 3.35 | 0.042 |
| | CTSDG [14] | **33.26** | **0.981** | **0.0070** | **3.15** | **0.038** |
| 20-30% | GConv [10] | 27.51 | 0.937 | 0.0183 | 8.41 | 0.095 |
| | EC [11] | 27.94 | 0.942 | 0.0171 | 8.06 | 0.085 |
| | PRVS [12] | 28.60 | 0.949 | 0.0151 | 7.02 | 0.082 |
| | MED [13] | 29.15 | 0.954 | 0.0155 | 6.34 | 0.074 |
| | CTSDG [14] | **29.63** | **0.959** | **0.0130** | **6.19** | **0.070** |
| 30-40% | GConv [10] | 25.15 | 0.898 | 0.0272 | 11.45 | 0.132 |
| | EC [11] | 25.51 | 0.902 | 0.0259 | 11.54 | 0.122 |
| | PRVS [12] | 26.38 | 0.919 | 0.0227 | 10.03 | 0.116 |
| | MED [13] | 26.49 | 0.921 | 0.0235 | 9.00 | 0.109 |
| | CTSDG [14] | **27.14** | **0.931** | **0.0201** | **8.88** | **0.104** |
| 40-50% | GConv [10] | 23.27 | 0.850 | 0.0375 | 15.02 | 0.174 |
| | EC [11] | 23.49 | 0.847 | 0.0368 | 16.71 | 0.163 |
| | PRVS [12] | 24.53 | 0.879 | 0.0316 | 13.98 | 0.153 |
| | MED [13] | 24.39 | 0.876 | 0.0333 | 12.59 | 0.148 |
| | CTSDG [14] | **25.14** | **0.894** | **0.0286** | **12.83** | **0.143** |
| 50-60% | GConv [10] | 20.45 | 0.736 | 0.0587 | 19.12 | 0.233 |
| | EC [11] | 20.51 | 0.719 | 0.0589 | 25.12 | 0.222 |
| | PRVS [12] | 21.99 | 0.793 | 0.0483 | 21.47 | 0.203 |
| | MED [13] | 21.28 | 0.768 | 0.0544 | **17.98** | 0.204 |
| | CTSDG [14] | **22.24** | **0.802** | **0.0465** | 19.27 | **0.201** |

## 4. Conclusion

In this paper, we detail the principles of six classical irregular face inpainting methods based on deep learning, experiment on CelebA-HQ, and quantitatively and qualitatively compare and evaluate the performance of the results. Experiments have shown that reasonable results can be produced when the hole is small. When the mask radio is large, each algorithm has different limitations. GConv [10] with two stages suffers from over-smoothing. EC [11] uses edge priori to guide the inpainting and fails to generate a reasonable visual structure resulting in distortion of the face. PRVS [12] with progressive structure generation produces a reasonable visual structure, but suffers from face distortion. The structure and texture information in MED [13] share a single generator, resulting in the failure to generate clear texture and other problems such as artifacts. CTSDG [14] has overall superior performance but edges are blurred.

## Acknowledgements

This work was supported by the Electrical Information (Computer Technology) Postgraduate Joint Training Base of Chongqing Normal University and Chongqing Century Keyo Technology Co., Ltd. It was supported partially also by the Research Project of Higher Education Reform of Chongqing Education Commission under Grant 201017S.

## References

*[1] Esedoglu S, Shen J. Digital inpainting based on the Mumford–Shah–Euler image model[J]. European Journal of Applied Mathematics, 2002, 13(4): 353-370.*
*[2] Liu D, Sun X, Wu F, et al. Image compression with edge-based inpainting[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2007, 17(10): 1273-1287.*
*[3] Darabi S, Shechtman E, Barnes C, et al. Image melding: Combining inconsistent images using patch-based synthesis[J]. ACM Transactions on graphics (TOG), 2012, 31(4): 1-10.*
*[4] Huang J B, Kang S B, Ahuja N, et al. Image completion using planar structure guidance[J]. ACM Transactions on graphics (TOG), 2014, 33(4): 1-10.*

ISSN 2616-5775 Vol. 5, Issue 3: 70-77, DOI: 10.25236/AJCIS.2022.050310

*[5] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.*
*[6] Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: Feature learning by inpainting[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2536-2544.*
*[7] Iizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics (ToG), 2017, 36(4): 1-14.*
*[8] Yu J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5505-5514.*
*[9] Liu G, Reda F A, Shih K J, et al. Image inpainting for irregular holes using partial convolutions[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 85-100.*
*[10] Yu J, Lin Z, Yang J, et al. Free-form image inpainting with gated convolution[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4471-4480.*
*[11] Nazeri K, Ng E, Joseph T, et al. Edgeconnect: Generative image inpainting with adversarial edge learning[J]. arXiv preprint arXiv:1901.00212, 2019.*
*[12] Li J, He F, Zhang L, et al. Progressive reconstruction of visual structure for image inpainting[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5962-5971.*
*[13] Liu H, Jiang B, Song Y, et al. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer International Publishing, 2020: 725-741.*
*[14] Guo X, Yang H, Huang D. Image Inpainting via Conditional Texture and Structure Dual Generation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 14134-14143.*
*[15] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 586-595.*
*[16] You Yang, Kesen Li, Sixun Liu, Ling Luo, Bin Xing. Perceptual Face Inpainting with Multi-column Gated Convolutional Network[J]. Journal of Electronic Imaging, 31(1), 013022 (2022). https://doi.org/10.1117/1.JEI.31.1.013022*