Object Removing via Multi-Column Gate Convolutional Neutral Network

Haipeng Huang[1, *, +], Haikun Yuan[2, +]

[1]Computer science, Nanjing University of Posts and Telecommunications, 10293, China

[2]Robotic Engineering, Xidian University, 10701, China

* Corresponding Author Email: b20032228@njupt.edu.cn

+These authors contributed equally

Abstract. Object removal is to remove unwanted objects in a photograph. It can be realized by the inpainting technique with irregular mask. Generative Image Inpainting with Contextual Attention [n] and Partial convolution cannot work well with irregular mask，so we present a Multi-column Gate Convolutional Neutral Network. It can work well with both regular and irregular mask. Our model adopts gate convolution which is more suitable in inpainting task than traditional convolution. Special context loss, semantics loss and spatial reconstruction loss are used for better performance. We train the model with CelebA and Landscape. A comparison is made between our model and some models published on CVPR etc, our model's results can generate more realistic images than existing ones for object removal.

Keywords: Gate convolution, Multi-convolution network, Implicit diversified MRF loss.

1. Introduction

Filling missing pixels of an image, often referred to as image inpainting or completion, is an important task in computer vision. Exemplar-based measure [2,7,8,9,10] and CNN-based measure are two of the methods applied commonly in inpainting tasks. The former is effective in repairing cracks in small regions. But when the area of the covered region is bigger than 10%, the result is unsatisfying. So, in large-scale inpainting tasks, we use CNN-based measures to estimate the covered pixels.

In recent years, generative networks based on CNN are becoming the predominant approaches to inpainting tasks. Context encoder [3] points out the deficiency of the Exemplar-based [2] method, novelty proposed a Generative Adversarial network and developed an improved model of the encoder-decoder network. The results are more compelling compared with the Exemplar-based method. But the methods referred to above don't perform well in inpainting irregular strokes.

Instead of vanilla convolution, Liu G al et. [4] exploited partial convolution to fix the image with irregular holes. If there are invalid pixels in the previous layer, the results of partial convolution are transferred and normalized, meanwhile, the mask is updated. Irregular patches distortion and blur can be overcome to a certain degree by using partial convolution. However, partial convolution can be viewed as unlearnable hard-gating. So, gate convolution is proposed to improve the flexibility and accuracy of inpainting. The unlearnable hard gate is replaced by the learnable soft gate which can be updated automatically from data.

According to the above, we customize a GAN-CNN network with gate convolution. With the multi-column encoder, it can learn a dynamic feature selection mechanism that can represent the feature in all locations, and improve the ability to fix irregular masks. Specifically, we used gate convolution in our model instead of vanilla convolution in the encoder part. With the implementation of the implicit Markov random fields, our model could also produce fast and high-quality textures. Thus, our model could generate higher-quality free-form inpainting images compared with previous models. The contributions of this works are as follows:

Use multi-column structure which include three different receptive fields and special resolutions instead of single column structure. This can extract different levels of information.

We apply both gate convolution and partial convolution to our model and we found that gate convolution could generate more visual-compelling results.

Gate convolution is applied into the network which is more suitable in inpainting irregular mask.

Use a new loss function called MRF-loss to make the texture and generated part more various, and some subtle details can also be restored.

2. Related work

2.1. Global and Local discriminator

Since the GAN network is used in inpainting tasks, global and local discriminator structures [5] can more realistically generate the image than one discriminator. Unlike patch-based approaches, it can also fix the image with content that does not appear elsewhere in the image.

2.2. Partial convolution

Because most of the masks in inpainting tasks are irregular, Partial convolution [4] is a novel method that distinguishes pixels between valid and invalid. The invalid pixels which are covered by masks are not computed in the convolution layer. Taking all pixels into consideration will cause blurs if the mask is irregular.

2.3. Coarse-to-fine framework

The classic GAN network tends to yield semantic information and texture with a smooth structure. The coarse-to-fine framework [3] proposed by Yu et al. [3] is a contextual attention module. The fine-level learns the features which are about explicitly matching and attending to relevant background patches. Unlike the common structure which can learn merely low-level information, this framework can produce a semantic structure with diverse textures.

3. Method

Our model consists of two parts: Generator parts and discriminator parts. Generator parts are used to produce the patch. The generator includes 3 parallel branches encoder-decoder for image inpainting. The input of the model takes an image X and a binary mask $\boldsymbol{M}$ (0 represents a valid pixel and 1 represents an invalid pixel). There are 4 steps after the images are fed into the network:

The image is sent to 3 branches to produce results respectively and then the results are concentrated to get the feature map with $\boldsymbol{F}$ with the same resolution.

The shared decoder module transforms the feature map $\boldsymbol{F}$ into a natural image $\widehat{\mathbf{Y}}$.

In the training phase the output $\widehat{\mathbf{Y}}$ is sent to 2 discriminators. The local discriminator judges the authenticity of patches and the global discriminator judges the authenticity of the whole image.

Then the ground truth image and the output image $\widehat{\mathbf{Y}}$ are sent to a pretrained VGG-19 network to calculate the MRF loss.

Only the Generator part is used in the testing phase, Discriminator and VGG-19 are only used in the training phase.

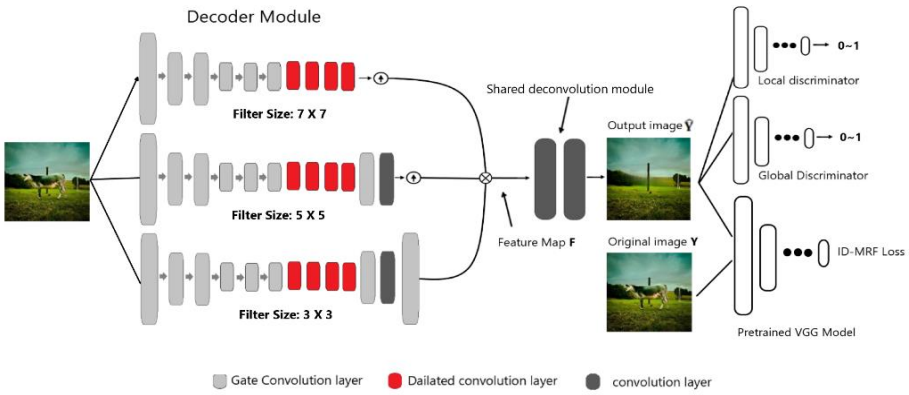


Figure 1. network structure

3.1. Multi-column encoder-decoder structure:

Inpainting involves different levels of representations, so we use 3 parallel encoder-decoder branches [1], the filter sizes of which are 3×3, $5 \times 5, 7 \times 7$ to extract features from the input image. With multi-column encoder-decoder structure, the network can draw features with various respective fields and spatial resolutions. Our structure overcomes the issue that a single encoder-decoder structure may ignore too large or too small semantics information in the input image. Our structure can also overcome the limitation that coarse-to-fine structure has: the errors that occur at the coarse level will influence the refinement. With parallel branches and concentrated parts, the information will complement each other instead of simply inheriting information.

3.2. Implicit diverse Markov Random Fields (ID-MRF)

Reliable Similar Patches Searching can make the model generate more realistic details [1]. However explicit nearest-neighbor search is computationally-heavy and it may harm the process when missing areas originally contain different contexts. So, our network uses implicit diversified Markov random fields instead, which are only taken in training phrases. We use MRF loss to minimize the difference between generated content and corresponding nearest neighbours from ground truth in the feature space to get a more realistic texture. If we use a similarity measure to find the nearest patches, it will cause one patch to appear repeatedly in the resulting image. To avoid this issue, we propose an optimized MRF loss in our network. We use the relative distance measure between the local feature and the target feature set.

Two images are sent to the MRF network (VGG19), one is the generated image, and the other is the ground truth image. The relative similarity is a formula to measure the similarity between generated neural patches and grand truth's neural patches. $\boldsymbol{v}$ and $\boldsymbol{s}$ are the patches from $\widehat{\boldsymbol{Y}}_{\boldsymbol{g}}$ and $\boldsymbol{Y^L}$. In our model, we extract $\boldsymbol{v}$ and $\boldsymbol{s}$ from $conv4\_2$ and $conv3\_2$. And the relative similarity formula is defined as:

$$RS(v,s) = \exp\left[\left(\frac{\mu(v,s)}{max_{r\in\rho_v(\boldsymbol{Y^L})}\mu(v,r)+\varepsilon}\right)/h\right] \tag{1}$$

Where $\mu(:,:)$ is the cosine similarity. $\boldsymbol{r} \in \rho_v(\boldsymbol{Y^L})$ means $\boldsymbol{r}$ belongs to $\boldsymbol{Y^L}$ excluding v. h and ε are two positive constants. The bigger of the $RS(v,s)$ is, the more similar v and s is.

$\overline{RS}(v,s)$ means the average similarity of all the patches.

$$\overline{RS}(v,s) = RS(v,s)/\sum_{r\in\rho_v(\boldsymbol{Y^L})} RS(v,r) \tag{2}$$

Finally, the MRF loss is defined as

$$\mathcal{L}_M(L) = -\log\left(\frac{1}{Z}\sum_{s\in \boldsymbol{Y^L}}\ \sum_{v\in \hat{\boldsymbol{Y}}_{\boldsymbol{g}}} max\,\overline{RS}(v,s)\right) \tag{3}$$

The MRF loss is computed on several layers in VGG-19. We use $Conv\ 3\_2$ and $Conv\ 4\_2$ to calculate MRF loss as

$$\mathcal{L}_{mrf} = 2\mathcal{L}_M(conv\ 4\_2) + \mathcal{L}_M(conv\ 3_2) \tag{4}$$

This algorithm can solve the problem like using the same pattern to fix the image. This will cause the generated part is lack of flexibility and variety, which will make the generated part unnatural. But using the relative similarity can make the generated content more flexible.

3.3. The optimization of the reconstruction loss

Reconstruction loss compares the differences between every pixel in the input and output image. However, directly using Reconstruction loss doesn't reasonable in the inpainting task because the pixels near the boundary are more strongly constrained compared to those pixels in the center. For this reason, we add a weight to each pixel to balance the spatial difference. We set the known pixels' weight as 1 and the unknown pixels' weight as a value related to the distance to the boundary. We propagate the value of weight by Gaussian filter g whose size is 64x64 and the standard deviation is 40. We generate $\bar{M}$ to create a weight mask M_ω as

$$M_w^i = \left(g * \bar{M}^i\right) \odot M \tag{5}$$

Where $\bar{M}^i = 1 - M + M_w^{i-1}$ and $M_w^0 = 0$ and $\odot$ is Hadamard product. The final reconstruction loss is defined as

$$\mathcal{L}_c = \left|\left|\left(Y - G([X,M],\theta)\right) \odot M\right|\right|_1 \tag{6}$$

We exploit the loss function that combines the pixel with spatial location by considering both known and unknown pixels.

3.4. Gate convolution

The vanilla convolution is suitable for object classification and detection because all the input pixels are valid. But in image inpainting, the input contains both valid and invalid pixels. So partial convolution [5] novelly propose a method that classifies all the spatial pixels into valid and invalid (1&0), and the outputs are only derived from the valid pixels.

$$x' = \begin{cases} W^T(X \odot M)\dfrac{sum(1)}{sum(M)} + b, (if\ sum(M) > 0) \\ 0 \qquad\qquad\qquad\qquad , (otherwise) \end{cases} \tag{7}$$

$$m' = \begin{cases} 1 & , if\ sum(M) > 0 \\ 0 & , otherwise \end{cases} \tag{8}$$

The equation shows that the all the invalid pixels will be replaced by valid pixels after several times convolution, so that Gate convolution could improve the quality of inpainting on irregular mask.

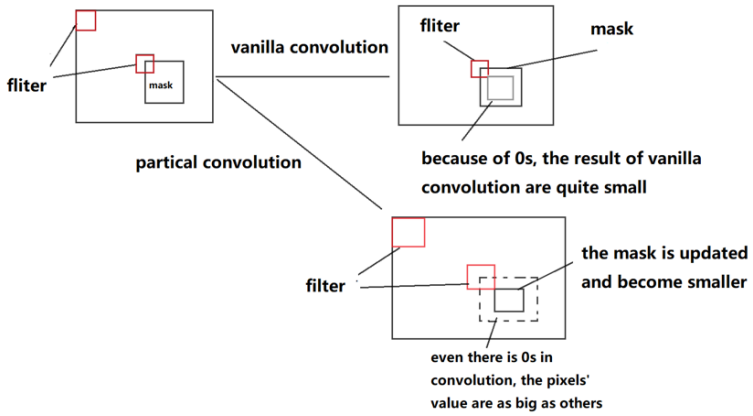


Figure 2. Visualization of Partial convolution

But there are some remaining issues [6]: (1) All the pixels are classified into 0 and 1. (2) The number of valid pixels in the previous layer is unknown. (3) All the channels using the same mask lack of flexibility. So, the partial convolution is called "unlearnable single channel hard-gating".

The gate convolution can solve the above questions by introducing an additional "Gating matrix"m which is transferred from feature matrix x by convolution. The gate convolution chooses the most suitable gate for the current feature matrix x. So, it is also called "learnable multi-channel feature soft-gating". Compared with partial convolution, the soft gate is more flexible and it can value every spatial location even in deep layers.

$$m_{x,y} = \sum\sum x_g \cdot I(fliter) \tag{9}$$

$$r_{x,y} = \sum\sum x_f \cdot I(fliter) \tag{10}$$

$$O_{x,y} = m_{x,y} \odot r_{x,y} \tag{11}$$

Where $m_{x,y}$ is the elements in soft gate matrix, $r_{x,y}$ means the results of vanilla convolution, and $O_{x,y}$ is the output of the gate convolution.

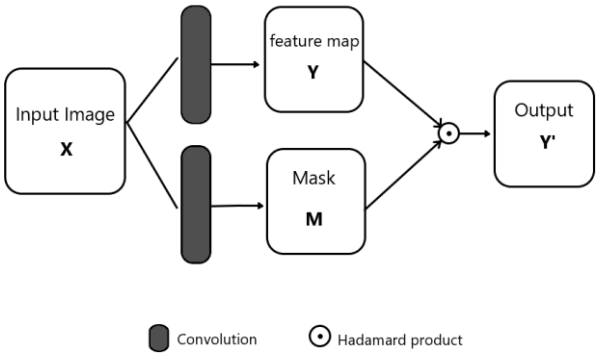


Figure 3. The purpose of Gate convolution: the input image is sent to two convolution layer and get the result respectively, the output gating values will be set between zeros and ones.

4. Experiments and Results

4.1. Datasets

We use two datasets for training and testing. One uses 4000 pictures chosen randomly from celeba-256 and the other uses pictures from Landscape download from Kreas. The size of input images is (128×128), and the mask is random strokes or random rectangles (128×128). The two types of masks are trained separately. We train our model with reconstruction loss for 40 times. After our

models converges (40 epochs), we add MRF-loss and adversarial loss for finetune. The learning rate is 1e-4 and the number of parameters is 21.925 million.

4.2. Implementation Details

Our implementation is with CUDA4.13.0 and PyTorch 3.9. The hardware is NVIDIA RTX2060 and CPUi7 3.6GHz, the batch size is 6, and the learning rate is 1e-4. It takes 19.31s to run 5 batches, 643ms per picture (training with MRF-loss and adversarial-loss).

4.3. Qualitative evaluation

As shown in figure n, comparing with GMCNN [1], Global & Local discriminator (Global&Local), Context Attention (DeepFillV1). It is obvious that our model performs better in stroke inpainting and can solve the problem of blur and distortion. Table 2 is comparison of PSNR&SSM of 4 methods. Although the inpainting task is not suitable to be evaluated by the peak signal-to-noise ratio (PSNR) or structural similarity (SSIM), we still give them for reference.

Testing the model in the scenery dataset, Figure n is the comparison of four methods to move the same target. It is obvious that if the target is big, the global and local discriminator and deepfillV1 can't repair it very well. The middle texture is vague.

Table 1. Comparison of different model

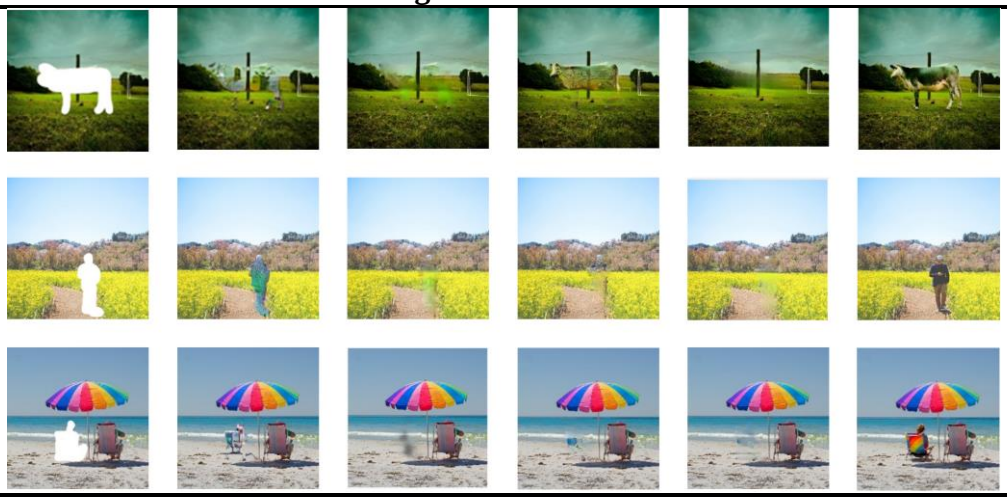
Input picture	**Global&Local**	**DeepFillV1**	**partial convolution**	**ours**	**ground truth**
					

Table 2. Qualitative evaluation of four methods

Model	**Global&Local**	**DeepfillV1**	**Partial conv**	**Ours**
PSNR	24.65	23.78	25.70	23.26
SSIM	0.8650	0.8580	0.8476	0.8739

4.4. Amblition study

We train the vanilla model and gate convolution model separately, and the model's structure and parameters are the same. Then we train them with the same dataset until converge. Table 3 is the comparison of the results. According to the figure, we find that our model can restore more details than partial convolution and our model can inpaint the image more naturally while the partial convolution model cannot restore the boundary smoothly.

Table 3. Comparison of Gate convolution and partial convolution, our model could generate more compelling results compared with previous model.

Input	Partial conv	Gate convolution	Ground Truth
			
			
			

5. Conclusion

This paper uses multi-column branches instead of single branches to extract various spatial features. we apply gate convolution to the model to improve the patches' visual authenticity on irregular inpainting tasks compared to directly using vanilla convolution. We also optimize the reconstruction loss and MRF loss in our model. Our model obtains more realistic and natural results compared to previous measures on irregular mask inpainting.

Limitations: this model was only trained with 30,000 images. We will train our model with more images in the future. Our model can't restore the objects whose most of the body is covered (for example, the body of the dog is covered but its head is not covered). In the feature, Our plan is to add edge restoration to our model to deal with this issue.

References

[1] Yi Wang Xin Tao Xiaojuan Qi Xiaoyong Shen Jiaya Jia: Image Inpainting via Generative Multi-column Convolutional Neural Networks In: arXiv:1810.08771v1 [cs.CV] 20 Oct 2018

[2] K. He and J. Sun. Image completion approaches using the statistics of similar patches. TPAMI, 36(12): 2423–2435, 2014.

[3] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S. Huang: Generative Image Inpainting with Contextual Attention In: arXiv:1801.07892v2 [cs.CV] 21 Mar 2018

[4] Guilin Liu Fitsum A. Reda Kevin J. Shih Ting-Chun Wang Andrew Tao Bryan Catanzaro: Image Inpainting for Irregular Holes Using Partial Convolutions In: arXiv:1804.07723v2 [cs.CV] 15 Dec 2018

[5] SATOSHI IIZUKA, EDGAR SIMO-SERRA, HIROSHI ISHIKAWA: Globally and Locally Consistent Image Completion in ACM Transactions on Graphics, Vol. 36, No. 4, Article 107. Publication date: July 2017

[6] Jiahui Yu Zhe Lin Jimei Yang Xiaohui Shen Xin Lu Thomas Huang: Free-Form Image Inpainting with Gated Convolution In: arXiv:1806.03589v2 [cs.CV] 22 Oct 2019

[7] K. He and J. Sun. Statistics of patch offsets for image completion. In ECCV, pages 16–29. Springer, 2012

[8] J. Jia and C.-K. Tang. Inference of segmented color and texture description by tensor voting. TPAMI, 26(6): 771–786, 2004.

[9] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum. Image completion with structure propagation. In TOG, volume 24, pages 861–868. ACM, 2005.

[10] K. He and J. Sun. Image completion approaches using the statistics of similar patches. TPAMI, 36(12): 2423–2435, 2014.