Proc. International Conference on Mechanical, Automotive and Mechatronics Engineering (ICMAME 2023)
29-30 April 2023, Dubai, UAE

Semantic Image Inpainting with Generative Adversarial Models and Skip Connections

Nikita Sharma
ABV-Indian Institute of Information Technology and Management
Gwalior, India
nikitaa081998@gmail.com

Anuraj Singh
ABV-Indian Institute of Information Technology and Management
Gwalior, India
anuraj@iiitm.ac.in

***Abstract*—Semantic Image Inpainting is a challenging and promising research problem where parts of an image are masked or corrupted and need to be filled based on the semantic and visual information available from the image. Image Inpainting can help revive ancient scriptures, paintings, and recover corrupted parts of images without the guidance of a subject matter expert. Existing computer vision techniques using neighborhood based gap filling provide unsatisfactory results. In this paper, a novel visual feature learning model driven by context-based pixel prediction is employed, which develops the missing part of the image by conditioning on the available data. We adopt a Generative model based approach to derive inference about the missing part of the image from the context of the entire image. A novel loss function based on the linear combination of generative loss and contextual loss provides promising results for semantic image inpainting. Experiments on datasets show that our method successfully predicts information in large missing regions and achieves pixel-level photorealism.**

I. Introduction

Images are one of the most vivid and frequently used form of communication used to exchange information. Humans learn to perceive and express their thoughts, ideas and expressions the best via images. Due to unforeseen circumstances like network or data corruption in case of digital image transfer or with time and tide, in case of physically stored images, they can get corrupted or a part of it might get distorted. This may lead to loss in important information. Heritage residing in Historical Artifacts and scriptures can become irrecoverable and hence, lost forever. Image Inpainting refers to the re-construction or revival or distorted or corrupted parts of the image in a manner that the information lost by image corruption can be recovered. Image Inpainting has been a challenging real-world problem and various techniques have been developed for achieving reasonable and pragmatic outputs. It can revive ancient scriptures, historical heritage and timeless art without the involvement of a subject matter expert. Performing image reconstruction manually requires a lot of time and effort. Thus, these techniques can prove to be influential for many real-life applications.

A. Semantic Image Inpainting

Semantic Image Inpainting attempts to perform Image reconstruction by taking into account the context of the image as well as the pixels surrounding the missing region. When pixel-based algorithms for image inpainting incorporates the semantics of the image, the accuracy and efficiency of the algorithm is bound to increase.

B. Motivation

Semantic Image Inpainting is a relatively lesser explored field for image reconstruction which follows the idea of how human brain perceives the surroundings and try to re-construct the scene in their mind. Two kinds of information are imminent for the same, information required to analyse the objects in the image scene, and then to understand their context with respect to one another and to the missing region. Existing computer vision techniques use neighborhood based pixel-wise color adjustment which provides unsatisfactory results since they are unable to capture the semantics. Neural Networks have been influential in determining hidden and complex relationships among the data points. Auto-Encoders are neural networks which try to reconstruct an input image by learning it's feature representation in a compact manner by passing the input through a series of layers and at last, a bottleneck layer. The decoder part takes the output of bottleneck as input and tries to reproduce the initial input data. Combining the Encoders with a generative model will incorporate the semantics of an image and can produce promising image inpainting model.

II. Related Work

Image Inpainting being an application-driven and dominant research field, there has been some significant work related to image reconstruction performing patching based and context based image inpainting. Some of the modern techniques also involve a combination of contextual and perceptual information.

A. Computer Vision based techniques

When image restoration is the main objective, various computer-vision based techniques are employed. These techniques mostly perform pixel-based nearest neighbourhood localities filling where the missing pixels were filled based on the pixel gradients of their neighbours. This approach was adopted by Xu et al in [1] and by Erkan et al in [2]. A similar technique also employed kinect depth maps in [3]. Ding et al use texture matching for image inpainting in [4]. Another technique involves patch searching and filling where the algorithm has access to a large dataset of images and the missing region of the input image is searched and matched with patches from various images. The image patch with the highest similarity value is selected and it's patch is copied and pasted over the corrupted region. Such researches were accelerated by better searching algorithms suggested in [5] by Xiong et al and improved edge detector introduced in [6].

B. Context based Machine learning techniques

Image Inpainting approaches involving contextual properties as well have becoming widely popular and dominant in the field. The techniques mentioned in the previous section took into account only the localitites of pixels. They did not have any provision to incorporate the semantics of the image or how the objects in the image are relevant to each other and the scene portrayed in the image. Context based techniques have therefore gained wide-spread support. Convolutional Neural Networks provided quintessential results in the field of image classification and segmentation. On the same lines, Lin et al proposed a deep convolutional neural network with residual connections for the image-inpainting task in [7]. The resultant image was however, not sharp and did not have appropriate contrast. Initial works involving Auto-encoder neural network architecture were context based Auto-encoders which reconstruct the image by learning the semantics of the image and it's feature representation via a bottleneck layer. The reconstructed image is the creation of the auto-encoder and the semantic information retained by it. This technique was used by Yadong et al in [8] and [9] which also used edge detector with encoder to address any edge based deformities and context blending across boundaries in the image reconstruction. Generative models introduced by Goodfellow et al in [10] is an interesting method for image generation by using adversarial losses between two loosely connected neural networks. Few such research works involving Generative Adversarial Networks were [11] and [12] for traditional image inpainting and crack detection and inpainting respectively. This research work is mainly inspired by work executed by Pathak et al in [13]. It combined the approach of context-encoders with generative adversarial network where the encoder was coupled with the generator, and a loss function which was a linear combination of two losses: reconstruction loss and adversarial/contextual loss was used. The above mentioned issues of current models not being able to accommodate the semantics of an image efficiently are addressed in the model discussed in this study by introducing a better and efficient architecture of Conditional Generative Adversarial Networks with Auto-Encoders coupled by a linearly combined loss function. Re-iterating the research gaps that we discovered while exploring various semantic image inpainting:

- The traditional patch based or pixel based approaches miss out on the context of the missing or distorting part, therefore it is based heavily on existing datasets and the similarity of test images and images in the existing dataset.
- All the techniques that explore Semantic Image Inpainting do not look into tuning losses as per context, rather they focus on tuning losses based on the L2 loss.
- An understanding of Generative Adversarial Networks can be applied with Autoencoders to accommodate context but the given concept is not explored to a saturation point and a lot of possibilities reside under the umbrella of neural networks.

C. Objectives

1) To develop an intuitive and influential approach for Semantic Image Inpainting, involving a combination of Context Encoders and Generative models.
2) Use a combination of Adversarial and content-based loss to accommodate the semantics of an image as well as the context of the missing region in an efficient manner, rather than using neighbourhood based techniques which only take into account pixel localities of the corrupt region.
3) To develop a robust and efficient model in it's entirety providing results comparable to the state-of-the-art approaches.

III. METHODOLOGY

In this section, a detailed overview of the low-level design and the architecture of the model to enhance the quality of semantic image inpainting is discussed. The architecture discussed in this paper uses Context-Encoders and Generative Adversarial Networks in a loosely coupled model. The combination of two networks along with a combined loss function to incorporate contextual and perceptual information is an attempt to enhance the performance for efficient practical applications.

A. Dataset

The dataset used in Image Painting is basically images with some regions of it corrupted or missing. The images can be randomly crawled from the internet. The dataset used in this project was collected from a subset of the **ImageNet Dataset and the Paris-Street-View Dataset**. The ImageNet Dataset consisted of over a million images for an image classification problem and were classified into 1000 categories. For our task, we do not require the labels. A subset was also taken from the Paris-Street-View Dataset which consists of images that mostly consisted streets and shops from paris. A glimpse of the dataset is given in fig

B. Dataset Pre-Processing

For the task of Semantic Image Inpainting, we need to drop few regions of the image so as to perform gap filling with the help of our model and compare it with the original image. Therefore, we mask some parts of the images with the help of a binary mask M. After masking the image with M, the regions representing 0 are the corrupted regions and the regions with 1 have their original pixels. Masking strategies that are being used are the following three:

Fig. 1. Central Region Masking

Fig. 2. Random Regions Masking

Fig. 3. Random Blocks Masking

1) *Central Region:* A part of the central region is being masked resulting the missing region to be 1/4th the size of the image.
2) *Random Region:* Regions are being masked in arbitrary shapes resulting the missing region to be 1/4th the size of the image.
3) *Random Blocks:* Overlapping blocks at random positions are being masked resulting the missing region to be 1/4th the size of the image. Instances of the masked images of the Dataset are given in fig 1, fig 2 and fig 3.

C. System Architecture

This study is mainly about exploring the best possible ways to unravel the possibilities that semantic image inpainting opens up for practical applications. The architecture proposed in this work is based on Generative Adversarial Networks, where two loosely coupled networks work in mutualcompetition an benefit from each other's performance.

1) *Generative Adversarial Networks:* Generative Adversarial Networks(GAN) are a new and interesting advancement over generative models for image generation. Under this architecture, two neural networks function and compete which each other which leads to mutual improvisation in performance of both networks. GAN consists of two neural networks, The Generator and The Discriminator. The Generator tries to create compelling life-like images from random or white noise and tries to fool the Discriminator into considering this image as a real image. The Discriminator tries to distinguish between real and fake images. The function used to optimize the performance is a Min-Max Objective function. It helps to increment the probability of the discriminator to understand the difference between real and fake images and thus, in turn compels the generator to create better real-like images to try and fool the discriminator in a more effective manner. The GAN architecture used in this work also consist of skip connections between the layers like a deep residual network (ResNet).
Min-Max Objective function-

$$L(gan) := maxE[f(D(x))] + E[g(D(G(z)))].$$

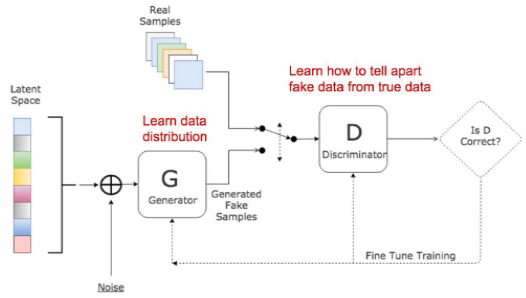

Fig. 4. GAN: Basic architecture

2) *Generator Architecture:* The generator network is developed by grouping B identical residual blocks in a consecutive fashion. Each of these blocks have of 2 Convolutional layers with 33 kernels and 64 feature maps. To reduce co-variate shifts, batch normalisation layer is added. Skip connections help to pass un-altered information from lower to the higher layers. For the activation function in each node of the network, a parametric ReLU activation function is used.

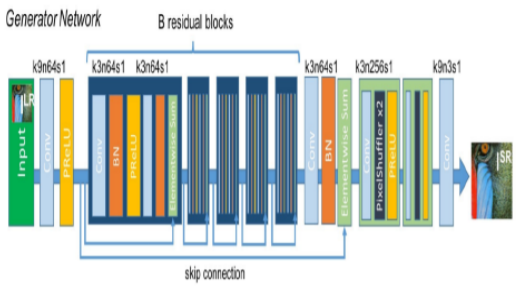

Fig. 5. An instance of Generator Architecture with residual blocks

3) *Discriminator Architecture:* The network of the discriminator has 8 convolutional layers of 3 3 filter kernels, growing at a steady pace by a factor of 2 from 128 to 512 kernels. Rather than using pooling layers, strided convolutions is used since it is a kind of convolution operation over the given input image matrix and can help improve the learning capabilities of the network.However, pooling just selects the maximum or average pixel value from a locality and does not contribute to learning any new information. We finally get 512 feature maps which are then followed by 2 fully connected layers. Sigmoid activation function is used which indicates 1 as real and 0 as fake image.

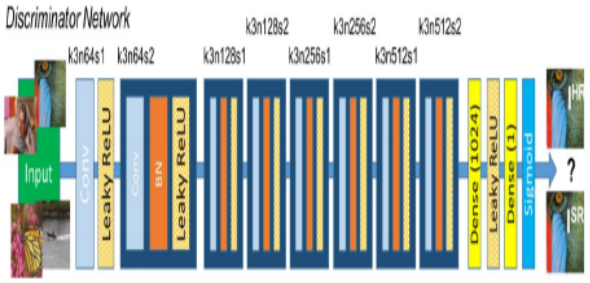

Fig. 6. An instance of Discriminator Architecture

4) *Loss Function:* The problem statement at hand needs a loss function tailored to address the semantics as well as the perceptual information of the image so as to produce realistic images where the network reconstructs the image their missing region has been filled by the network. The Mean Squared Error(MSE) which evaluates the average summation of squared pixel-wise difference of the predicted and expected values is not fit for the task because the averaging of values lead to blurred and blunt pixel gradients which is not desirable. Thus the loss function for the problem needs to be a linear combination to grasp the contextual aspects and semantics of the image. The proposed joint loss is a linear combination of content loss and adversarial loss. The joint loss is depicted as:

$$L = \lambda_{content} L_{content} + \lambda_{adv} L_{adv}$$

The content loss used in this research is dependent on the perceptual similarities representing the differences between corresponding pixels between the activation layers of an AlexNet Model. The loss is thus the Euclidean distance between the feature representations of an input image with its missing regions and the original image itself. The layers to be chosen needs to be experimented with for best possible results.

The adversarial loss tends to support the images generated which are more realistic. The generator generates images from random or white noise and thus, it can be anything. To generate an image similar to our corrupted image with missing regions re-constructed, we need to condition the generator on the corrupted image. The Adversarial loss is therefore, probabilities of discriminator over the training samples. Adversarial loss-

$$L_{adv} := maxE[log(D(x)) + log(1 - D(f(1 - M) \odot x))]$$

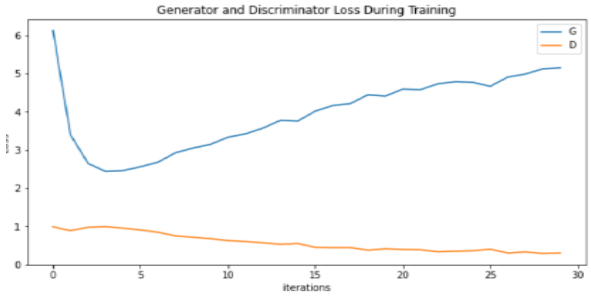

Fig. 7. Generator and Discriminator loss during training

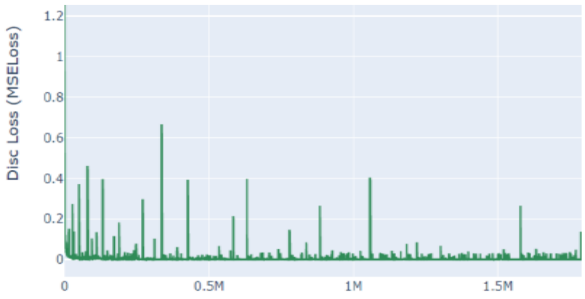

Fig. 8. Generative Adversarial Loss

IV. RESULTS

This section discusses the various experiments pertaining to the proposed hypothesis and their findings.

The architecture proposed in this research work is expected to be at par with the state-of-the-art approaches. The model is expected to provide an image that has been semantically inpainted and resembles the original image as closely as possible. The joint loss function is expected to help capture the contents and the context of the image efficiently and produce high quality images with photo-realism with their corrupted or missing regions filled according to the semantics of the

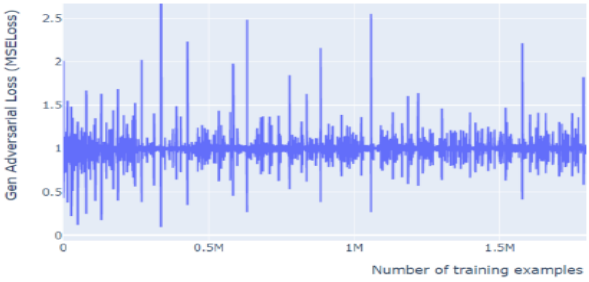

Fig. 9. Discriminator loss

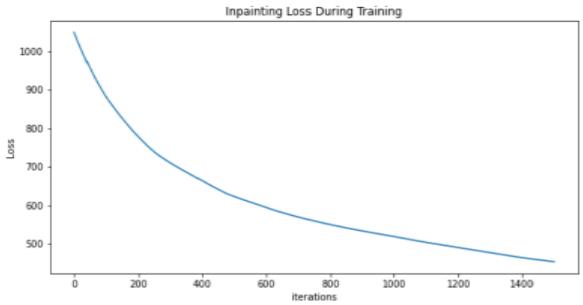

Fig. 10. Inpainting loss

Fig. 11. Output Image 1

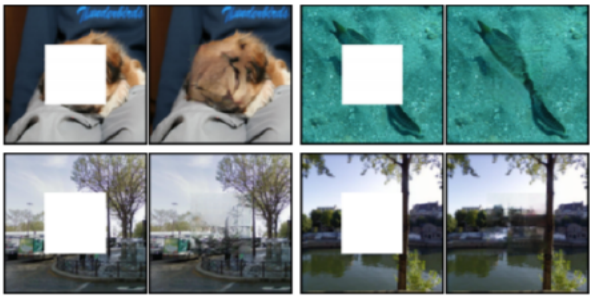

Fig. 12. Output Image 2

image and the pixel values of the locality. The architecture is expected to converge with an optimal output image so that the value of the joint loss function can be minimised as much as possible to make the model's output image resemble the original image. Our proposed network depicts our baseline model with an enlarged dataset and an optimized loss function. In this experiment, the architecture is formed as a combination of GANs and ResNets and uses a joint loss that is a combination of perceptual loss based on AlexNet and an Adversarial loss for keeping a check on filling missing regions as close to reality as possible. The range of the LR input images was scaled to [0, 1]. We use Adam with 1 = 0.9 for optimization. With a learning rate of 10-4 and 106 update iterations, the SRResNet networks were prepared. To prevent undesired local optima, we used the trained ResGAN network as an initialization for the generator when training the actual GAN. The loss function after tuning the coefficients for both losses turn out to be given by:

$$L = \lambda_{content} L_{content} + 10^{(-3)} * \lambda_{adv} L_{adv}$$

V. FINAL OUTPUT IMAGES

Listed in this section are output images that resulted from the final training of the ResNet+GAN architecture.

VI. CONCLUSION

A brief analysis of image inpainting is presented in this article. Sequential-based (classical approach without learning), CNN-based and GAN-based methods are among the approaches that have been presented. We also try to collect methods for dealing with various forms of image distortion, such as text, objects inserted, scratches, and noise, as well as different types of data, such as RGB, RGB-D, and historical images. Learned features, such as deep learning, are a strong alternative to these traditional features because they have greater generalisation potential in more complex scenarios. These models must be trained on a large amount of data in order to be accurate. To do so, we assemble the most commonly used datasets for training these models. To summarise the various analyzed cases and their results, we present a summary in the form of tables for each category of approaches, presenting their evaluation based on the types of data, datasets, and metrics used for each approach. To summarise, there is no tool that can inpaint all forms of image distortion, but learning techniques can help. In conclusion,there is no singular method for dealing with every type of distortion of images but using several deep learning approaches involving GANs provide some very promising results for analysis. Amongst the approaches mentioned above, the Approach with GANs and ResNets combined over a joint loss gives the most promising results.

ACKNOWLEDGMENTS

This work is supported by the Mathematical Research Impact Centric Support scheme (MTR/2020/000477), Science and Engineering Research Board (SERB), India.

REFERENCES

[1] Y. Xu and S. Wang, "Image inpainting base on color differences and structure differences," in *Proceedings of 2013 3rd International Conference on Computer Science and Network Technology*. IEEE, 2013, pp. 364–368.

Fig. 13. Output Image 3

[2] U. Erkan, S. Enginoğlu, and D. N. Thanh, "An iterative image inpainting method based on similarity of pixels values," in *2019 6th International Conference on Electrical and Electronics Engineering (ICEEE)*. IEEE, 2019, pp. 107–111.

[3] M. Stommel, M. Beetz, and W. Xu, "Inpainting of missing values in the kinect sensor's depth maps based on background estimates," *IEEE Sensors Journal*, vol. 14, no. 4, pp. 1107–1116, 2013.

[4] D. Ding, S. Ram, and J. J. Rodríguez, "Image inpainting using nonlocal texture matching and nonlinear filtering," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1705–1719, 2018.

[5] X. Xiong, H. Zhao, and C. Tang, "Image inpainting used in specific target replacement," in *2008 International Conference on Information and Automation*. IEEE, 2008, pp. 354–356.

[6] C. De and F. Y. Shih, "Improved image inpainting using maximum value edge detector," in *2012 IEEE Southwest Symposium on Image Analysis and Interpretation*. IEEE, 2012, pp. 77–80.

[7] D. Lin, G. Xu, Y. Wang, X. Sun, and K. Fu, "Dense-add net: An novel convolutional neural network for remote sensing image inpainting," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 4985–4988.

[8] W. Yadong, H. Yongguo, and Z. Hongying, "Image encoding and decoding scheme based on digital image inpainting technology," in *2009 International Conference on Information Technology and Computer Science*, vol. 1. IEEE, 2009, pp. 51–54.

[9] T. Ružić and A. Pižurica, "Context-aware patch-based image inpainting using markov random field modeling," *IEEE transactions on image processing*, vol. 24, no. 1, pp. 444–456, 2014.

[10] M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, I. J. Goodfellow, and J. Pouget-Abadie, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 2672–2680, 2014.

[11] X. Wu, J. Li, S. Lang, X. Zhang, C. Shi, X. Li, K. Jia, and M. Zou, "Perceptual generative adversarial network for image inpainting with regular and irregular patch," in *2019 2nd International Conference on Safety Produce Informatization (IICSPI)*. IEEE, 2019, pp. 46–50.

[12] S. Jaidilert and G. Farooque, "Crack detection and images inpainting method for thai mural painting images," in *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2018, pp. 143–148.

[13] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.