Image Completion Based on Gans with a New Loss Function

To cite this article: Ying Huang et al 2019 J. Phys.: Conf. Ser. 1229 012030

View the article online for updates and enhancements.

You may also like

- Radiation dose efficiency of multi-energy photon-counting-detector CT for dual-contrast imaging
  Liqiang Ren, Kishore Rajendran, Cynthia H McCollough et al.

- Task-driven optimization of CT tube current modulation and regularization in model-based iterative reconstruction
  Grace J Gang, Jeffrey H Siewerdsen and J Webster Stayman

- Task-driven optimization of the non-spectral mode of photon counting CT for intracranial hemorrhage assessment
  Xu Ji, Ran Zhang, Guang-Hong Chen et al.

Recent citations

- Jireh Jam et al
Image Completion Based on Gans with a New Loss Function

Ying Huang1,2, Maorui Wang1,2, Ying Qian1, Shuohao Lin1 and Xiaohan Yang1
1 School of Software Engineering, Chongqing University of Posts and Telecommunications, China
2 Chongqing Software Quality Assurance System Engineering Technology Research Center, China
huangying@cqupt.edu.cn

Abstract. Recently, many approaches based on deep learning have demonstrated amazing capabilities in varieties of challenging image tasks, such as image classification, object detection, semantic and instance segmentation, and so on. These methods are more capable of extracting deeper features than traditional methods, which is critical for different kinds of image tasks. Similarly, these methods are gradually applied to the work of image completion of natural images. In consideration of the fact that most of the current methods would lead to blurring and fake results, we propose an image completion method based on the generation adversarial network. We use the network structure of the encoder and decoder to obtain the high-level feature information of the image and generate reasonable pixel values to fill the missing regions. Besides, we construct a new joint loss function based on SSIM evaluation indicators, which can retain the similarity between two images as much as possible. Our proposed method can keep the completion regions consistent with the surrounding pixels, which makes the images look more realistic. We evaluate on our datasets with our proposed method and compare with other methods in this paper, our results are sharper and more realistic than the previous ones.

1. Introduction
With the achieved remarkable effects of deep learning in recent years, the frontiers of visual field research have been totally occupied by deep learning. More and more researchers have begun to focus on the deep learning field. In the field of visual, the study is being researched around deep learning, and the effectiveness of these methods have been proved. These successes depend on a large amount of manually labeled data to some extent, but in practical application, having such a large amount of labeled data may not be easy, we need new methods from less labeled data and learn the specific features. Therefore, we need to design a more efficient network model to extract the deep features of the image better to complete our target tasks.

In reality, images are inevitably occluded, and we can see this phenomenon in our daily life. We cannot directly observe the missing part of the image, so we need to fix the missing part. For our human beings, we can rely on the information around the image and our understanding to artificially repair the image, even if we have never seen the original image, we can repair the image due to the natural image has a high degree of structural texture and naturalness. However, because of the inherent nature and complexity of the image, the general image restoration methods may result in blurred and unrealistic effects. At present, this is still a challenging task.

In order to overcome the above problems, we propose a method based on the generation adversarial
network [1,2] to predict the missing information in the image. Our model includes an encoder to obtain a potential feature representation and a decoder to generate reasonable pixel to fill the missing regions. The entire network structure uses U-Net [3] as the backbone, which adds a skip-connection between encoder and decoder. It is helpful for decoder to take advantage of potential features to recover missing regions. Our main contributions are as follows:

i) We use the conventional loss function based on SSIM and L1 combined with the adversarial loss function to improve the structural texture and authenticity of the repaired image.

ii) We consider the edge information and the details of the image, add padding information in the network to avoid missing edge information, and replace the pooling operations with convolutional operations with strides.

iii) We use the mini-batch discriminator to optimize the training to increase the diversity of generated sample.

2. Related work
With the rapidly explosive growth of big data and the improvement of hardware computing power in recent years, deep models have performed well in various tasks. At the same time, computer vision tasks such as image classification and object detection [4,5], as well as semantic and instance segmentation [6,7] etc. These models can be used to extract deeper feature to handle complex tasks. We take advantage of these excellent features to complete image restoration work. The following is a brief introduction to the recent work related to this article.

2.1. Deep neural networks
The first neuron model was proposed by McCulloch and Pitts in 1943. Until the concept of deep neural network has been proposed in 2006, the neural network has regained its attention in academia and industry, and in 2012 the network proposed by Alex [8] opened up the explosive growth of deep learning. At present, neural networks and deep learning provide the best solutions for many problems in the fields of image recognition, speech recognition and natural language processing. In recent years, continuous improvement of deep neural networks has also improved the difficulty of training for deep networks, and put forward some training techniques, such as the operation of Batch Normalization [9] and Dropout [10], so that deeper networks can also be trained.

2.2. Generation adversarial networks
Deep models are mainly divided into generation models and discriminant models. Discriminant models such as classification have made great progress in recent years. However, until the generation adversarial network was proposed by Goodfellow in 2014, which brings the generation models a new life. The network consists of a generator and a discriminator. The generator learns to capture the potential distribution of real data samples and uses noise as input to get the data samples. The discriminator acts as a two-classifier, which is mainly used to distinguish the samples obtained from the real data samples. The whole optimization process is a problem of minimax game. The optimization goal is to achieve Nash equilibrium, the discriminator cannot distinguish the samples generated by the generator, that is, the generator can estimate the true distribution of the data samples.

2.3. Image completion
The traditional image restoration methods are mainly based on local and non-local information. Most existing methods are designed to fix a single image. For example, based on the method of total variation [11], the smooth continuity of images is taken into consideration, which is effective for removing small holes. The PatchMatch [12] was one of the most popular repair methods at that time due to high quality and efficiency, which searched for similar patches in the image. Further improvement is to find the image with the largest similarity in the entire training set to repair the missing area [13], however, it is necessary to include correct information in the input image for all individual image. Recently, the research based on CNN image restoration has become popular. For
example, Context Encoder [14] adopts the simplest overall content constraint, that is, the L2 distance between the prediction map and the original image. In contrast, the main idea of On Demand Learning [15] is to exploit the feedback mechanism to self-generate training examples where they are needed. With a deeper understanding of CNN and GAN, we can use the feature representation to generate some useful information to fill missing areas. The use of GAN to generate transformed images in recent years [16, 17, 18] demonstrates its potential for capturing the deeper features and generating data. Similarly, the effectiveness of image filling has been demonstrated in our research experiments.

3. The methods

In order to fill the missing areas in the image, we need the GANs to generate some pixels to achieve the image completion. Compared with the conventional image inpainting methods, the pixel value generated in the repaired region may not be derived from the image. The traditional generation adversarial network was shown in Figure 1. The generator includes an encoder and a decoder, the encoder takes the image with the missing area and some noises as input and learns the latent feature representation, and then reconstructs the learned latent features to restore the original image through the decoder. Finally, the reconstructed image and the target image are sent to the discriminator. By continuously optimizing the generator and the discriminator, the discriminator cannot distinguish the generated image from the target image, that is, the repaired image is sharper and more realistic.

3.1. Network structures

The overall structure of this paper is an encoder-decoder pipeline [19]. The encoder gradually reduces the spatial dimensions of the image to extract the deeper features, and the decoder gradually repairs the details and spatial dimensions to reconstruct the repaired image. There is a skip-connection between the encoder and decoder, which helps the decoder to better fix the details. To make the network train more efficiently, there are no fully connected layers in the structure between encoder and decoder (see Figure 2). In the deconvolution of the expanding path, each step will add a feature map from the corresponding shrink path. In our experiments, we replace all the pooling operations into convolution operations to extract the deeper and more potential feature representations. In order to maintain the image’s edges, we use padding to fill the edge information in the convolution process, which is very important to maintain the edge texture information around the missing area.

![Figure 1. The overall generation adversarial network structure.](image1)

![Figure 2. The Network structure uses the encoder and decoder, add the skip connection to combine the feature information into the decoder part.](image2)
3.2. Loss function

The innovation of the loss function of the neural network in the computer vision field is relatively unsatisfactory compared to the various kinds of network architectures. Then most models still use the L2 loss function [20], and it is undeniable that the L2 loss function performs well on some tasks, however this may be disappointed for others. The L2 loss function tries to minimize the Euclidean distance between the output and the corresponding data, but tends to average multiple patterns in the prediction results, which thus leading to the ambiguous results. We explore with L1 loss which produces sharper and more realistic images. The L1 loss function is defined as:

\[
L_{l1}(G) = E_{x,z} \left[ \|y - G(x,z)\|_1 \right]
\]

(1)

The learning process of the generation adversarial network is to train a generator G and a discriminator D at the same time. The generator network is continuously learned until the distribution of input noise is close to the probability distribution of real data; on the other hand, the discriminator is able to distinguish whether an image comes from training data or a generator. The whole learning process is like a two-player game, in which discriminator takes the predicted images from generator and real images and tries to distinguish them, while generator tries to confuse discriminator by producing as realistic samples as much as possible. The objective for conditional GANs can be formulated as follows:

\[
L_{GAN}(G,D) = E_{x,y} \left[ \log D(x,y) \right] + E_{x,z} \left[ \log(1 - D(x,G(x,z))) \right]
\]

(2)

Among the traditional image evaluation indicators, there are similar indicators such as PSNR [21] and SSIM [22] to judge the relationship between the two images, respectively. Here, in order to take human visual perception into consideration, we set the loss function based on SSIM from the perspective of image composition. The human vision system, which combines human subjective perception, is sensitive to structural information and insensitive to high-luminance and texture-contrast regional distortion. The calculation formula of SSIM is:

\[
\text{SSIM}(p) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \times \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}
\]

(3)

Where \(\sigma_x, \sigma_y\) are the mean values of the image, \(\sigma_x, \sigma_y\) are the standard deviation, \(\sigma_{xy}\) is the covariance, and \(C_1, C_2\) represents a small constant.

Accordingly, the SSIM-based loss function is defined as follow:

\[
L_{SSIM}(p) = \frac{1}{N} \sum_{p \in p} 1 - \text{SSIM}(p)
\]

(4)

The previous methods have found that it would perform better for combining the generation adversarial network loss function with a more traditional loss function. The task of the generator is not only to deceive the discriminator, but also to get the output as close to real data as possible. In order to make up for the shortcomings that L1 loss cannot measure the structural similarity, we combined with the adversarial loss function, the L1 loss function and the SSIM-based loss function, we get the final objective loss function, which is defined as:

\[
L_{\text{total}} = \arg \min_{G} \max_{D} L_{GAN}(G,D) + \alpha L_{l1}(G) + \beta L_{SSIM}
\]

(5)

3.3. Implementation details

We used a total of 2015 real street images, all images are resized to 256x256, and an irregular size mask is randomly generated. Our goal is to fill these missing areas, make it look close to the real image. In the generator, we use a U-Net based structure which changes the pooling to a convolution operation. To maintain image edge information, we set the convolution's padding as same attribute. Similarly, we have ignored the fully connected layer and channel-wise fully connected layer to connect the feature, so large number of parameters are reduced. Besides, we replace all the ReLU
activation functions with Leaky LeRU, and add dropout operations in the high layers of the decoder to filter the other useless features. In the discriminator, we use the mini batch method to let the discriminator judge the current sample to be true or false, while paying attention to other samples. About parameter setting in this paper, we set all the weight initialization methods to obey the random positive distribution with a mean value equals 0 and a variance value equals 0.02. We set the L1 loss weight in the joint loss function to 50, the SSIM loss weight to 0.8, and the adversarial loss weight to 1.0. On the NVIDIA GeForce GTX 1070Ti GPU machine, we set the batch size to 4 which have better results according to our hardware conditions, and the training process is optimized using the stochastic gradient descent solver Adam.

4. Experiments and analysis
This section mainly evaluates the performance of our method on the dataset and analyzes it based on the experimental results, indicating the superiority of the proposed method. First we introduce our dataset and compare it with other methods to prove that the image completion results are better in our proposed method.

4.1. Dataset
We totally collected 2015 real streetscape road images, which includes in a variety of situations, such as sunny, foggy and rainy days, as well as day and night, we collect scenarios as many as possible in real scenes, which proves the robustness and stability of the proposed method. Similarly, our datasets are not same with the public face datasets, because the human face dataset has similar structural information, and our streetscape roads are randomly variable, which puts more demands on the learning ability of the network. We first resize the collected image dataset to a fixed size of 256×256, and then randomly generate some mask areas of unfixed size in the main view area of the image, and finally randomly shuffles the total datasets. After that, 80% was selected as the training set, and the remaining 20% was used as the test set. We try to make the data distribution of the training are similar with the test sets, so that the model can handle with many image filling tasks in various situations.

4.2. Evaluation
First, we compare our approach to the best results achieved by the adversarial loss function in the Context Encoder. Next, we compare the results in the improved version of the On Demand Learning method. Our comparison results demonstrate the effectiveness of the proposed structural optimization and loss function optimization. These two image restoration methods work well for the central square block, but cannot be well fixed which produces a satisfactory result when the size is not fixed. Figure 3 shows the comparison between the proposed algorithm and the previously proposed methods. From the subjective analysis, our proposed algorithm performs better in terms of detail consistency and content authenticity, In addition, we compare the contrast experiments for increasing the SSIM loss function separately. We can see the repair effect on the missing region boundary in detail, which reduce the boundary block effect. We magnify the detailed repaired part to analyze (see Figure 4). The method proposed in this paper is better in terms of detail, and reduces the block effects. At the same time, we also judge the completion results from qualitative analysis and visual effects. We use three quality evaluation indicators: MSE, SSIM and PSNR values for analysis. We first calculated the results of all test images, and then randomly selected 10 images from the test results for specific analysis. It contains the repaired results in five weather conditions, each contains two images, corresponding to daytime, nighttime, foggy, rainy and sunny days. The experimental results of all test images are shown in Table 1. The statistical results alone are shown in Figure 5, Figure 6, and Figure 7. We can see from the results that the method proposed in this paper performs best in foggy and sunny days, which is superior to other methods, and other cases are also better, but there is a slight deficiency in rainy days. For the visual effect analysis, the proposed method for repairing the effect is better than the other two methods in the edge texture. The repaired area and the surrounding area show stronger continuous consistency, so the effect of the completion is sharper and closer to the real image.
Figure 3. The method proposed in this paper is compared with the method proposed before. (a) Image, (b) Context Encoder, (c) On Demand Learning, (d) L1+GANs, (e) ours and (f) ground truth.

Figure 4. It can be observed from the details that the method proposed in this paper is better in terms of detail consistency and content authenticity. (a) Context Encoder, (b) On Demand Learning, (c) L1+GANs, (d) ours.

We can see from the magnified repair area that the proposed method performs better than the previous algorithm in terms of detail consistency and authenticity, indicating that the proposed method is really better.
Table 1. The comparison of different algorithms for test results.

| Evaluation index | Context Encoder | On Demand Learning | L1+GANs | Our Algorithm |
|------------------|-----------------|--------------------|---------|----------------|
| SSIM             | 0.95            | 0.95               | 0.96    | **0.97**       |
| PSNR             | 37.8            | 37.36              | 39.04   | **39.63**      |
| MSE              | 11.02           | 11.78              | 9.15    | **8.99**       |

Figure 5. Select 10 images for analysis, the comparison index of the evaluation index SSIM.

Figure 6. Select 10 images for analysis. The comparison results of the evaluation index PNSR are shown.

Figure 7. Select 10 images for analysis. The figure is the comparison index of the evaluation index MSE.
From the perspective of line graph analysis, the SSIM and PSNR values of the repaired image and the target image which the proposed algorithm obtained are mostly better than the previously proposed method, and the MSE value is lower than the previous value, indicating the effectiveness of our proposed method.

5. Conclusion
We propose a generation adversarial network method based on context structure, and combine the new SSIM-based loss function to complete the training optimization of the network. Finally, the pixel values of the missing regions are synthesized to complete the image filling task, and which achieves a good repair effect.

Our model is able to repair images which produces sharper and more realistic results. Compared with the current methods of Context Encoder and deep generation models, the experimental results show that we perform better on edge and texture consistency, and generate more reasonable pixel values in the missing area. And our proposed model may also show great potential in other aspects such as super-resolution and denoising. As future work, we plan to design a sub-task of the discriminator to focus on the repair of the local missing area, and we will also design a better bottleneck networks in our algorithms to improve the training speed and training effects.

Acknowledgments
The authors wish to thank the associate editors and anonymous reviewers for their valuable comments and suggestions on this paper.

References
[1] Goodfellow I J, Pougetabadie J, Mirza M, et al. Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems, 2014, 3:2672-2680.
[2] Mirza M, Osindero S. Conditional Generative Adversarial Nets[J]. Computer Science, 2014:2672-2680.
[3] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. 2015, 9351:234-241.
[4] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks[J]. 2017.
[5] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// International Conference on Neural Information Processing Systems. MIT Press, 2015:91-99.
[6] Li Y, Qi H, Dai J, et al. Fully Convolutional Instance-aware Semantic Segmentation[J]. 2016:4438-4446.
[7] Noh H, Hong S, Han B. Learning Deconvolution Network for Semantic Segmentation[C]// IEEE International Conference on Computer Vision. IEEE Computer Society, 2015:1520-1528.
[8] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.
[9] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. 2015:448-456.
[10] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.
[11] Chan T F, Shen J. Mathematical Models for Local Nontexture Inpaintings[J]. Siam Journal on Applied Mathematics, 2001, 62(3):1019-1043.
[12] Barnes C, Shechtman E, Finkelstein A, et al. PatchMatch:a randomized correspondence algorithm for structural image editing[J]. Acm Transactions on Graphics, 2009, 28(3):1-11.
[13] Bertalmio M, Sapiro G, Caselles V, et al. Image inpainting[C]// SIGGRAPH ’00 : Proceedings of the, Conference on Computer Graphics and Interactive Techniques, New York, Ny, USA. ACM Press/Addison-Wesley Publishing Co. 2000:417-424.
[14] Pathak D, Krahenbuhl P, Donahue J, et al. Context Encoders: Feature Learning by Inpainting[Cl]. Computer Vision and Pattern Recognition. IEEE, 2016:2536-2544.

[15] Gao R, Grauman K. On-demand Learning for Deep Image Restoration[C]. IEEE International Conference on Computer Vision. IEEE Computer Society, 2017:1095-1104.

[16] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks[J]. 2017:2242-2251.

[17] Isola P, Zhu J Y, Zhou T, et al. Image-to-Image Translation with Conditional Adversarial Networks[J]. 2016:5967-5976.

[18] Choi Y, Choi M, Kim M, et al. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation[J]. 2017.

[19] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks[J]. 2014, 4:3104-3112.

[20] Larsen A B, Sønderby S K, Larochelle H, et al. Autoencoding beyond pixels using a learned similarity metric[J]. 2015:1558-1566.

[21] Bjontegaard G. Calculation of Average PSNR Differences between RD-curves[J]. 2001.

[22] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity.[J]. IEEE Trans Image Process, 2004, 13(4):600-612.