Base-Detail Image Inpainting

Ruonan Zhang[1]
zhangrn@pcl.ac.cn
Yurui Ren[2]
yrren@pku.edu.cn
Jingfei Qiu[1]
qiujf@pcl.ac.cn
Ge Li*[2]
geli@ece.pku.edu.cn

[1] Peng Cheng Laboratory
No.2, Xingke 1st Street, Nanshan
Shenzhen, China

[2] Peking University Shenzhen graduate school
University Town, Nanshan District
Shenzhen, China

Abstract

Recent advances in image inpainting have shown exciting promise with learning-based methods. Though they are effective in capturing features with some prior techniques, most of them fail to reconstruct reasonable base and detail information, so that the inpainted regions appear blurry, over-smoothed, and weird. Therefore, we propose a new "Divider and Conquer" model called Base-Detail Image Inpainting, which combines the reconstructed base and detail layers to generate the final subjective perception images. The base layer with low-frequency information can grasp the basic distribution while the detail layer with high-frequency information assists with the details. The joint generator overall would benefit from these two as guided anchors. In addition, we evaluate our two models over three publicly available datasets, and our experiments demonstrate that our method outperforms current state-of-the-art techniques quantitatively and qualitatively.

1 Introduction

Image inpainting(a.k.a. image completion or image hole-filling), involves filling in the corrupted areas of images. Since humans have an uncanny ability to zero in on visual inconsistencies, the filling-in must be perceptually plausible. Thus, how do humans plausibly fill the hole in the image? First, we should know there are four typical cases: (1) Single texture. (2) Multiple textures. (3) Single or Multiple textures. (4) Content with strong semantics. Many satisfactory fillings [1, 2, 4, 8] exist in case (1) and (2). Recently, many approaches [9, 19, 25] try to find good solutions for case (3) and (4), but the filling-in is very contrived in case (4) and thus remains challenging.

Broadly speaking, traditional approaches can be divided into two groups: diffusion-based methods [2, 4] which propagate background data into the missing region by adopting a diffusive process typically modeled using different operators, and patch-based methods [1, 8], which fill corrupted regions with patches from a collection of source images (search space) that maximize patch similarity. While they do a better job for cases (1) and (2), they fail to reconstruct complex details and also need post-processing to blend the image.

(a) Corrupted input images (b) Coarse base layer (c) Coarse detail layer (d) Final result

Figure 1: An example of the proposed method.

With the development of deep learning and wide-spread GAN [6], a series of learning-based methods rushed in [9, 14, 19, 24, 25] have successively found remarkable results for complex scenes. These methods can generate corrupted regions with meaningful structures, but struggle to generate high-frequency details accurately.

This paper is motivated by the observation that many existing techniques generate blurry and weird filling-ins without base and detail information. At the same time, the "Divide and Conquer" model in algorithms gives us a big gift for solutions. The original retinex-based methods [13] successfully use this idea for image enhancement. From this point of view, in image inpainting, we could decompose the corrupted images into base and detail layers and reproduce them separately, and after that, combine them to obtain the desired result. Then, the joint base-detail generator can balance in a more definite direction fed with the coarse reconstructed base and detail layer information. Meanwhile, the network feels relaxed by targeting at the smaller divided parts which is easy to converge.

In the proposed architecture shown in Fig.2, the base reconstructor is solely focused on describing the basic distribution, while the detail reconstructor mainly reproduces the local details. They are then jointly sent to the base-detail generator as good guides to get the final perceptually plausible results as Fig.1(d). In the meantime, two models both with and without switchable normalization (SW) [16] are introduced in the base-detail generator. Two different scaled patch-based GANs [10] are also used to predict real *vs.* fake for overlapping image patches with two different sizes. Finally, we evaluate our proposed model on standard datasets Celeba [15], Places2 [26], and Paris [3], and experiments demonstrate that our results are subjectively and objectively competitive against several current state-of-the-art schemes. Our contributions can be summarized in three:

- We design a base-detail backbone architecture, which contains the base reconstructor, the detail reconstructor, and the base-detail generator.

- The base reconstructor reproduces the base layer reflecting the basic distribution of images with low-frequency, while the detail reconstructor captures more detail and light intensity of images with high-frequency.

- The joint base-detail generator combines of base and detail reconstructors and fulfill the good potential of combining multiple losses to measure comprehensively. The introduced two models both with and without SW perform well. The final experimental results competitive compared with other state-of-the-art approaches.

2 Related Work

2.1 Image Inpainting

Traditional works can be mainly divided into two categories: diffusion-based and patch-based methods with low-level features. Diffusion-based methods [2, 4] propagate neighboring information. into the missing holes, which are limited to the local recovery of holes. Patch-based methods [1, 8] emphasize picking out the most similar patches between the source and target images(which usually are the same) to reconstruct and blend the holes. Consequently, these methods excel at recovering highly-patterned stationary textures but struggle at reconstructing non-stationary and locally-unique data. Meanwhile, most of them need an image blending technique as post-processing.

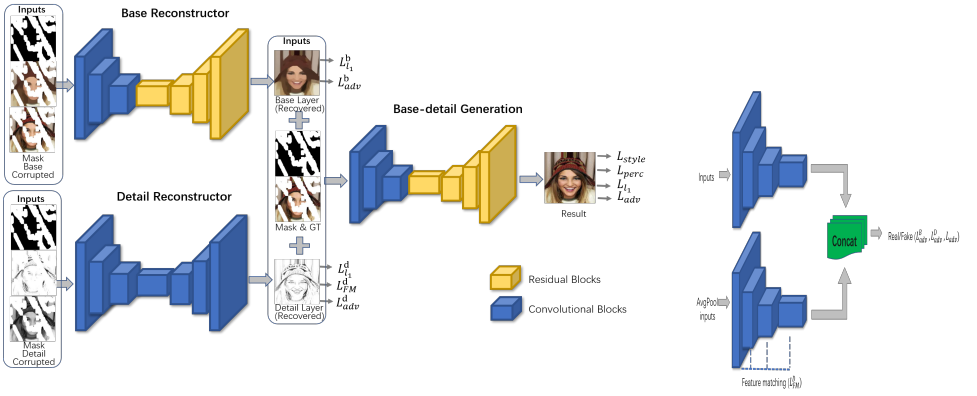
Recently, learning-based approaches with GAN have emerged as a promising paradigm for image inpainting. Initially, Context Encoders [19] uses an encoder-decoder architecture to fill-in holes. However, itsoutput contains visual artifacts due to the bottleneck in the channel-wise fully connected layer. Based on the Context Encoders, a series of related works [9, 24], [24] improved the architecture for high resolution images by using textures to boost the missing areas. [9] adds global and local discriminators to enhance the performance. These methods are limited to fixed center masks without analysis of irregular holes.

More recently, [14] introduced "partial convolution" (PConv) with irregular holes, where convolution weights are normalized by the mask area of the window. Context Attention (CA) [25] proposes a coarse-to-fine model with a context attention layer based on [9]. The following-up refinement network can refine the raw predictions, but cannot resolve all the artifacts. Edge-Connect (Edge) [18] learns to hallucinate edges in the missing regions to improve the output results for some highly structured scenes. Yet, the distribution of edges cannot keep pace with the target images, or to say, the only edge extractor ignores the basic information making it difficult to generate reliable textures(e.g. Fig.3). We are motivated to use some simple decomposed layers to generate reasonable sparse information as priors for image inpainting. To better handle the unstable distribution problem, we propose a base-detail network architecture with two reconstructors to repair base and detail information, respectively. The base reconstructor generates the base layer holding the global distribution of images while the detail reconstructor generates the detail layer providing details in gray-level.

3 Base-Detail Architecture

In this section, we present the details of our network, which consists of two reconstructors and one generator as shown in Fig.2(a). The base reconstructor concentrates on the coarse-level recovering of global information. while the detail reconstructor aims at fine-level inpainting of local details. The discriminator is presented in Fig.2(b) with two different patch-based GANs [10].

Let G_b and D_b be the generator and discriminator, respectively, for the base reconstructor, and similarly, G_d and D_d for detail reconstructor and G_{bd} and D_{bd} for the base-detail generator.



(a) Overview of our base-detail architecture (b) The discriminator

Figure 2: The proposed base-detail image inpainting architecture (Best viewed in color)

3.1 Base Reconstructor

The base layer can be seen as the backbone of an image. Visually, it can be regarded as a colorful draft in the drawing. Thus, we employ a well-behaved smoothing filter [23] to represent the ground-truth (GT) of the base layer labeled as $\mathbf{I}_b$. Assuming GT, the corrupted input, the corrupted base layer, and the mask of the images are represented as $\mathbf{I_{gt}}$, $\mathbf{I_c}$, $\mathbf{I_{cb}}$ and $\mathbf{M}$, respectively. $\mathbf{I_c} = \mathbf{I_{gt}} \circ (1 - \mathbf{M})$, $\circ$ denotes the element-wise product. $m = 1, m \in M$ represents the missing regions, and vice versa. Then the base reconstructor can be written as

$$\mathbf{B} = G_b(\mathbf{I}_c, \mathbf{I}_{cb}, \mathbf{M}), \tag{1}$$

where $\widehat{\mathbf{B}}$ is the predicted base layer images, and $\mathbf{I}_{cb} = \mathbf{I}_b \circ (1 - \mathbf{M})$. Meanwhile, the reconstruction loss is $\mathcal{L}^b_{\ell_1} = \|\mathbf{B} - \mathbf{I}_b\|_1$, which represents the similarity between B and its GT $\mathbf{I}_b$. Here, the adversarial loss of base reconstructor applying [6], which is a good option to mimic the distribution of the target reconstructor, can be shown as

$$\mathcal{L}^b_{adv} = \mathbb{E}_{\mathbf{I}_c \sim \mathbf{P}_c}[\log(1 - D_b(G_b(\mathbf{I}_c, \mathbf{I}_{cb}, \mathbf{M})))] + \mathbb{E}_{\mathbf{I}_{gt} \sim \mathbf{P}_{gt}}[\log(D_b(\mathbf{I}_{gt}))], \tag{2}$$

Finally, we train the generator G_b and the discriminator D_b by the below equation via optimizing the weighted-sum of the aforementioned losses together.

$$\min_{G_b} \max_{D_b} \mathcal{L}^b(G_b, D_b) = \lambda^b_{\ell_1} \mathcal{L}^b_{\ell_1} + \lambda^b_{adv} \mathcal{L}^b_{adv}, \tag{3}$$

where λs are weighted parameters. We pick a two-scale PatchGAN [10] to predict images in a pyramid-level architecture which permits photo-realistic synthesis. This spatial patch option can be understood as a form of base structure loss by replacing point-wise feature statistics. Meanwhile, we apply spectral normalization (SN) [17] to all generators and discriminators in our experiments since SN effectively restricts the Lipschitz constant of the network to one as well as suppresses the sudden fluctuation of parameters.

3.2 Detail Reconstructor

The detail layer can be seen as the refinement of the base layer. Suppose GT of gray-level, the corrupted gray-level input, the corrupted detail layer, of the images are represented as

$\widehat{\mathbf{I}}_{gt}$, $\widehat{\mathbf{I}}_c$, $\mathbf{I}_{cd}$, respectively, and $\widehat{\mathbf{I}}_c = \widehat{\mathbf{I}}_{gt} \circ (1-\mathbf{M})$. The detail reconstructor can be summarized as

$$\mathbf{D} = G_d(\widehat{\mathbf{I}}_c, \mathbf{I}_{cd}, \mathbf{M}), \tag{4}$$

where $\mathbf{D}$ is the predicted detail layer images. $\mathbf{I}_{cd} = \mathbf{I}_d \circ (1-\mathbf{M})$. The $\mathbf{I}_d$ is calculated by

$$\mathbf{I}_d = \frac{\mathbf{I}_{gt}+\varepsilon}{\mathbf{I}_b+\varepsilon}, \widehat{\mathbf{I}}_d \approx \frac{\widehat{\mathbf{I}}_{gt}+\varepsilon}{\widehat{\mathbf{I}}_b+\varepsilon}. \tag{5}$$

Since the obtained detail layer $\mathbf{I}_d$ is a 3-channel-image with similar details in each channel, we use an approximation to calculate $\widehat{\mathbf{I}}_d$ from the gray-level base layer denoted as $\widehat{\mathbf{I}}_b$. The gray-level base layer is calculated by the standard transformation $\widehat{\mathbf{I}}_b = 0.299 * R + 0.587G + 0.114B$, where R, G, B are 3 channels in $\mathbf{I}_b$ and $\mathcal{L}^d_{\ell_1} = \|\mathbf{D} - \mathbf{I}_d\|_1$. To stably train the process, the feature matching loss referenced from [22] is adopted as

$$\mathcal{L}^d_{fm} = \mathbb{E}[\sum_{L}^{i=1} \frac{1}{N_i} \|D^i_d(\widehat{\mathbf{I}}_{gt}) - D^i_d(\widehat{\mathbf{I}}_d)\|_1] \tag{6}$$

where L represents the final convolution layer of the discriminator, and N_i is the number of ingredients in the i^{th} layer of the discriminator. The feature-matching loss forces the generator to produce outputs with representations that are similar to real ones by comparing the activation maps in the intermediate layers of the discriminator. Then, the adversarial loss of detail reconstructor can be similar expressed as $\mathcal{L}^d_{adv} = \mathbb{E}_{\widehat{\mathbf{I}}_{\mathbf{c}} \sim \mathbf{P}_c}[\log(1 - D_d(G_d(\widehat{\mathbf{I}}_{\mathbf{c}}, \mathbf{I}_{\mathbf{cd}}, \mathbf{M})))] + \mathbb{E}_{\widehat{\mathbf{I}}_{gt} \sim \mathbf{P}_{gt}}[\log(D_d(\widehat{\mathbf{I}}_{\mathbf{gt}}))]$. Then, the optimization process can be clearly summarized as

$$\min_{G_d} \max_{D_d} \mathcal{L}^d(G_d, D_d) = \lambda^d_{\ell_1} \mathcal{L}^d_{\ell_1} + \lambda^d_{fm} \mathcal{L}^d_{fm} + \lambda^d_{adv} \mathcal{L}^d_{adv} \tag{7}$$

Unlike fluctuated edge information used in [18], our detail layer calculated from the base, serving as a good aid to support the better perceptually plausible outputs.

3.3 Base-detail Generation

The base-detail generator is

$$\mathbf{J} = G(\mathbf{I}_c, \mathbf{J}_{bd}, \mathbf{M}), \tag{8}$$

where $\mathbf{J}$ is the final generated image and $\mathbf{J}_{bd}$ represents the combination of $\mathbf{B}$ and $\mathbf{D}$. The reconstructor loss is $\mathcal{L}^{bd}_{\ell_1} = \|\mathbf{J} - \mathbf{I}_{gt}\|_1$. In particular, we select two losses called perceptual loss $\mathcal{L}_{perc}$ and style loss $\mathcal{L}_{sty}$ recommended in [6, 10]. Both losses keep up with the meaningful image properties and provides more freedom to the regressor in generating jobs. The perceptual and style losses are commonly derived respectively as

$$\mathcal{L}_{perc} = \mathbb{E}[\sum_i \frac{1}{N_i} \|\phi_i(\mathbf{I}_{gt}) - \phi_i(\mathbf{J})\|_1], \mathcal{L}_{sty} = \mathbb{E}_j[\|G^\phi_j(\mathbf{J}) - G^\phi_j(\mathbf{I}_{gt})\|_1], \tag{9}$$

where ϕ_i is the activation map of i^{th} layer in VGG-19 [20], and 4 layers from relu1-1 to relu4-1 are used in test. Additionally, G^ϕ_j is a $J_j \times J_j$ Gram matrix constructed from ϕ_j, and style loss judges the main difference between covariances. Analogously, the adversarial loss of base-detail generator D_{bd} can be expressed as $\mathcal{L}_{adv} = \mathbb{E}_{\widehat{\mathbf{I}}_{\mathbf{c}} \sim \mathbf{P}_c}[\log(1 - D_{bd}(G_{bd}(\mathbf{I}_c, \mathbf{J}_{bd}, \mathbf{M})))] + \mathbb{E}_{\mathbf{I}_{gt} \sim \mathbf{P}_{gt}}[\log(D_{bd}(\mathbf{I}_{gt})]$. Finally, the overall loss function can be concluded as

$$\min_{G_{bd}} \max_{D_{bd}} \mathcal{L}(G_{bd}, D_{bd}) = \lambda^{bd}_{\ell_1} \mathcal{L}_{\ell_1} + \lambda^{bd}_{adv} \mathcal{L}_{adv} + \lambda^{bd}_{perc} \mathcal{L}_{perc} + \lambda^{bd}_{sty} \mathcal{L}_{sty}. \tag{10}$$

4 Experiments

4.1 Implementation Details

In this section, we conduct intensive evaluations in three public datasets, namely Celeba [15], Places2 [26], and Paris [3]. Places2 is the most challenging one with more than 10 million images divided into about 400 different scene categories. For each dataset, we use the same settings for comparable state-of-art methods, namely, Contextual Attention (CA) [25][1], Partial Convolution (PConv) [14], and Edge-Connect(Edge) [18][2].

With regard to the inputs of the base reconstructor, we extract the GT of the base layer $\mathbf{I}_b$ from [23] with default parameters. Then, three combined elements $\mathbf{I_c}$, $\mathbf{I}_{bc}$, M are sent to the base reconstructor. Referring to the detail reconstructor, the detail layer is first obtained by Eq.5 as an approximation. Second, the base reconstructor is fed with $\widehat{\mathbf{I}}_c$, $\mathbf{I}_{dc}$, M. Once obtaining the results of base and detail layer, the base-detail generator begins to work with associated inputs. The visualization of inputs is shown in Fig.2(a).

During the training process, the base and detail layer reconstructor G_b and G_d are trained parallel and independently with inputs until convergence. The inputs are resized to 256×256 and the batch size is 4. Both regular(one box with random position) with hole size 128×128 at random positions and irregular masks [14] are used. The parameters of λ_{ℓ_1}, λ_{adv}, λ_{perc} and λ_{sty} are 4, 1, 0.1, 250, respectively. The Adam [12] is selected with $\lambda_1 = 0.5$, $\lambda_2 = 0.9$ and $learningrate = 10^{-4}$ that is lowered every 100,000 iterations.

Our full model is implemented on pyTorch v1.0, CUDNN v7.0, CUDA v9.0, and run on NVIDIA(R) GPU GTX 1080 Ti. More details can be found in our supplementary materials.

4.2 Comparisons

In the setup above, we quantitatively and qualitatively check out all the methods.

Quantitative comparisons In image inpainting, measuring results is tough since there is a lack of standard metrics. In order to compare the results all-round, two types of metrics are assessed referring to image quality methods: (1) Distortion, and (2) Perceptual quality metrics. Type (1) involves structural similarity index (SSIM) and Peak signal-to-noise ratio (PSNR) assuming that the target and the source images are the same. PSNR is a bottom-up method in pixel-level of which values are higher the better, while SSIM is a top-down method considering the local information based on the perceptual consistency in a human visualization system (HVS) model. Thus, these two are commonly used together to estimate the image quality. Nevertheless, they ignore global distribution awareness to hold the overall perspective. Then, Type (2) helps with perceptual awareness by using the high-level Fréchet Inception Distance (FID) [7] score. Usually, we take the pre-trained inception-V3 model to calculate the score.

The compared results over Places2 are shown in Tab.1. We test the results on available given models of the compared methods CA, Edge, and PConv. Our work without SW has the best performance in Places2, which means the G and D compete well with each other.

Qualitative comparisons Some typical images are exhibited in Fig.3. Our generated images are much more accord with subjective visual perception, while others are not that satisfactory. Note that the larger the missing area, the better our performance is compared with other methods. CA is unstable since it cannot grasp the essential sketch of the image.

[1] https://github.com/JiahuiYu/generative_inpainting
[2] https://github.com/knazeri/edge-connect

Figure 3: Comparison of qualitative results with state-of-art methods. (From top to bottom) GT images, corrupted input images with the mask area increasing from left to right, CA [], Edge [], ours.

Table 1: The performance compared with state-of-art methods on Places2 with irregular (e.g. 10-20% means that the area of masks is up to 10-20% of the whole images) and regular (fixed) masks.

Method	L1 %					PSNR					SSIM					FID				
Mask %	10-20	20-30	30-40	40-50	fixed	10-20	20-30	30-40	40-50	fixed	10-20	20-30	30-40	40-50	fixed	10-20	20-30	30-40	40-50	fixed
CA	2.05	3.52	5.07	6.62	3.69	24.22	21.01	18.99	17.61	20.46	0.893	0.805	0.717	0.627	0.773	6.55	13.59	23.25	34.70	7.81
PConv	1.14	1.97	3.01	4.11	-	28.015	24.895	22.450	20.860	-	0.869	0.778	0.685	0.589	-	-	-	-	-	-
Edge	1.31	2.26	3.25	4.39	-	27.32	24.31	22.33	20.70	-	0.942	0.890	0.830	0.757	-	3.47	5.49	9.04	14.33	-
Ours	**1.02**	**1.76**	**2.59**	**3.55**	**3.06**	**29.71**	**26.34**	**24.03**	**22.21**	**21.85**	**0.964**	**0.927**	**0.880**	**0.821**	**0.801**	**2.80**	**3.74**	**5.50**	**8.97**	**7.74**

Edge seems puzzled and sinks into wrong directions, e.g the wrong position of eyes in Fig.3 row 4. Obviously, our method effectively operates on different image scenes and get well-generated images. More comparisons can be found in our supplementary materials.

4.3 Ablation Studies

In this section, we present two kinds of additional experiments to enrich our demonstration. First, we evaluate the efficiency of both single and combined reconstructors. Then, we present the influences of the two proposed models both with and without SW.

Base-only and Detail-only Ablation For further illustrating, we verify the generated images based on base-only and detail-only reconstructors separately with the base-detail generator. From the numerical values in the Tab.2, we can conclude that the joint of the base and detail reconstructors perform better than just one of them. The base-only or the detail-only could sometimes produce good results, is mainly because the inputs I_c in base-detail generator leave some latent messages to boost, but both of them are unstable seen from FID score. From this point of view, there is a great deal of effort behind slight progress. As the base layer grasp the greater proportion of the global view and with the help of the

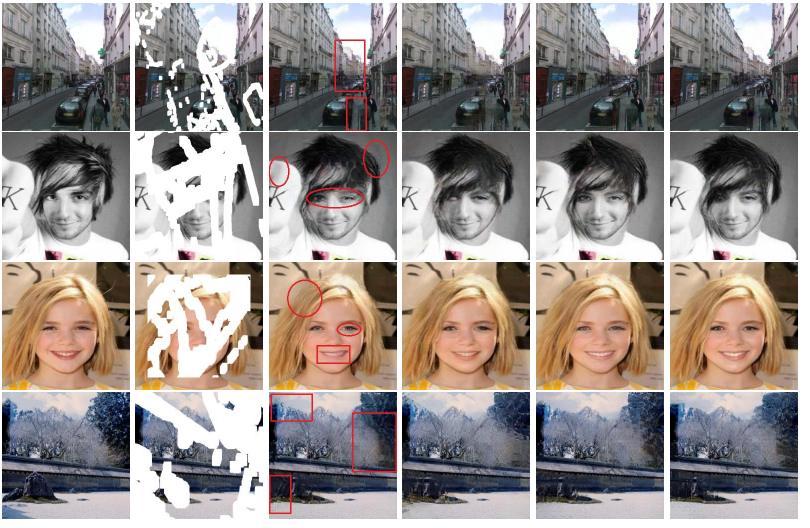


Figure 4: (From left to right) GT images, corrupted images(**Mask area is increasing top-down**), base-only outputs, detail-only outputs, w/o SW outputs, w/ SW outputs of our methods. Best compared with red circle regions.

Table 2: Ablation studies on Celeba with irregular (e.g. 10-20% means the area of masks is up to 10-20% of the whole images) and regular (fixed) masks.

Method	L1 %					PSNR					SSIM					FID				
Mask %	10-20	20-30	30-40	40-50	fixed	10-20	20-30	30-40	40-50	fixed	10-20	20-30	30-40	40-50	fixed	10-20	20-30	30-40	40-50	fixed
only base	0.72	1.2	1.77	2.48	2.47	33.20	29.72	27.12	24.96	23.41	0.981	0.962	0.936	0.901	0.870	0.98	1.43	2.20	3.43	4.20
only detail	0.74	1.25	1.84	2.57	2.43	33.88	29.41	26.84	24.70	23.71	0.958	0.961	0.934	0.897	0.872	0.99	1.45	2.25	3.55	3.09
w/o sw	0.69	1.14	1.69	2.36	**2.04**	33.73	30.25	27.61	25.43	**24.97**	0.982	0.964	0.939	0.905	**0.901**	0.94	1.31	1.83	2.62	**1.82**
w/ sw	**0.61**	**1.05**	**1.59**	**2.26**	2.49	**34.34**	**30.54**	**27.75**	**25.50**	24.68	**0.985**	**0.969**	**0.945**	**0.912**	0.833	**0.86**	**1.16**	**1.63**	**2.40**	1.88

detail layer, the joint generator can polish images in a smaller search space which leads to a clear direction without hesitation. Visible results in Fig.4($3-5^{th}$*columns*) also display the performance, better viewed by overlapping exchange. The inpainted images with base-only reveal the basic skeleton of an image with smoothing and blurry performance, while detail-only brings local sharp details without holding the global distribution of images. The joint performs better than either of them.

With and Without SW Here, we test two cases: (1)without SW, the basic method proposed in this paper with Instance normalization (IN) [21]; (2)with SW, replacing IN with SW in all layers of the base-detail generator. IN can keep the independence between each image instances, which increases the performance of generating image instances. Referring to the base-detail generator, using IN alone may not be enough. Then we select the latest method switchable normalization (SW) to assess. Different from IN, SW is a normalization technique that is able to learn different normalization operations for different normalization layers in a deep neural network, which is self-adaptive and generalized, that is to say, SW picks the reliable normalization for each layer in convolution blocks. Some results are presented in Tab.2. Notably, with SW gives a slight improvement in performance. In a sense, SW plays an important role in generating images in a global view with normal-friendly used in each layer. Visible results in Fig.4($5-6^{th}$ columns) also show the performance. Also, note that different normalization makes a slight difference in style and perception.

5 Conclusion

In this paper, we put forward a new "Divider and Conquer" adversarial model called Base-Detail Image Inpainting, which consists of a reconstructed base and detail layer to generate the final perceptually plausible results. The base reconstructor is designed to grasp the basic distribution of structure and color, while the detail reconstructor gives more local details reflecting the intensity. Meanwhile, we illustrate the base and the detail reconstructors are existing side by side and playing a part together. For different purposes, the different losses are used in each generator. Furthermore, SW is proven to produce better results in some cases compared with IN. Lastly, the results quantitatively and qualitatively verify the effectiveness of our method compared with current state-of-the-art techniques. In future work, we will do further explorations from the three aspects: (1) More reliable subjective and objective measurements, especially the subjective ones, because everyone may have a different subjective perspective. (2) The big hole (the area of hole >50% of the images) image inpainting with effective methods. The big hole gives us much more imagination with little constraints, thus, how to obtain useful constraints from an image becomes challenging. (3) Transform the proposed method to other inpainting tasks, e.g. video inpainting, point-cloud inpainting.

References

[1] C. Barnes, E. Shechtman, et al. Patchmatch: A randomized correspondence algorithm for structural image editing. *Acm Transactions on Graphics*, 28(3):1–11, 2009.

[2] M. Bertalmio, G. Sapiro, et al. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 417–424, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co. ISBN 978-1-58113-208-3. doi: 10/dcvpvb.

[3] S. Doersch, C.and Singh et al. What makes paris look like paris? *Communications of the ACM*, 58(12):103–110, 2015.

[4] S. Esedoglu and J. Shen. Digital inpainting based on the mumford–shah–euler image model. *European Journal of Applied Mathematics*, 13(4):353–370, August 2002. ISSN 1469-4425, 0956-7925. doi: 10/dr5892.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition*, 2016.

[6] I. Goodfellow, J. Pouget-Abadie, et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[7] M. Heusel, H. Ramsauer, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv:1706.08500 [cs, stat]*, June 2017. arXiv: 1706.08500.

[8] J. Huang, S. B. Kang, et al. Image completion using planar structure guidance. *ACM Transactions on Graphics*, 33(4):1–10, July 2014. ISSN 07300301. doi: 10/f6cszm.

[9] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):1–14, July 2017. ISSN 07300301.

[10] P. Isola, J. Y. Zhu, et al. Image-to-image translation with conditional adversarial networks. 2016.

[11] J. Johnson, A. Alahi, and F. F. Li. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980 [cs]*, December 2014. arXiv: 1412.6980.

[13] E H Land. The retinex theory of color vision. *Scientific American*, 237(6):108–128, 1977.

[14] G. Liu, F. A. Reda, et al. Image inpainting for irregular holes using partial convolutions. *arXiv:1804.07723 [cs]*, April 2018. arXiv: 1804.07723.

[15] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[16] P. Luo, J. Ren, et al. Differentiable learning to normalize via switchable normalization. *arXiv:1806.10779 [cs]*, June 2018. arXiv: 1806.10779.

[17] T. Miyato, T. Kataoka, et al. Spectral normalization for generative adversarial networks. *arXiv:1802.05957 [cs, stat]*, February 2018. arXiv: 1802.05957.

[18] K. Nazeri, E. Ng, et al. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv:1901.00212 [cs]*, January 2019. arXiv: 1901.00212.

[19] D. Pathak, P. Krahenbuhl, et al. Context encoders: Feature learning by inpainting. *arXiv:1604.07379 [cs]*, April 2016. arXiv: 1604.07379.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556 [cs]*, September 2014. arXiv: 1409.1556.

[21] D. Ulyanov, V. Lebedev, et al. Texture networks: Feed-forward synthesis of textures and stylized images. 2016.

[22] T. C. Wang, M. Y. Liu, et al. High-resolution image synthesis and semantic manipulation with conditional gans. 2017.

[23] Li X., Qiong Y., et al. Structure extraction from texture via natural variation measure. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2012.

[24] C. Yang, X. Lu, Z. Lin, et al. High-resolution image inpainting using multi-scale neural patch synthesis. *arXiv:1611.09969 [cs]*, November 2016. arXiv: 1611.09969.

[25] J. Yu, Z. Lin, J. Yang, et al. Generative image inpainting with contextual attention. *arXiv:1801.07892 [cs]*, January 2018. arXiv: 1801.07892.

[26] B. Zhou, A. Lapedriza, et al. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.