Inpainting with Sketch Reconstruction and Comprehensive Feature Selection

Siyuan Li[1][0000−0002−2354−4233], Lu Lu[1], Zhijing Li[2], Kepeng Xu[1], Matthieu Claisse[3], Wenxin Yu[1*], Gang He[1], Gang He[4], and Zhuo Yang[5]

[1] Southwest University of Science and Technology
[2] Accenture Japan Ltd
[3] cole internationale des sciences du traitement de l'information
[4] Xidian University
[5] Guangdong University of Technology
[*]yuwenxin@swust.edu.cn

Abstract. With the advent of the convolutional neural network, learning-based image inpainting approaches have received much attention, and most of these methods have been attracted by adversarial learning and various loss functions. This paper focuses on the enhancement of the generator model and guidance of structural information. Hence, a novel convolution block is proposed to comprehensively capture the context information among feature representations. The performance of the proposed method is evaluated on Place2 test dataset, which outperforms the current state-of-the-art inpainting approaches.

Keywords: Image Inpainting · Deep Learning · Feature Selection · Edge Guidance.

1 Introduction

Image inpainting, also known as image completion, is the process of restoring the missing parts in a damaged image. It plays a vital role in all sorts of computer vision tasks. For example, it can be used in removing some specific objects from an image. However, because the corrupted region only can be inferred through its neighborhood, it is a challenging task to recover the details of the corrupted region to completely match the original image. Therefore, making full use of context information is a commonly used means of making the generated image visually sensible. In addition, especially in the case where the damaged region contains complex gradient information, the fine structure in the filled region can guide the inpainting model to produce sharper results and away from the blurred edges, so as to improve inpainting quality and make filled region reasonable. Therefore, if we first paint the missing area with the fine structure or precise edges, the final results guided by the repaired sketch will be greatly improved.

In consideration of these pieces of knowledge, this paper proposes a two-stage, learning-based image inpainting apporach (Figure 1) with enhanced generator model and a new type of convolution block. Similar to one of the most advanced works [12], the two stages of our proposed method are sketch reconstruction and image completion phases. Although both of the two stages also introduce the generative adversarial network (GAN) [3], this paper put more attention to the enhancement of the generator model rather than the discriminator model.

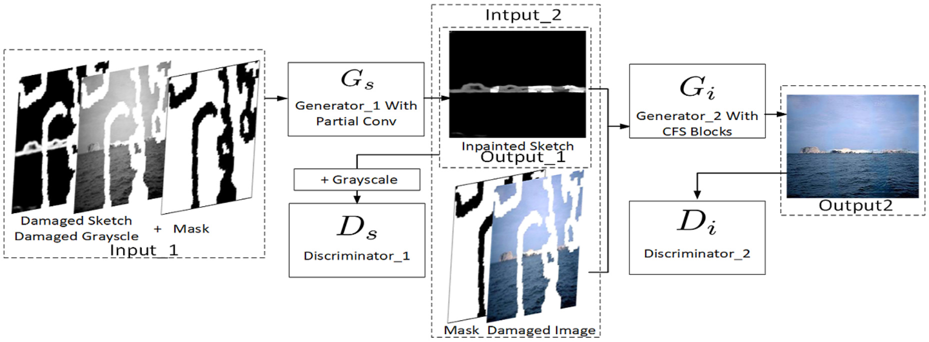

Fig. 1. The pipeline of the proposed approach.

The first stage, sketch reconstruction phase, aims to recover the gradient information in the missing area from corrupted sketch maps. In this paper, the Holistically-nested edge detection (HED) [16] is introduced to generate the sketches maps of the original images, the HED method, which is one of state-of-the-art learning-based edge detector, has the ability to resolve the challenging ambiguity around the object boundary, therefore the edges generated by HED are almost connected and represent the basic outline of objects. It is worth noting that Kamyar's work [12] has conducted some experiments that also tried to use the edge maps generated by HED, however, their edge generator model fails to achieve better accuracy than using Canny edge detector [1] in the first phase. Even with poor edge prediction, the final inpainted results guided by HED sketches are not worse than those led by Canny detector. Therefore, the sketch map generated by HED is considered as guiding information in this paper, and we enhance the sketch generator model by increasing the convolution layers and applying some of the popular techniques to competent the sketch reconstruction task.

In the second phase, the goal of image completion networks is exploiting the repaired sketch maps and corrupted raw RGB image to color the sketch in the filled region. This paper applies a new convolution module named Comprehensive Feature Selection Block (CFS Block) in the second phase to comprehensively capture the saliency of context information among the feature maps. And the weight of the proposed module can be automatically updated during the training phase by the backpropagation.

The models we presented are evaluated on the test dataset of Place2 [20], compared with those state-of-art inpainting approaches, the produced results quantitatively achieve great improvement. In addition, we compare produced results against [12] to observe the improvement in the qualitative aspect. To sum up, our paper makes the following contributions:

- *The introduction of the edge sketch produced by HED, which better represents the rough shape of objects in images.*
- *A reinforced edge generator that can repair or hallucinate (imagine) the sketch map through rest of sketch.*
- *An integrated inpainting generator with a novel convolution block – CFS Block.*

2 Related Work

In the past few years, a variety of learning-based approaches have flourished the field of image inpainting. Most of these learning-based methods have the ability to fill in the damaged area by hallucinating (imagining) or estimating some novel objects that may exist in the real world. This benefit from the context encoder [14] work, which is one of the first inpainting method applying deep learning.

The context encoder [14] embeds the corrupted image into the high-level feature maps with low-dimensional, which the decoder then use to reconstruct the predicted image. However, due to its monotonic loss function and simple network structure, the generated images are usually blurry and contains many inconsistent artifacts. But on the other hand, the encoder-decoder neural network architecture has been commonly used as a generator model in the learning-based approaches for image inpainting during recent years.

Since Goodfellow proposed GAN [3], using adversarial learning to enhance the quality of images has become popular. Iizuka [5] typically proposed two discriminators to reinforce the global and local consistent of restored images. However, the training process is fragile and hard to converge due to the use of raw discriminator without optimization measures, and their outcomes largely rely on the post-processing to eliminate style inconsistencies around the boundaries of filled region.

In order to adapt to the characteristics of inpainting, Liu et al. [9] renormalize the traditional convolution layer into a Partial Convolution layer which calculation is based on the mask maps. This convolution layer only calculates the valid pixel in image or feature map, while the validity of pixels is determined by the corresponding mask map. Since the perceptual loss and style loss are introduced into this work, their model is able to eliminate the color inconsistencies around the borderline. Although their result is visually plausible, their structural information in the filled region is mismatch the neighborhood.

Yu et al. [18] proposed Gated Convolution and use hand-written sketches to guide the inpainting process. Their work assumes their coarse estimate recovered by the first stage are reasonably consistent with the original image, and their Gated Convolution in each layer only selects convolved feature values meanwhile lacking a channel-wise selection mechanism. For example, the original image, mask map, and their user-guided sketch should have different importance for inpainting work, however, the Gated Convolution ignores calculating the weights of these features.

Kamyar et al. [12] proposed two-stage inpainting schemes which firstly restores the damaged edge map and then colors the image with the recovered edge map. However, they suppose the edge inpainting process is a relatively easy task and don't put enough attention on the edge generator model. But the result of edge inpainting is crucial to the final image quality. Some of the recovered edge maps with artifacts lead the blurriness.

Although our work is close to the combination of Kamyar's [12] and Yu's [18], our work proposed a novel convolution block to comprehensively select the features among the input features and convolved features in current module meanwhile enhancing the generator model through incorporating lots of popular technology to assure the accuracy of sketch and texture prediction.

3 Approach

The proposed scheme divides the learning-based image inpainting process into two phases. At each stage, we create a generator model to produce target image meanwhile establishing a discriminator model that feedback to the generator to help produce high-quality results. In the first phase, the corrupted sketch and grayscale image are concatenated as feature map, the features and the binary mask map which use 1 represent the non-damaged region are fed into the generator with Partial Convolution, then the generator predicts a complete sketch as output. At the second stage, the restored sketch and damaged RGB image are considered as features together, they are inputted to a new generator built by CFS Blocks, then the second generator aims to get a complete RGB image.

In this section, we describe the detailed architecture of proposed networks in each phase and the detailed design of the Comprehensive Feature Selection Block (CFS Block) and briefly analyze what CFS Block actually doing.

3.1 Networks

The general architecture of the proposed model for each phase is illustrated as Figure 2. The intention of the first 4 convolutions (CFS Block or Partial Conv) with stride 2 in both generators is extracting high-level representation and compress the feature to low resolution. Vice versa, the last 4 deconvolution layer including upsample and convolution (CFS Block or Partial Conv) are designed to level up the resolution and restore the image. According to the literature [19], batch normalization [6] deteriorates the color coherence of restored image. Thus this paper doesn't add any batch normalization to neither generator or discriminator, but the Leaky ReLU Activations [17] is stiil used in each layer except output layer.

As mentioned in [10], Spectral normalization can further stabilize the training process through utilizing the maximum singular value of the weight matrix

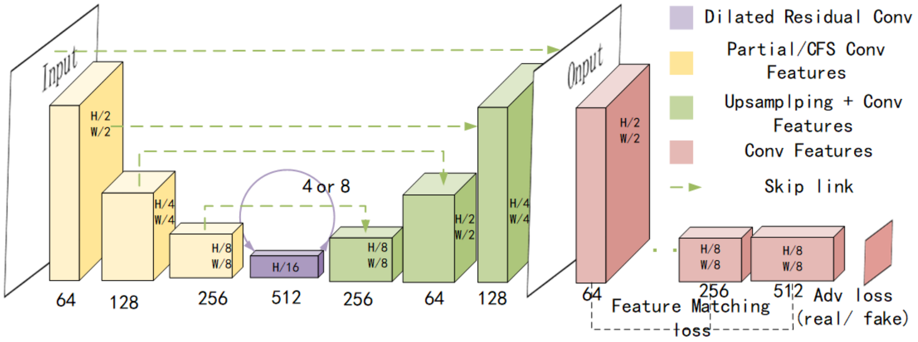

Fig. 2. The design of networks in one of stages.

to reduce each weight matrix, which limits the Lipschitz constant of functions to 1. In spite of spectral normalization was originally used only in discriminators, Odena [13] has recently shown that spectral normalization can keep generators away from dramatic changes of parameters and gradient. Therefore, we apply spectral normalization to the generators and discriminators in both phases.

The structures of proposed discriminators in different phases (sketch reconstruction phase and image completion phase) are exactly the same. In the first 3 layers of discriminators, we use vanilla convolution with stride 2 to extract the high-level representation of repaired image (or real image). Then following the two convolution layer with same padding reduce the number of channels to 1 after increasing the number.

Compared with the discriminators, the difference between the generators is greater. In sketch reconstruction phase, the task of generator is relatively easy, thus we apply Partial Convolution [9] to each convolution layer in this generator instead of applying CFS Block, another difference is that we set 4 dilated residual block with Partial Convolution in the middle part of generator rather than 8 dilated residual blocks with CFS in image completion generator.

Partial Convolution In the image inpainting work of Liu [9], the Partial Convolution is masked and redefined to be conditioned on non-damaged pixels. Their mask-based convolutions method with an automatic mask updating mechanism outperformed other irregular image inpainting works at that time. Therefore, all calculations are changed slider by slider in the Partial Convolution. For a brief representation, one of sliding operation of Partial Conv (at every kernel-size location in feature map/ original image) is expressed as Eq. (1).

$$x^{'} = \begin{cases} W^T(x \odot m)\frac{K_x K_y}{sum(M)} + b & if\ sum(m) > 0 \\ 0 & otherwise \end{cases} \tag{1}$$

In which W denotes the filter weights in vector form and b denotes the filter bias, x is the vector of flattened feature values in the current sliding window, M is x's corresponding slider in binary mask, $\odot$ represents element-wise multiplication, and K_x K_y correspond to the width and length of filter. This operation causes the convolution in the current slider to only process valid pixels, meanwhile balancing the magnitude of convolved value in the sliders which has invalid pixels. And the mask-updating function of [9] in each slider can be formalized as Eq. (2) [9].

$$m^{'} = \begin{cases} 1 & if\ sum(m) > 0 \\ 0 & otherwise \end{cases} \tag{2}$$

This updating process happens in each Partial Conv layers except the last layer, the updated mask determines whether pixels are valid in the next hidden layers.

3.2 Comprehensive Feature Selection Block

The inspiration of the front part of CFS Block comes from classical Gated Recurrent Units (GRU) [2] and Gated Convolution [18], whileas the tail of the module directly introduce the Squeeze-and-Excitation [4] block. The proposed integrated CFS Block aims to not only emphasize spatial relationships but also to further concern about channel correlation in feature selection process.

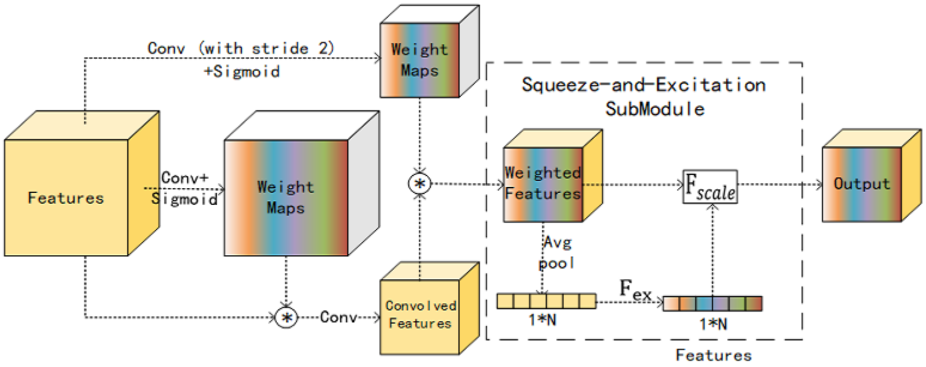

Fig. 3. The internal structure of Comprehensive Feature Selection Block.

As shown in Figure 3, we adopt two convolution layer followed by sigmoid activation from the input features to calculate the weights of input features and the weights of convoluted features. Before the convolution, the input features are multiplied by their corresponding weights, and the convolved values also are multiplied by their weights. The slider-wise process is described as follows.

$$g = \sigma(W_g^T x + b_g) \tag{3}$$

$$G = \sigma(W_G^T x + b_G) \tag{4}$$

$$f = G \cdot \phi(W_f^T (g \cdot x) + b_f) \tag{5}$$

In which the g represents the gating value of the original feature in one of sliding windows and G is the gating values (weight) of the convolved feature map, the σ denotes sigmoid function and ϕ represents the Leaky ReLU activation with the slope of 0.2. The f corresponds to the weighted feature computed by those foregoing units. The meaning of other symbols can be referred to the interpretation of Eq. (1) .

In the Squeeze-and-Excitation[4] submodule, it utilizes the average pooling layer to extract the global information in each channel, then applies two fully-connected (FC) layers with nonlinear activation function to get the weight of each channel, finally, the weighted features are multiplied by the weight of channels to get the final output of the whole module. If denote the input of this submodule as $I = [i_1, i_2, ..., i_C] = f$ (C is the number of feature channels), these processes can be formalized as follows:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} i_c(i, j) \tag{6}$$

$$s = F_{ex}(z, W) = \sigma(W_2 \psi(W_1 z)) \tag{7}$$

$$o_c = F_{scale}(i_c, s_c) = s_c i_c \tag{8}$$

Here $H \times W$ is the spatial dimensions of i_c, ψ refers to the ReLU [11] activation, $z \in \mathbb{R}^C$, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ (r is set to 16 in experiments). F_{scale} function as channel-wise multiplication between the scalar s_c and the cth channel of input $i_c \in \mathbb{R}^{H \times W}$. Thus, the output $O = [o_1, o_2, .., o_C]$ in this submodule also refers to the output of CFS Block

Discussion about CFS Block All of the parameters in CFS Block (except r) are learnable in the training process, this means that all the gating values (weight maps) in CFS Block can automatically be updated from data, it enables the generator to learn weight maps dynamically thus can select features both in input feature maps and the features in the next level (Gating values have also been proved significantly improve inpainting results in [18]). Because the importance of the damaged image, masks and sketches are obviously different, it is reasonable to set an additional gated convolution to select the input features rather than directly applying cross-level learning pattern [18] which may increase the difficulty of learning. Furthermore, properly selecting the feature maps is critical for image inpainting, the Squeeze-and-Excitation submodule is set to reinforce the ability of the model to notice some essential channels in the computed features. However, the number of parameters in CFS Block is a bit large, this is the reason why the relatively easy task, the sketch inpainting task, adopts Partial Conv instead of CFS Block.

3.3 Loss Function

For convenience of formulizing, this paper denotes the generator and discriminator in sketch reconstruction as G_s and D_s, denotes the generator and discriminator in image completion as G_i and D_i, and the binary mask that labels the valid pixels as 1 (invalid pixels as 0) is expressed as M, the ground truth images in dataset is represented as $\mathbf{I}_{gt}$, damaged image can be represented as $\acute{\mathbf{I}} = \mathbf{I}_{gt} \odot M$, the complete sketch generated by HED is S_{gt}, the incomplete sketch is $\acute{\mathbf{S}} = \mathbf{S}_{gt} \odot M$, the composite sketch is described as $\mathbf{S}_{comp} = \mathbf{S}_{pred} \odot (1-M) + \mathbf{S}_{gt} \odot M$

The sketch generator predicts the complete sketch by considering the concatenation of damaged grayscale image and damaged sketch as the features, meanwhile inputting the mask M for Partial Conv. While in the image completion generator, it concatenates the inpainted sketch, damaged image and mask as the input feature.

$$\mathbf{S}_{pred} = G_s(\left[\acute{\mathbf{I}}_{gray}, \acute{\mathbf{S}}\right], M) \tag{9}$$

$$\mathbf{I}_{pred} = G_I\left(\left[\acute{\mathbf{I}}, \mathbf{S}_{comp}\right]\right) \tag{10}$$

On account of the feature-matching loss [15] is introduced to one of loss terms for further optimizing the generator, the total loss for sketch generator is interpreted as Eq. (10).

$$\min_{G_s} \max_{D_s} \mathcal{L}_{G_s} = \min_{G_s} \left(0.1 \max_{D_s} (\mathcal{L}_{D_s}) + 10\mathcal{L}_{FM}\right) \tag{11}$$

The task of the discriminator is to distinguish whether the input sketch in discriminator belongs to the grayscale image of the corresponding ground truth image, this paper adopts the hinge loss as the target function of discriminator, which train the discriminator more strictly.

$$\begin{aligned} \mathcal{L}_{D_s} = \mathbb{E}_{\mathbf{S}_{gt}} &\left[\psi(1 - D_s(\mathbf{S}_{gt}, \mathbf{I}_{gray}))\right] \\ &+ \mathbb{E}_{\mathbf{S}_{pred}}\left[\psi(1 + D_s(\mathbf{S}_{pred}, \mathbf{I}_{gray}))\right] \end{aligned} \tag{12}$$

The role of feature-matching loss term is similar to the perceptual loss [7] in the task of inpainting, it can be formulized as :

$$\mathcal{L}_{FM} = \mathbb{E}\left[\sum_{i=1}^{L} \frac{1}{N_i} \left\| D_s^{(i)}(\mathbf{S}_{gt}) - D_s^{(i)}(\mathbf{S}_{pred}) \right\|_1\right] \tag{13}$$

where L is the number of layers in the discriminator, while N_i corresponds to the number of total activation units in the ith convolution layer, and $D_s^{(i)}$ represents the feature values of activation units in the ith convolution layer.

Image completion generator adopts L1 distance and the perceptual loss ($\mathcal{L}_{style}$ and $\mathcal{L}_{perc}$) [7]. It enables the model to learn the high-level representation and remove the checkboard artifacts from the predicted image, the total loss of completion generator is express as

$$\mathcal{L}_{G_i} = \mathcal{L}_{\ell_1} + 0.1\mathcal{L}_{D_i} + 300\mathcal{L}_{perc} + 300\mathcal{L}_{style} + 10\mathcal{L}_{FM} \tag{14}$$

where $\mathcal{L}_{D_i}$ is similar to $\mathcal{L}_{D_s}$ however the input of $\mathcal{L}_{D_i}$ has a slight difference, it just input the $\mathbf{I}_{gt}$ and $\mathbf{I}_{pred}$ without any grayscale images.

4 Experiments

All of the experiments in this paper are conducted in the dataset of Place2 [20], which training dataset contains 1803460 images with the resolution of 256x256 and test set contains 328000 images. The sketches are inferred from RGB images through HED [16] approach. The irregular mask dataset used in this paper comes from the work of Liu [9]. With these data groups (image sketch mask), We train the proposed models on single NVIDIA 1080TI with a batch size of 6 until the generators converge using Adam optimizer [8].

4.1 Quantitative Results

The quantitative results in the test dataset of Place2 are shown in Table 1, this table also shows some results produced popular inpainting methods in comparison. In the case of all different ratios of the damaged region, the table indicates that our results outperform the others in PSNR (Peak Signal-to-Noise) and SSIM (Structural Similarity) metric, especially in the case of small masks. In addition, the unique sketch prediction task achieved 77% accuracy (Binary Classification problem), thus the quantitative improvement is benefited from the enhanced sketch generator and the HED-style sketch, which can not only describe the obvious edges but also describe the subtle structure.

Table 1. Comparison of quantitative results. CA means Context Attention[19] method, the GLCIC represent the result of Iizuka's[5], PConv refer to Lius'[9], and EdgeCnt correspond to EdgeConnect approach[12], these data are taken from this paper[12].

Mask		CA	GLCIC	PConv	EdgeCnt	Ours
10-20%	PSNR	24.36	23.49	28.02	27.95	**30.85**
	SSIM	0.893	0.862	0.869	0.920	**0.951**
20-30%	PSNR	21.19	20.45	24.90	24.92	**26.87**
	SSIM	0.815	0.771	0.777	0.861	**0.900**
30-40%	PSNR	19.13	18.50	22.45	22.84	**24.17**
	SSIM	0.739	0.686	0.685	0.799	**0.841**
>=40	PSNR	17.75	17.17	20.86	21.16	**21.74**
	SSIM	0.662	0.603	0.589	0.731	**0.7695**

4.2 Qualitative Results

The EdgeConnect approach[12] has recently shown surprising advancement in image inpainting, therefore we compare our result with this state-of-art work (illustrated in Figure 4). The red box in the figure indicates that the proposed approach with CFS Block can produce a clearer color image with the more obvious edges than EdgeConnect, the 5th column show the generated sketch in the proposed approach, the orange lines in the image represent the inpainted line by generator, it demonstrates the sketch generator with Partial Conv already can handle the sketch reconstruction task.(More results can be found in Figure 5)

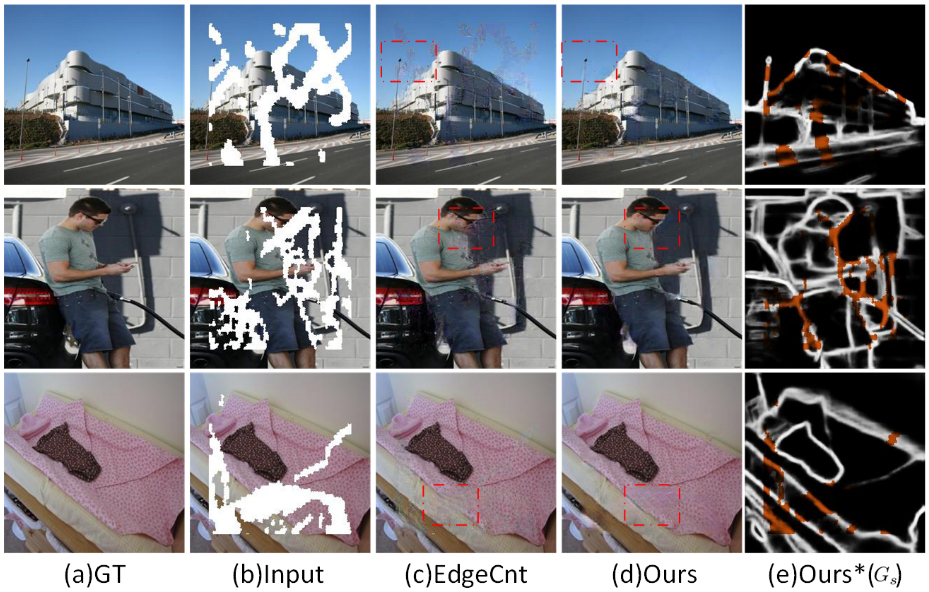
(a)GT (b)Input (c)EdgeCnt (d)Ours (e)Ours*(G_s)

Fig. 4. The qualitative comparison of results. (a) Ground Truth image. (b)Corrupted image. (c) EdgeConnect[12]. (d)Ours. (e) Restored sketches of Ours.

5 Conclusions

This paper proposed a two-stage image inpainting approach with a novel feature selection mechanism – CFS Block, and proves that the enhanced sketch generator and the proposed comprehensive feature selection mechanism can significantly improve the inpainting results. The qualitative comparisons show that the proposed approach produces visually more pleasing results, and the objective evaluations in various sizes of masks demonstrate the superiority of the proposed approach.

6 Acknowledgements

This research was supported by【2018GZ0517】【2019YFS0146】【2019YFS0155】 which supported by Sichuan Provincial Science and Technology Department. Supported by State Key Laboratory of ASIC & System【2018KF003】, Science and Technology Planning Project of Guangdong Province【2017B01011

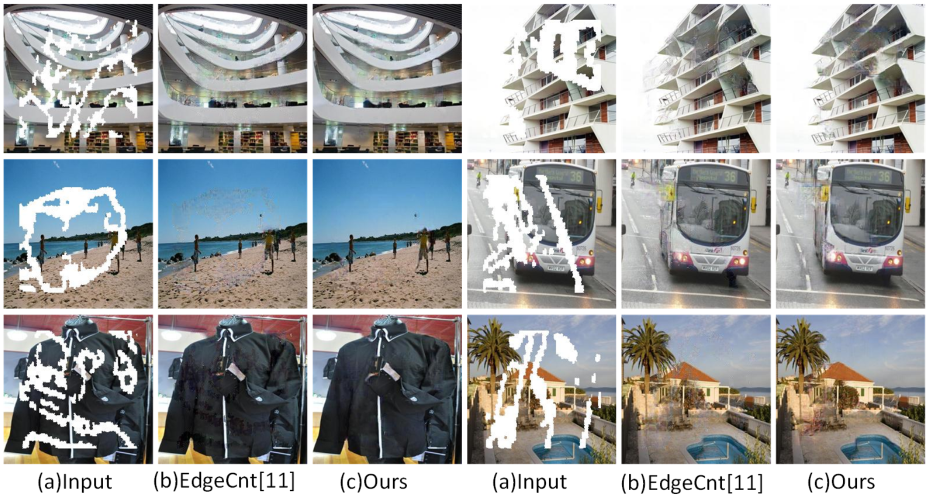

Fig. 5. The additional qualitative comparison. (a) Corrupted image. (b) EdgeConnect[12]. (c) Ours.

0007】. Supported by Postgraduate Innovation Fund Project by Southwest University of Science and Technology（19ycx0050）.

References

1. Canny, J.: A computational approach to edge detection. In: Readings in computer vision, pp. 184–203. Elsevier (1987)
2. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR (2014), http://arxiv.org/abs/1406.1078
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
4. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. CoRR (2017), http://arxiv.org/abs/1709.01507
5. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics (ToG) **36**(4), 107 (2017)
6. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
7. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 85–100 (2018)

10. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
11. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 807–814 (2010)
12. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. CoRR **abs/1901.00212** (2019)
13. Odena, A., Buckman, J., Olsson, C., Brown, T.B., Olah, C., Raffel, C., Goodfellow, I.: Is generator conditioning causally related to gan performance? arXiv preprint arXiv:1802.08768 (2018)
14. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
15. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8798–8807 (2018)
16. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1395–1403 (2015)
17. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. CoRR **abs/1505.00853** (2015), http://arxiv.org/abs/1505.00853
18. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. arXiv preprint arXiv:1806.03589 (2018)
19. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
20. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)