SERBIAN JOURNAL OF ELECTRICAL ENGINEERING
Vol. 14, No. 3, October 2017, 379-386

UDC: 004.932.4+004.934.72 **DOI: https://doi.org/10.2298/SJEE1703379H**

Combination of Single Image Super Resolution and Digital Inpainting Algorithms Based on GANs for Robust Image Completion

Sparik Hayrapetyan[1], Gevorg Karapetyan[2], Viacheslav Voronin[3], Hakob Sarukhanyan[2]

Abstract: Image inpainting, a technique of completing missing or corrupted image regions in undetected form, is an open problem in digital image processing. Inpainting of large regions using Deep Convolutional Generative Adversarial Nets (DCGAN) is a new and powerful approach. In described approaches the size of generated image and size of input image should be the same. In this paper we propose a new method where the size of input image with corrupted region can be up to 4 times larger than generated image.

Keywords: Inpainting, Deep learning, Super-resolution.

1 Introduction

There are different applications of image inpainting such as restoration corrupted photographs, text and watermark removal. One of the challenging applications is filling large missing or corrupted areas in digital image, where the missing region contains elements that cannot be reconstructed via information available in single image. For such problems popular patch based image completion methods such as exemplar based on inpaiting cannot work [9, 10].

In previous work [7], researchers have considered filling large missing large regions with usage of GANs, where the input image size is the same as generated one.

However, we found the approach of generating the same sized image at computation expense and it is not robust. Since the size of input image may vary so the previous approach is impossible to use one generative model for input images of different sizes. Though for same sized input and generated images the previous approach works well.

[1] American University of Armenia, Marshal Baghramyan Ave. 40, Yerevan, 0019, Republic of Armenia;
E-mail: sparik_hayrapetyan@edu.aua.am

[2] Institute for Informatics and Automation Problems of NAS RA, P. Sevak str. 1, Yerevan, 0014, Republic of Armenia ; E-mails: gevorgk@ipia.sci.am; hakop@ipia.sci.am

[3] Don State Technical University, Gagarina sq. 1, Rostov-on-Don, 344000, Russia;
E-mail: voroninslava@gmail.com

This paper presents a new and efficient scheme that combines the advantages of approaches to image inpaining and single image super resolution. By using upscaling of the generated image we can successfully fill the missing region of original image is up to 4 times larger size. The approach allow significantly increase computation power necessary for generating large images. We have applied our algorithm on variety of face images. Generative model was trained on CelebA dataset which has 200 thousands samples.

The remaining sections are organized as follows: Section 2 describes the principles of generative adversarial networks. Image inpainting using generative adversarial networks is described in Section 3. Section 5 describes single image super resolution with the help of generative adversarial networks. The approach combining single image super resolution and digital inpainting presented in Section 6. Finally, the paper is concluded in Section 7.

2 Generative Adversarial Networks

Generative Adversarial Networks [2] (GANs) are a recent idea in the field of machine learning. The most common usage of GANs is generating realistic samples from high-dimensional probability distributions for which more traditional methods are infeasible. For example, GANs can learn to generate high-quality images of human faces given a dataset of human faces.

Traditional GANs consist of two neural networks, a generator and a discriminator. The generator learns to map from a low-dimensional probability distribution to the desired distribution (e.g. human faces), while the discriminator learns to distinguish between samples of the real sample distribution from the fake distribution learned by the generator network. The two networks are then trained in parallel, helping to improve each other.

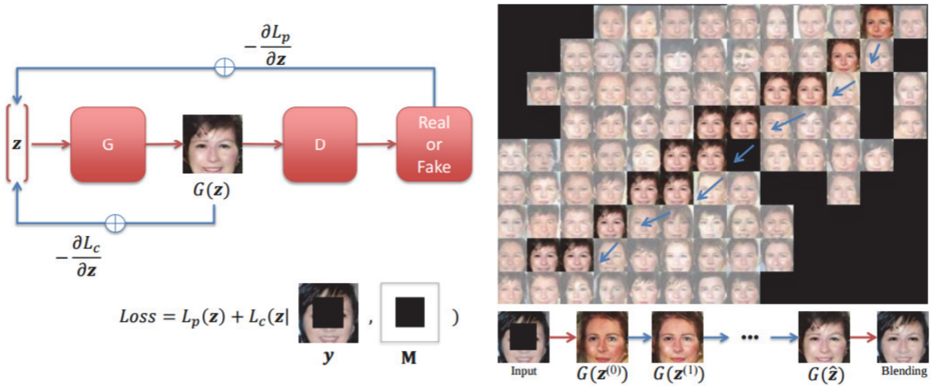
The task of a GAN is to learn a mapping from a low-dimensional latent space to the desired high-dimensional space (for which we have samples to train on). After the mapping is learned, each value Z out of the latent space corresponds to a single sample in the desired high-dimensional space (e.g. images of human faces). In this way, instead of sampling directly from the high-dimensional space, which can be difficult or impossible, we can sample a value from the latent space, and map it onto the desired space.

While Generative Adversarial Networks are still in their infancy, they are an active field of research in machine learning, and improvements in GAN architectures and training are being introduced very frequently.

For example, DCGANs [6] (Deep convolutional GANs) are an extension of standard GANs that were known to produce better results on generating images. Improvements of DCGANs over simple GANs include using stridden convolutions instead of pooling layers, and batch normalization almost everywhere.

3 Generative Adversarial Networks for Image Inpainting

DCGANs have been applied to the image inpainting task by Yeh et al. [7]. Given trained DCGAN with discriminator D and generator G, an image is to be inpainted and a binary mask indicating the points to be completed, they apply back-propagation to the input data Z of the generator to find $\hat{z}$ that completes the best image.


$-\frac{\partial L_p}{\partial z}$
z
G
$G(z)$
D
Real or Fake
$-\frac{\partial L_c}{\partial z}$
$Loss = L_p(z) + L_c(z|$,)
y
M
Input
$G(z^{(0)})$
$G(z^{(1)})$
$G(\hat{z})$
Blending

(a) (b)

Fig. 1 – *Usage DCGAN for image inpainting.*

More specifically, they formulate the process of finding $\hat{z}$ as an optimization problem. Let y be the corrupted image and M be the binary mask with size equal to the image, to indicate the missing parts with $M_{ij} = 0$. $\hat{z}$ is then defined as

$$\hat{z} = \arg\min_z \mathcal{L}_c(z \mid y, M) + \mathcal{L}_p(z)\,, \tag{1}$$

where $\mathcal{L}_p$ denotes the perceptual loss, and is defined as

$$\mathcal{L}_p(z) = \lambda \log(1 - D(G(z)))\,, \tag{2}$$

where λ is a constant denoting how important perceptual loss compared to $\mathcal{L}_c$, the contextual loss, and is usually chosen to be a small number of magnitude ≈ 0.003.

Contextual loss is a measure of similarity between the unmasked portions of the real image and the generated image. It can be defined in many ways, but Yeh et al. found empirically that an importance-scaled $l1$-norm works fine. So contextual loss is defined as,

$$\mathcal{L}_c(z \mid y, M) = \left\| W \odot (G(z) - y) \right\|_1\,, \tag{3}$$

where W_{ij} basically denotes how important it is to get pixel at (*i, j*) right. Yen et al. define *W* as

$$W_{ij} = (1 - M_{ij}) \sum_{x,y \in N(i,j)} \frac{1 - M_{xy}}{|N(i,j)|},$$

where $N(i,j)$ is the set of neighbors of (i,j) in some chosen window size.

Because of this definition, pixels that are far away from masked areas will play no role in contextual loss, as they do not usually help to paint masked pixels.

After finding $\hat{z}$, Poisson blending is also applied to blend the generated image $G(\hat{z})$ to the masked image to achieve a more photo-realistic inpainted image [5].

4 Advantages and Disadvantages of GANs for Inpainting

While using generative methods such as GANs instead of traditional methods for the image inpainting task is quite an appealing idea, but they also have drawbacks.

Firstly, GANs must be trained on a large dataset for producing good results. Also, inpainting by means of using GANs is a lot slower than using traditional methods.

However, we should note that GAN research is at its peak now, due to the inpainting method suggested by Yen et al. it does not depend on any specific architecture, it is progressive as it gets better results so the GAN architectures are better. Other possible improvements areas include the importance-weight matrix *W* and the loss functions.

5 Generative Adversarial Networks for Image Super-Resolution

5.1 Image super-resolution

Another image processing problem dealing with neural networks is image super-resolution (SR).

Given an image with low resolution, the problem of image super-resolution is to enhance the resolution of the given image. The basic baseline of the image super-resolution problem is bicubic interpolation. This problem was previously solved by using sparse-coding based methods [8].

Like in the case of many computer vision tasks, Deep Convolutional Networks have been applied to the SR task by Dong et al. in 2014 who claimed to achieve a state-of-the-art performance [1]. They apply bicubic interpolation to upscale the image to the desired size, and then feed it into a convolutional network trained to learn mapping from an upscaled low-resolution image to its high-resolution counterpart [1].

5.2 Image super-resolution using GANs

In 2016, Ledig et al. proposed a new architecture for the single image super-resolution (SISR) task based on Generative Adversarial Networks [4]. They incorporate recent innovations of deep learning into their architecture.

The architecture of the generator network *G* is based on a ResNet [3] with 3x3 convolutions and 64 feature maps followed by batch normalization layer and Parametric ReLU activation in each block. The input to the generator network is the low-resolution image and the generator learns to map the LR image to the HR image directly. [4]

The discriminator network *D* learns to distinguish between the super-resolved image and the ground-truth high-resolution image. Each convolutional layer in the discriminator halves the image size using stridden convolutions and has double the number of feature maps compared to the previous layer. Following the recent tendencies, this network also avoids pooling layers and employs batch normalization and LeakyReLU activations. The convolutional layers are followed by dense layers and a sigmoid function to map the output to probability. [4]

Notice that the architecture of the discriminator network is identical to the architecture used for image inpainting described in the previous section.

In contrast with other works in the SR domain that usually use the MSE of the super-resolved and the real HR images, Ledig et al. use a weighted sum of the adversarial loss and a content loss.

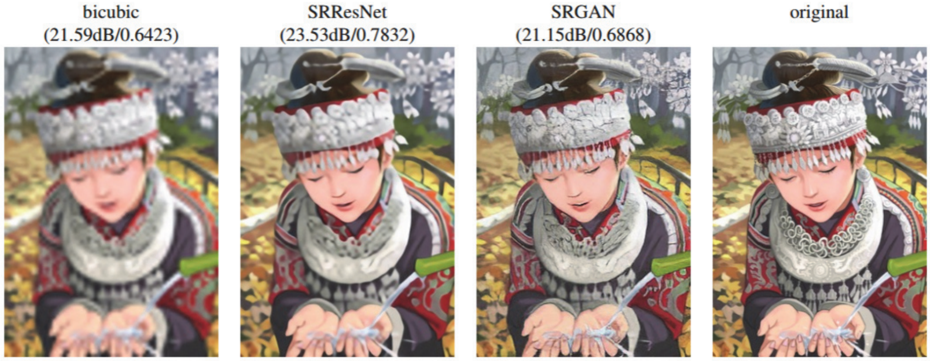


Fig. 2 – *From left to right: bicubic interpolation, deep residual network optimized for MSE, deep residual generative adversarial network optimized for a loss more sensitive to human perception, original HR image. Corresponding PSNR and SSIM are shown in (4× upscaling).*

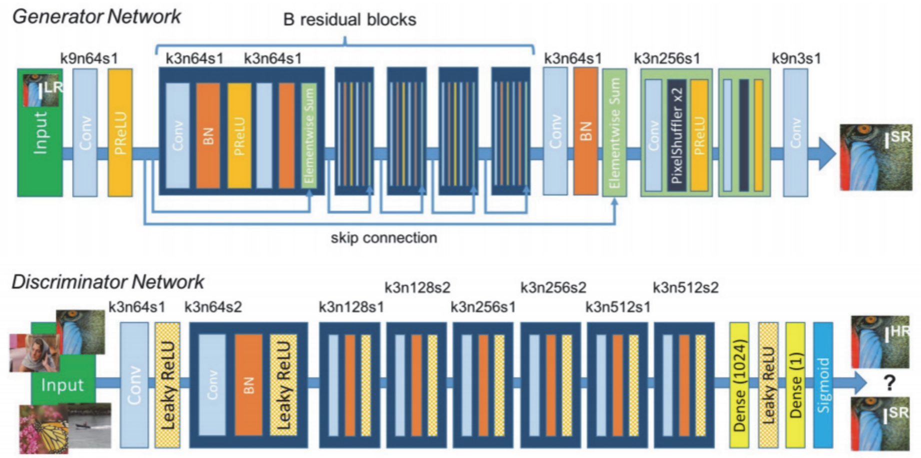


Fig. 3 – *The generator and discriminator network architectures of SRGAN.*

As in the case of the image inpainting task, the loss is decomposed into two parts, an adversarial loss and a content loss.

The adversarial loss is just the loss of the discriminator network, and can give the answer as how much the discriminator believes that the given image is a HR image.

After experimenting with some content loss functions, they argue that the usual pixel-wise MSE is not the perfect loss function for the SR task. Instead, they define various versions of VGG loss based on the *i*-th convolution before the *i*-th max-pooling layer of the VGG network, denoted as $\phi_{i,j}$. The proposed content loss function is defined as

$$l_{VGG_{i,j}} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G(I^{LR})) \right)^2 ,$$

where I^{HR} is the real high-resolution image, I^{LR} is the low-resolution image, and, consequently, $G(I^{LR})$ is the super-resolved image. $W_{i,j}$ and $H_{i,j}$ are the width and the height of $\phi_{i,j}(I)$, respectively [4].

According to the authors, using this loss function instead of pixel-wise MSE we can avoid the problem of ever smooth textures in the super-resolved image [4]. They specify to state that $l_{VGG_{5,4}}$ performed the best way.

While optimizing MSE as the loss function is beneficial in terms of PSNR score because minimizing MSE is the same as maximizing PSNR, Ledig et al. conducted a MOS (Mean opinion score) test by asking human respondents to evaluate the quality of the super-resolved images obtained by different methods on different datasets. While SRGAN did not particularly excel at getting a low PSNR score, it got the highest MOS scores on all datasets.

This suggests that PSNR is not a good measure for the image super-resolution task either.

6 Combining Inpainting and Super-Resolution

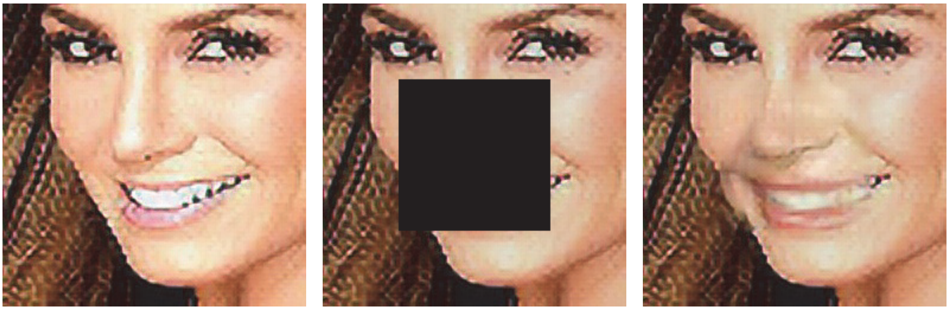
The GAN inpainting framework described above is specifically designed to inpaint fixed-size images, usually with equal width and height dimensions. While a new, slightly modified architecture can be designed for inpainting images of each specific size, it is highly inconvenient and expensive to do. Moreover, current GAN architectures are not adjusted to generating large realistic images, as most papers use 64×64 image size. Also, training the network for large image sizes is both harder and demands much more resources.

One way to overcome this issue is to combine the two GAN architectures described above.

Let us say that we have trained versions of both GANs. Then we have a mechanism to inpaint 64×64 images, and another mechanism for robust upscaling of a 64×64 image to an 256×256 image using the 4× upscaling SRGAN. Then, we can inpaint 256×256 (or smaller) image following the procedure below:

1. Downscale the image to 64×64;
2. Inpaint the image with the GAN inpainting procedure;
3. Super-resolve the image to 256×256 size, and downscale it if you need to match the original size;
4. Merge the inpainted regions with the original image.

You can see in Fig. 4 that inpainting, then super-resolving give quite good results.



(a) The original image;

(b) Input to the inpainting algorithm (upscaled);

(c) The inpainted and super-resolved image;

Fig. 4 – *The result of using inpainting combined with super-resolution versus the original image.*

7 Conclusion

We implemented both networks in Theano, and closely followed all the implementation advices mentioned by the authors. As it is mentioned by many authors, GANs are quite hard to train. We closely monitored the loss process and the generated samples to make sure the networks were properly trained.

We used $\lambda = 0.1$ as the perceptual loss weight for inpainting in our experiments, and found that, in most cases, it works better than $\lambda = 0.001$ as it was suggested by the authors.

The inpainting GAN network was trained on the celebA dataset with cropped and aligned images, for 25 epochs.

The SRGAN network was trained on the ImageNet dataset as in the original paper. All the preprocessing was done and all hyper-parameters were chosen according to the original paper.

We found that, inpainting besides being very slow for many practical purposes, using the suggested pipeline does not perform consistently even when it runs several times on the same image. Moreover, it is not robust to slight changes in the image.

8 References

[1] C. Dong, C.C. Loy, K. He, X. Tang: Image Super-Resolution using Deep Convolutional Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 38, No. 2, Feb. 2016, pp. 295 – 307.

[2] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio: Generative Adversarial Networks, ArXiv:1406.2661, June 2014.

[3] K. He, X. Zhang, S. Ren, J. Sun: Deep Residual Learning for Image Recognition, IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27-30 June 2016, pp. 770 – 778.

[4] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi: Photo-Realistic Single Image Super-Resolution using a Generative Adversarial Network, ArXiv:1609.04802, Sept. 2016.

[5] P. Perez, M. Gangnet, A. Blake, Poisson Image Editing. ACM Transactions on Graphics, Vol. 22, No. 3, July 2003, pp. 313 – 318.

[6] A. Radford, L. Metz, S. Chintala: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, ArXiv:1511.06434, Nov. 2015.

[7] R.A. Yeh, C. Chen, T.Y. Lim, A.G. Schwing, M. Hasegawa-Johnson, M.N. Do: Semantic Image Inpainting with Deep Generative Models, ArXiv:1607.07539, July 2016.

[8] J. Yang, J. Wright, T.S. Huang, Y. Ma: Image Super-Resolution via Sparse Representation, IEEE Transactions on Image Processing, Vol. 19, No. 11, Nov. 2010, pp. 2861 – 2873.

[9] V. Voronin, V. Marchuk, S. Makov, V. Mladenović, Y. Cen: Spatio-temporal Image Inpainting for Video Applications, Serbian Journal of Electrical Engineering, Vol. 14, No. 2, June 2017, pp. 229 – 244.

[10] V.V. Voronin, V.I. Marchuk, E.A. Semenishchev, S. Makov, R. Creutzburg: Digital Inpainting with Applications to Forensic Image and Video Processing, IS&T International Symposium on Electronic Imaging — Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications, San Francisco, CA, USA, 14-18 Feb. 2016, pp. MOBMU-289.1 – MOBMU-289.7.