Learning to Inpaint by Progressively Growing the Mask Regions

Mohamed Abbas Hedjazi, Yakup Genç
Gebze Technical University
Kocaeli, Turkey
mahedjazi,yakup.genc@gtu.edu.tr

Abstract

Image inpainting is one of the most challenging tasks in computer vision. Recently, generative-based image inpainting methods have been shown to produce visually plausible images. However, they still have difficulties to generate the correct structures and colors as the masked region grows large. This drawback is due to the training stability issue of the generative models. This work introduces a new curriculum-style training approach in the context of image inpainting. The proposed method increases the masked region size progressively in training time, during test time the user gives variable size and multiple holes at arbitrary locations. Incorporating such an approach in GANs may stabilize the training and provides better color consistencies and captures object continuities. We validate our approach on the MSCOCO and CelebA datasets. We report qualitative and quantitative comparisons of our training approach in different models.

1. Introduction

Image inpainting is a technique that allows filling in missing regions/holes or removing unwanted objects/artifacts in an image. This task is easy for humans since they can understand the image structure representing the scene, even when a significant portion of the scene is not visible. However, it is a very challenging task for a computer. It is applied in many problems including localization and segmentation [1], video compression [2], 3D shape inpainting [3], depth inpainting [4][5] and face verification [6].

Recently, deep learning methods [7, 8, 9, 10] applied Generative Adversarial Networks (GANs) [11] to fill in masked regions by learning from large image datasets. They outperform the traditional inpainting methods [12, 13, 14] both qualitatively and quantitatively. However, some of these methods [7] fill in the center of the image, that may fail to inpaint variable size regions. Furthermore, they suffer from artifacts around the inpainted regions and need post-processing steps to correct the resulted image [8]. Therefore, understanding the structure and different objects in the scene helps to achieve a high quality image completion.

Although GANs fit the inpainting problem very well, they suffer from stability problems that lead to mode collapse and over-fitting. To address these limitations, [15] provides architectural guidelines and optimization hyper-parameters that leads to better synthesis results. Moreover, a multi-stage generation approach introduced in [16] creates high-quality images by progressively adding layers to the generator and the discriminator. Furthermore, [17] improves [16] by controlling the visual features of the image in different scales through the adaptive normalization layer [18]. Some works addressed the loss functions improving the training stability including Wasserstein distance [19], Least Squares [19] and Energy-based GANs [20].

Another attempt to stabilize the training of GANs is to employ a curriculum learning (CL) approach [21]. It achieved a lot of success in many tasks, including natural language processing [22] [23], image recognition [24] and generation [25]. CL is a setting in which it gradually reveals training samples to the model from the easiest to the most difficult. Inspired by this idea, we propose a curriculum-style strategy to progressively train an effective generator by growing the size of the masked regions in the context of image inpainting. The intuition was that the generator and the discriminator networks solve the inpainting problem starting from simpler to much harder inpainting regions. By simpler, we mean small masked regions with basic structures which can be filled easily without the need for global object structures. On the other hand, harder means larger mask regions that need both local and global understanding of the scene.

We validate our approach using several models of different architectures and loss functions. The first one is our customized model that is trained using two networks: a deep residual convolutional generator [26], and a multi-scale discriminator that criticizes the quality and the relevance of the completed image in different scales. In the generator, we replace the vanilla convolutions with the gated convolutions introduced in [10]. They proved that it is a good replacement of vanilla convolutions in the context of image inpainting. The other methods are two of the state-of-the-art models [7] and [10]. We conduct two experiments: fixed versus progressively growing masked regions on the previously stated models. Additionally, to show the effectiveness of our approach, we check if a simple reconstruction loss is sufficient to stabilize the generator for the first training iterations. During training, we use a fixed masked region then gradually increase the adversarial loss weight. We report qualitative and quantitative results on the MSCOCO [27] and CelebA [28] datasets. The quantitative metrics include L1, L2, PSNR, Inception score (IS) and Frichet Inception Distance (FID) quality metrics

Our contributions are as follow:

- We propose the progressively growing of the masked regions as a GAN stabilization technique for image inpainting task.
- We compare the usage of fixed versus progressively growing mask regions using different architectures and loss functions, and we report the qualitative and quantitative results on two challenging datasets.
- We investigate other training stabilization setups and compare it against our approach.

2. Related Work

Traditional Inpainting methods such as the diffusion-based image synthesis propagates closest pixels around the masked regions to fill it in [29, 30]. Nevertheless, these methods have many limitations because they just complete texture patterns and do not understand the anatomy of the scene and the objects to be completed. Moreover, they cannot fill in large masked regions. On the other hand, patch-based methods can fill in large masked regions in images by searching for similar patches in the image [12, 13, 14]. However, these methods fail to fill in large holes in complex scenes especially when the texture to be filled is not present in the image. Furthermore, patch-based methods are slow, come at a large processing cost and are not based on a semantic understanding of the scene. Therefore, the inpainting task cannot be handled by traditional inpainting approaches since the missing region is very large for local-non-semantic methods to work well.

Deep Learning-based inpainting methods fill in masked values in an end-to-end manner by optimizing a deep encoder-decoder network to reconstruct the input image. But, it tends to produce blurry images and often proceeded by a post-processing step. To outperform this limitation, GANs [11] showed to be a great data distribution modeling technique synthesizing realistic-looking images. GANs train two networks against each other in a minimax game, the first one generates images from a random distribution (called generator) and the later which tries to distinguish between real and generated images (called discriminator).

In the context of image inpainting, [7] optimizes an encoder-decoder network to produce relevant content in a central rectangular hole in the image based on GANs. This approach can generate novel objects and textures in the image, but it lacks local consistency. To address this limitation, [8] extends it using a global and a local discriminator to ensure both the global coherence and the local image consistency. The drawback of this technique is the post-processing step that should be done using Poisson Image blending [31].

Another valuable work is [32] that presents a novel contextual attention layer to explicitly attend on related feature patches at distant spatial locations. [9] uses a stack of partial convolution layers and mask updating steps to perform image inpainting using an autoencoder without adversarial learning. The intuition was that regular convolutions treat both valid pixel values and masked values in the same manner while partial convolutions are conditioned only on valid pixels. The proposed architecture demonstrates the effectiveness of training image inpainting models on irregularly shaped holes.

Moreover, [10] introduces Gated Convolutions to overcome the limits of [9]. The latter is a hard-gating single-channel un-learnable layer multiplied to input feature maps. However, Gated Convolutions are learnable layers that learn a dynamic feature selection mechanism for each channel and each spatial location. Additionally, it allows user-input (a sketch) as an additional channel.

Curriculum Learning: is an effective approach to improve the training of neural networks. Unlike the traditional training approach of CNNs that uniformly samples mini-batches from the data distribution, [33] used CL to order the training samples by difficulty and creates mini-batches from them, this lets the network start with the easiest ones which improve both the accuracy and the learning speed. Further, [33] improves the generalization ability by increasing the dropout rate throughout training that gradually increases the difficulty of the problem.

Furthermore, [34] employed CL on GANs by making the discriminator solves harder problems during training. They augment the dimensionality of the sample space with additional random variables. This approach makes the task much difficult for the discriminator and prevents it from being over-confident.

In the context of image inpainting, [25] utilizes a progressive generative network to fill-in images with squared masks. The approach splits the task into different stages, where each one aims to do a part of the entire curriculum. After that, an LSTM framework is used to chain all of them

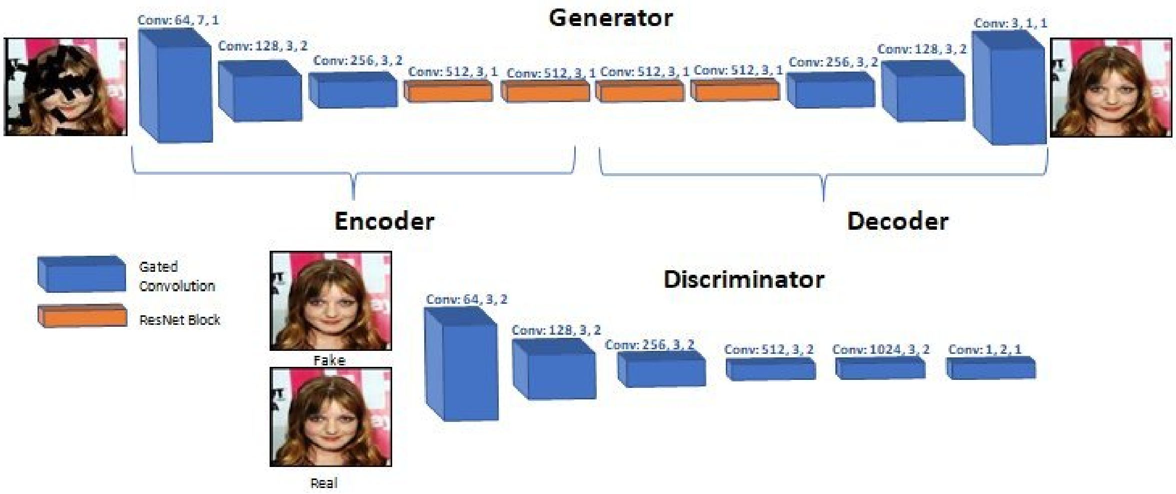
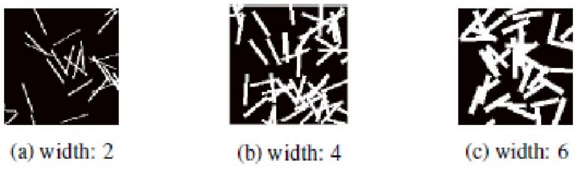


Figure 1. Illustration of the Generator (top) and the Discriminator (down) network architectures. The generator takes as input the masked image and the binary mask. It outputs the inpainted image. The discriminator takes either the ground truth (real) or the generated (fake) image as input and outputs either fake or real tensors.



(a) width: 2 (b) width: 4 (c) width: 6

Figure 2. Illustration of the progressively growing approach. After each K iterations, we progressively increase the masked region size until it reaches half the size of the image.

together. [35] utilize CL on the contour and image completion modules using different stages. The training starts using only the content loss, then they fine-tune it with a small weight for the adversarial loss. Finally, they fine-tune the whole module using 1:1 weights.

Unlike the previously mentioned methods, we adapt CL in the image inpainting task by progressively growing the masked regions during training. In test time, the generator network can fill irregularly shaped holes. To the best of our knowledge, no method uses a similar approach in the context of image inpainting.

3. Progressive Image Inpainting

As it is known, GANs are very hard to train due to its nature that depends on two networks having two sets of parameters optimized independently of each other. That leads to many problems, including mode collapse, non-convergence, and the vanishing gradients. The inpainting task is strongly affected by robust adversarial loss functions, stable architectures, and GAN stabilization techniques. We focus on the last point and propose a simple yet effective training technique to stabilize the training of GANs in the context of image inpainting. The process is as follows: the generator starts by solving a simple problem successfully, after each K iterations the masked region grows to a much harder problem till the region size reaches the half size of the image, as illustrated in Figure 2. By simple, we mean that the masked region contains basic structures (textures) while hard refers to masked regions that contain complicated structures and objects.

We claim that, in the beginning, the generator easily fills-in the narrow masked region since the adversarial loss is responsible for an easy problem that is simply a reconstruction in this case. Then, the problem's difficulty increases as we grow the width of the mask. Consequently, the generator can fill-in the half size of the masked region without much difficulty. That makes the adversarial loss stable in the next K training iterations. The training process continues this way till a specified maximum width. We will investigate this claim by reporting the quantitative results of each K iterations using different mask sizes.

4. Architectures and Training

To validate our approach, we use different models: our customized model illustrated in Figure 1, the Context-encoder model [7] and the Free-form inpainting model [10].

Our customized model: the generator has two sub-networks, an encoder network that down-samples the size of the input to 1/4 the original size followed by two residual blocks. We duplicate the number of filters after each gated convolution and residual block. The decoder network is the reverse order of the encoder. Instead of using transposed convolutions as generally done in decoders, we use bilinear interpolation before applying gated convolutions. The last convolution layer outputs an RGB image.

In the discriminator network, we use a multi-scale architecture that contains five convolution layers. It down-samples the feature maps size and increases the number of filters. The last two convolution layers have the same number of filters. The discriminator outputs an array of network layers in different scales.

Instead of using Batch Normalization [36] that seems to cause problems in the inference time, especially when the batch size is small, we use the Instance Normalization [37] that normalizes each batch independently across spatial locations. Additionally, it provides visual and appearance invariance, moreover it is agnostic to the contrast of the image. The loss functions include the LSGAN loss [38], an L1 loss between the non-masked regions in the ground truth and the generated image, an L1 loss between the masked region in the ground truth and the generated image, finally, we include the Perceptual loss using a pre-trained VGG network [39].

The Context-encoder model: optimizes an auto-encoder network to produce a rectangular hole in center of the image. The discriminator considers the later as fake, while the center of the ground truth image as real. The training requires two loss functions: a pixel-wise reconstruction loss and an adversarial loss [11].

The Free-form inpainting model: the generator has the same architecture as [32] followed by a refinement network without residual connections. The discriminator is a Patch-GAN that classifies image patches of size 70x70 as real or fake. Thus, there is no need for a global and local discriminator as in [8]. Furthermore, the networks do not add any normalization layer. It computes two loss functions: the Hinge loss and the reconstruction loss. It does not include any perceptual or style loss.

5. Experimental Setups

In this section, we describe the datasets, the experimental setups, and the comparisons planning.

Datasets: we experiment on a variety of challenging datasets including MSCOCO [27] and CelebA [28]. The first dataset contains cluttered scenes with a lot of changes in colors and structures. The later dataset contains cropped faces that have fewer structure changes. We train on 200k and 82k training images defined in CelebA and MSCOCO, respectively. We test the performance on 10000 random validation images (no available test set) for the CelebA dataset and 5000 test images for the MSCOCO dataset.

Experimental setups: our main experimental setup is to investigate the fixed size masks versus the progressive growing approach. We use constant weights for both the reconstruction and the adversarial loss. To prove/disprove our claim and hypothesis in section 3, we are planning to explore the following setups using a fixed mask size:

- Use a simple reconstruction loss for the first k iterations, then include the adversarial loss where both loss functions will have fixed weights.
- Fix the reconstruction loss during the whole training and increase the adversarial loss weight after each k iterations.

Comparison plan: unlike the common comparison showing the outperformance of their method against the SOTA, in our case, we aim to confirm the impact of our proposed training scheme (Progressive growing) and the two other setups described above on different architectures including our model, CE, and the Free-form inpainting model. To adapt our training approach to the CE model, we are planning to start the training process with a small rectangle in the middle, then progressively increase the rectangle size to reach the half size of the image. The Free-form inpainting model adds the sketch as an additional input to the model. To ensure a fair comparison, we only input the image and the mask.

We test on the MSCOCO and CelebA datasets for the different setups on our customized model and the Free-form inpainting model. We report the quantitative comparison using L1, L2, PSNR, Inception score, and Frichet Inception Distance. Furthermore, we show the output of our customized model versus the Free-form inpainting model on different training schemes in the qualitative comparison. Since the CE model input is a fixed central mask in the middle of the image, we do not compare it against the other models. Thus, we only report the qualitative and quantitative results of the different setups against each other. We do not perform any post-processing step for all the models. We use images of resolution 128x128 in both datasets.

We implement the models using Pytorch v1.1.0, CUDA v10.0, CUDNN v7.5.1, and the hardware GPU is NVIDIA GTX 1080 Ti. The training takes roughly five days per experiment.

References

[1] M Partha Sarathi, Malay Kishore Dutta, Anushikha Singh, and Carlos M Travieso. Blood vessel inpainting based technique for efficient localization and segmentation of optic disc in digital fundus images. *Biomedical Signal Processing and Control*, 25:108–117, 2016.

[2] Yu Gao, Gene Cheung, Thomas Maugey, Pascal Frossard, and Jie Liang. Encoder-driven inpainting strategy in multiview video compression. *IEEE Transactions on Image Processing*, 25(1):134–149, 2015.

[3] Weiyue Wang, Qiangui Huang, Suya You, Chao Yang, and Ulrich Neumann. Shape inpainting using 3d generative adversarial network and recurrent convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2298–2306, 2017.

[4] Lucian Petrescu, Anca Morar, Florica Moldoveanu, and Alin Moldoveanu. Kinect depth inpainting in real time. In *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, pages 697–700. IEEE, 2016.

[5] Dylan Seychell and Carl James Debono. Monoscopic inpainting approach using depth information. In *2016 18th Mediterranean Electrotechnical Conference (MELECON)*, pages 1–5. IEEE, 2016.

[6] Shu Zhang, Ran He, Zhenan Sun, and Tieniu Tan. Multitask convnet for blind face inpainting with application to face verification. In *2016 international conference on biometrics (ICB)*, pages 1–8. IEEE, 2016.

[7] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[8] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):107, 2017.

[9] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[10] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, volume 28, page 24. ACM, 2009.

[13] Connelly Barnes, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. The generalized patchmatch correspondence algorithm. In *European Conference on Computer Vision*, pages 29–43. Springer, 2010.

[14] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Trans. Graph.*, 31(4):82–1, 2012.

[15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[18] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[19] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[20] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

[21] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

[22] Tom Kocmi and Ondrej Bojar. Curriculum learning and minibatch bucketing in neural machine translation. *arXiv preprint arXiv:1707.09533*, 2017.

[23] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M Mitchell. Competence-based curriculum learning for neural machine translation. *arXiv preprint arXiv:1903.09848*, 2019.

[24] Nikolaos Sarafianos, Theodore Giannakopoulos, Christophoros Nikou, and Ioannis A Kakadiaris. Curriculum learning for multi-task classification of visual attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2608–2615, 2017.

[25] Haoran Zhang, Zhenzhen Hu, Changzhi Luo, Wangmeng Zuo, and Meng Wang. Semantic image inpainting with progressive generative networks. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1939–1947. ACM, 2018.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[29] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co., 2000.

[30] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004.

[31] Kaiming He and Jian Sun. Image completion approaches using the statistics of similar patches. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2423–2435, 2014.

[32] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.

[33] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. *arXiv preprint arXiv:1802.03796*, 2018.

[34] Rishi Sharma, Shane Barratt, Stefano Ermon, and Vijay Pande. Improved training with curriculum gans. *arXiv preprint arXiv:1807.09295*, 2018.

[35] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5840–5848, 2019.

[36] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[37] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[38] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.