Generative Image Inpainting using Deep Learning

Ujjwal Sharma (2017IMT-108)

Supervised by:

Prof. Rajendra Sahu

ABV-IIITM Gwalior
Gwalior-474 015, MP, India

November 7, 2020

Introduction

- Most common issues in the field of Deep Learning and Processing of images are Image Completion and Image Enhancement.
- In real life applications, the complexity of some computer vision tasks is increased due to some corrupted or missing values of pixels in images.
- Application of Image Completion and Image Enhancement separately causes poor results.
- Work is required not only on filling the missing image but also make it as similar as the existing image
- Many techniques exist that take care of low level features and try to handle Image Completion and Image Enhancement separately. However, these issues may occur concurrently [13].

Introduction

- Generative Networks pick local points from latent space which are basically random points relating to some distribution.
- Most of these techniques are overwhelmingly used in restoration of art paintings, surveillance and security systems at high level.
- In our thesis, Least Squares Generative Adversarial Network (LSGAN) [18] has been used for image completion and enhancement network used is inspired by Auto-Encoder Network which refines the quality of the image.

Literature Review

- Ahuja and Jia-Bin have proposed an advance-knowledge approach which applied contextual information for image completion [11].
- Zhao, Liu, and Hiang proposed deep neural networks to inpaint and denoise the corrupted image [26]. However, in this method, the handling of structure and texture was not satisfactory.
- Another approach was developed by Deepak Pathak using Context Encoders[19] to estimate missing areas with the help of surroundings. However, the results produced were blurry and included noise.

Literature Review

- The model put forward by Chen and Fu is also called Progressive Inpainting[22] using Generative Models where the models first predicted the full missing region.But this model could not handle the issue of large missing regions.
- In Variational Image Completion, CVAEs[2] are used for image completion but these also use conditional labels.However, on a practical scale there is an absence of labels and the results produced were blurry for very large regions.
- Chuanxia Zheng,Tat-Jen Cham,Jianfei Cai regarding the pluralistic inpainting has been developed that creates multiple possible solutions for the existing missing regions using GANs [27].However, this method is a bit slow and the training needs a lot of existing data.

Gaps in literature

- Structural Continuity
- Calculation of Correct Distribution
- Large Patches
- Time Required for Inpainting

Problem Statement

- Current methods of image inpainting suffer from major issues like large size of corrupted images or if the image of very high dimensions which resulted the algorithms to generate low quality and unsatisfactory images.
- Image inpainting is also difficult because a corrupted image may contain pixels which are of drastically different distributions than the pixels of non corrupted part of the image. As the information is permanently lost,it is impossible to generate exact missing portions. So, the results should provide a resemblance.

Objective

- The Aim is to further improve the performance the Generative Adversarial Networks[6] to reduce Perceptual and Contextual Losses[23] by adding networks for refining images.
- We intend to improve upon the previous work done by adding:
  - Least Squares Generative Adversarial Network Architecture.[18]
  - Auto-Encoder Refinement Network.[1]
- To maximize resemblance of completed image and actual image is the main objective of our Model.

Novelty

- Images are completed solely based on visual input provided by using Extracted Features and deriving Importance Matrix of the incomplete image.
- New structure that incorporates LSGAN combined with the output of Auto-Encoder Model.
- Better Pose Detection and Better Quality Image completion over existing GANs and stabler learning and adaptation process.
- Clubbing the problems of Image Completion and Image Refinement together using multipurpose models with better Quality of Image Completion
- The new Model maintains greater and better variational diversity for Image Completion for the given target distribution.

Methodology

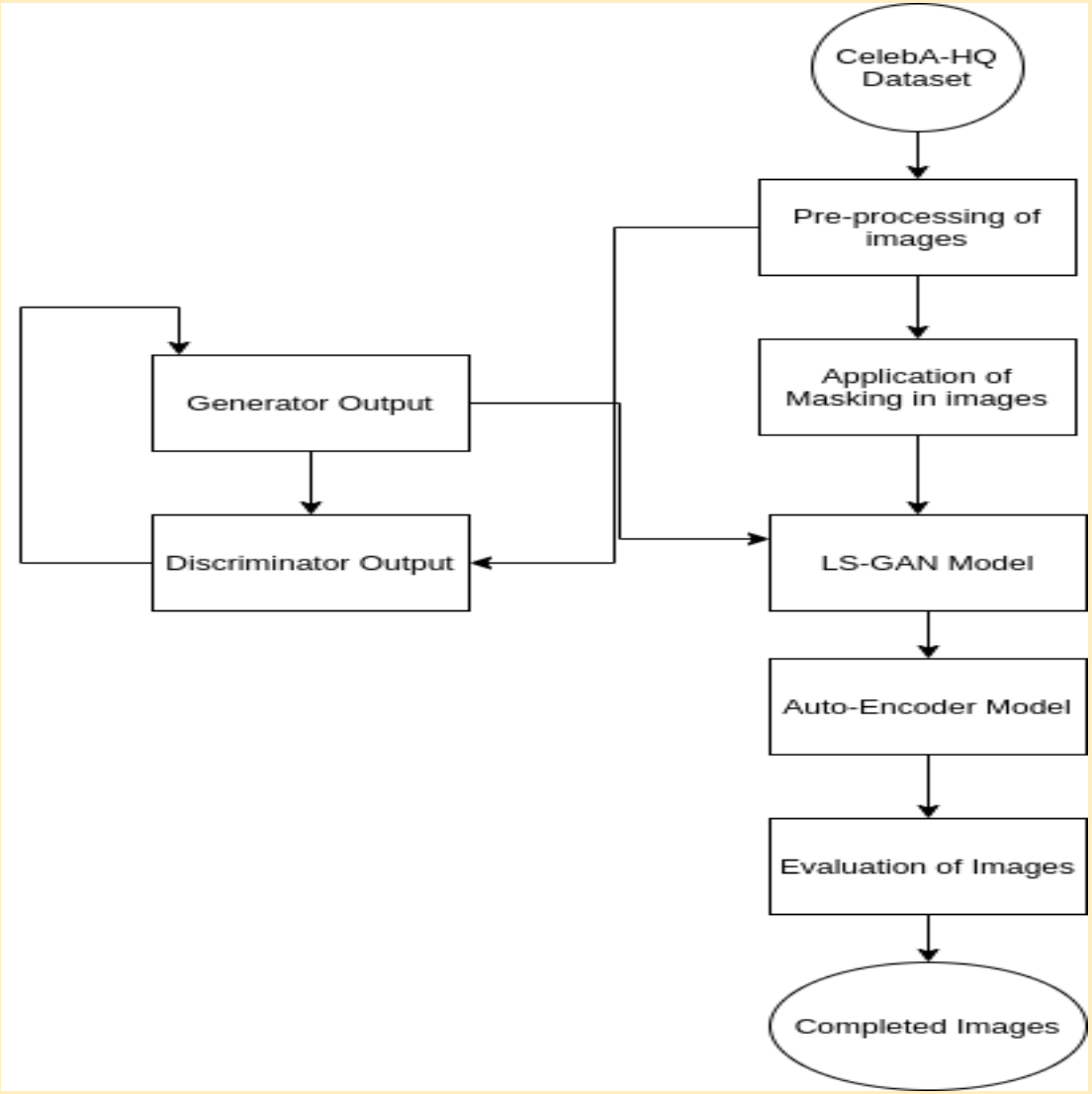

Figure 1: Proposed work-flow diagram

Methodology

- The Main feature that separates our work from the existing techniques is the combination of Least Squares Generative Adversarial Network along with AutoEncoder network both of which work toward reducing the Contextual and Perceptual Loss and provide realistically completed images.This is the Salient Feature of our thesis as no existing technique combines these 2 models for the same purpose.
- Image completion technique we quantify the existing loss functions and make a final function that helps in converging the input. Clipping additionally helps to reach the find the final distribution closest.
- $\hat{z} = argmin_z(Loss_{contextual}(z)) + \lambda Loss_{perceptual}(z)$
- $x_{completed} = Mask * y + (1 - Mask) * (Y_{AutoEncoder})$

Methodology

Generative Image Inpainting basically is combination of the following tasks together :-

- Analysing and Converting the Given Incomplete Image and generating Feature Maps and Importance Matrix.
- Using LSGAN to generate the image closest to the importance matrix
- Minimizing the Contextual Loss
- Using Auto-Encoders to generate the final completed image

Working of the model

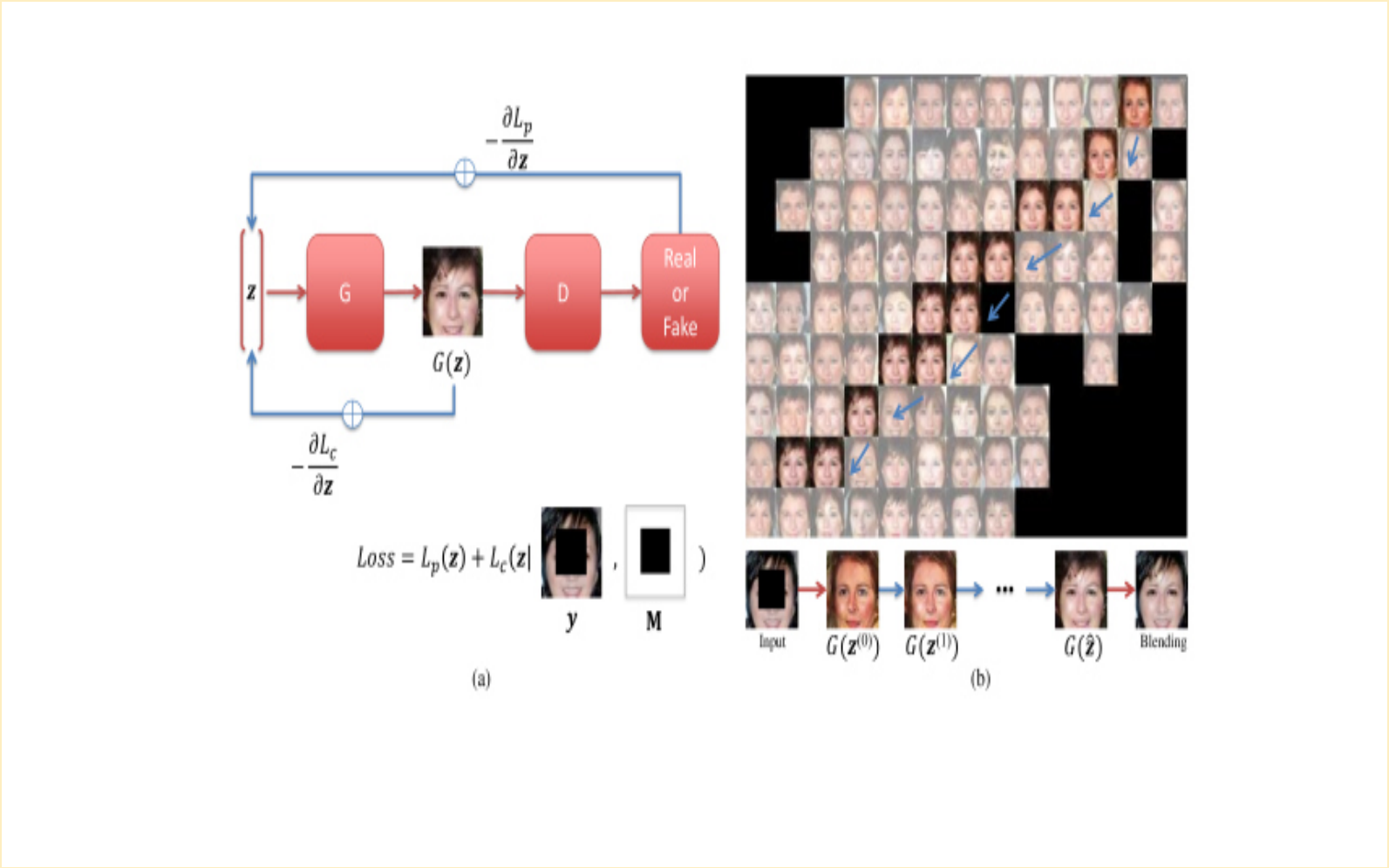

Figure 2: Proposed Image Completion Architecture in "Semantic Image Inpainting with Perceptual and Contextual Losses." using Perceptual and Contextual Losses Chen and Hu (2018) [11]

Combined Model Architecture

Layer	Size/Stride/Pad	Levels	Output
Generator Conv2DT 1024/lrelu/drop	32,3,3	1	4x4x1024
Generator Conv2DT 512/lrelu/drop	32,3,3	1	8x8x512
Generator Conv2DT 256/lrelu/drop	32,3,3	1	16x16x256
Generator Conv2DT 128/lrelu/drop	32,3,3	1	32x32x128
Encoder		4	1x8092
Decoder		4	64x64x3

Figure 3: Proposed Model Architecture with details of Layers

Implementation Details

- We implement our model using TensorFlow 1.5.
- We do not apply layer normalization as it tends to destabilize the LSGAN. We use Leaky-Relu Activation function which is the most effective in this case along with Adam Optimization[15].
- The LSGAN has been trained for 100,000 epochs to provide sharp results
- AutoEncoder Model has been trained till 2000 epochs. Batch Normalization has been used in the AutoEncoder Model to prevent the extent of overfitting.

LSGAN

- The benefits of LSGAN [25] is that unlike the normal Generative Adversarial Networks, it penalizes samples even when they are correctly converged.
- Generator
  - Generator model takes an input of 100 latent points. It provides an output of 64*64*3.
  - This output is similar to a fake generated image and this is passed to the discriminator which tries to classifies this as real or fake. This way, Contextual loss is reduced by the generator.
- Discriminator
  - Discriminator model is a normal neural network that aims at binary classification of the image.
  - This is usually trained separately but can also be trained along with the generator.

LSGAN

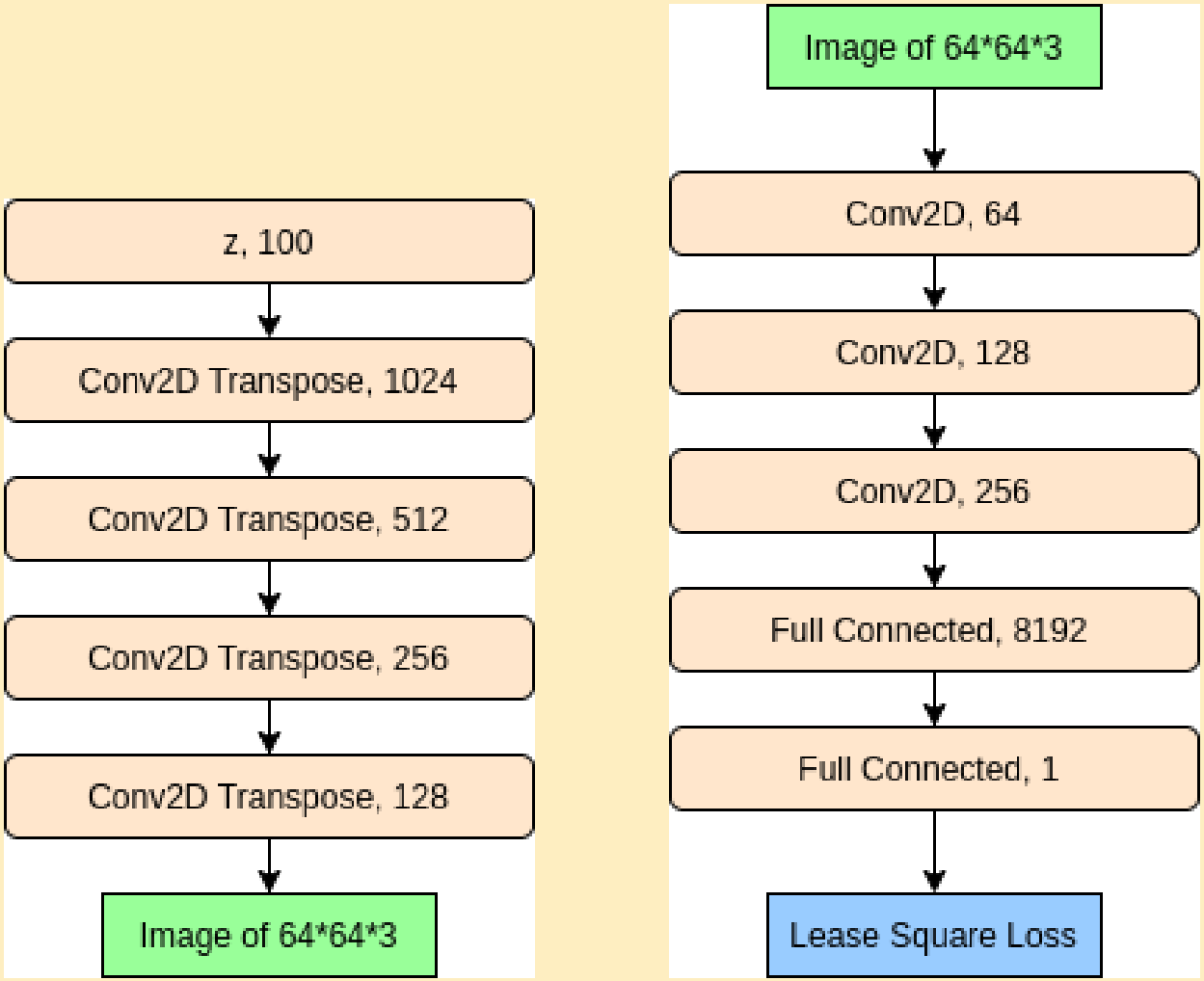

Figure 4: Generator and Discriminator Architectures

Proposed Auto-Encoder Model

- This model takes input in the form of image ie. 64*64*3 and responds with a new and refined image of the same dimensions ie. 64*64*3. This includes 2 parts and these are Encoder and Decoder.
- The function of Encoder part is to break the image of (64,64,3) into linear dimensions which means basically encoding all the data of the image to a relavant layer of size 8192. Now, all of these 8192 features are picked up by the Decoder Model which tries to recreate a refined image by using UpSampling.

Proposed Auto-Encoder Model

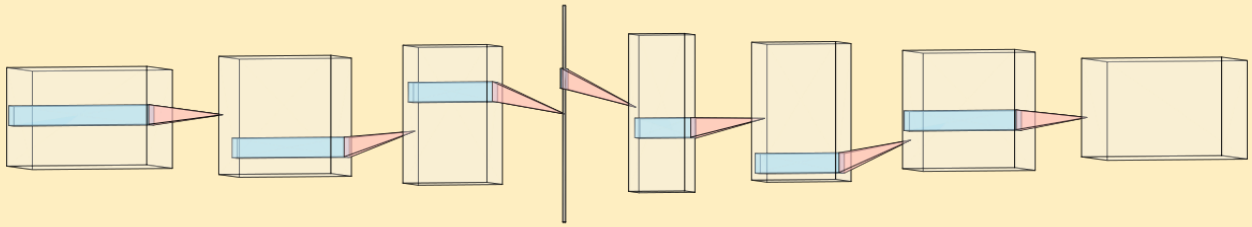

Figure 5: Proposed Auto-Encoder Architecture

Dataset and Preprocessing

The dataset used to train the model is the CelebA Dataset [17] which contains images of Faces of a wide variety of images. It consists of more than 200,000 images of celebrity faces that provides the vast feature scales

Before feeding the Images into our LSGAN model,the images are first normalized in the range of (-1 , 1).Out of these 200,000 images, only 20,000 images have been used to train and test the models.

Understanding Loss Functions

- Contextual Loss
  - Makes sure that the generated pixels by the generator (G(z)) are as similar as possible to the actual missing or uncorrupted pixels.[23]
  - $Loss_{contextual} = ||M.G(z) - M.y||_1$
- Perceptual Loss
  - The main focus of Perceptual Loss is to ensure that the output looks real. This means that the output has to be near the ground truth.
  - $Loss_{perceptual} = log(1 - C(G(z)))$
- Total Loss
  - $Loss_z = Loss_{contextual}(z) + QLoss_{perceptual}$

Understanding Evaluation Metric

- Peak Signal to Noise Ratio(PSNR)[10.]
  - It is measured in Decibels(dB). The higher the PSNR, the better the image has been recreated.
  -

$$MSE = \frac{1}{mn} \sum_{m} \hat{a} \sum_{n} (x_{mn} - y_{mn})^2$$

m number of real pixels

n number of generated pixels

x_{mn} pixel value

y_{mn} pixel value

$$PSNR(x, y) = \frac{10 \log_{10}(\max(\max(x), \max(y))^2}{|x - y|^2}$$

Understanding Evaluation Metric

- Structural Similarity Index(SSIM)[10.]
  - This metrix can depend on the complexity of the data. It can depend on constrast(C),structural term(S), luminance(I).
  - $SSIM(x, y) = [C(x, y)]^a * [S(x, y)]^b * [CIx, y)]^c$

Glimpses of training

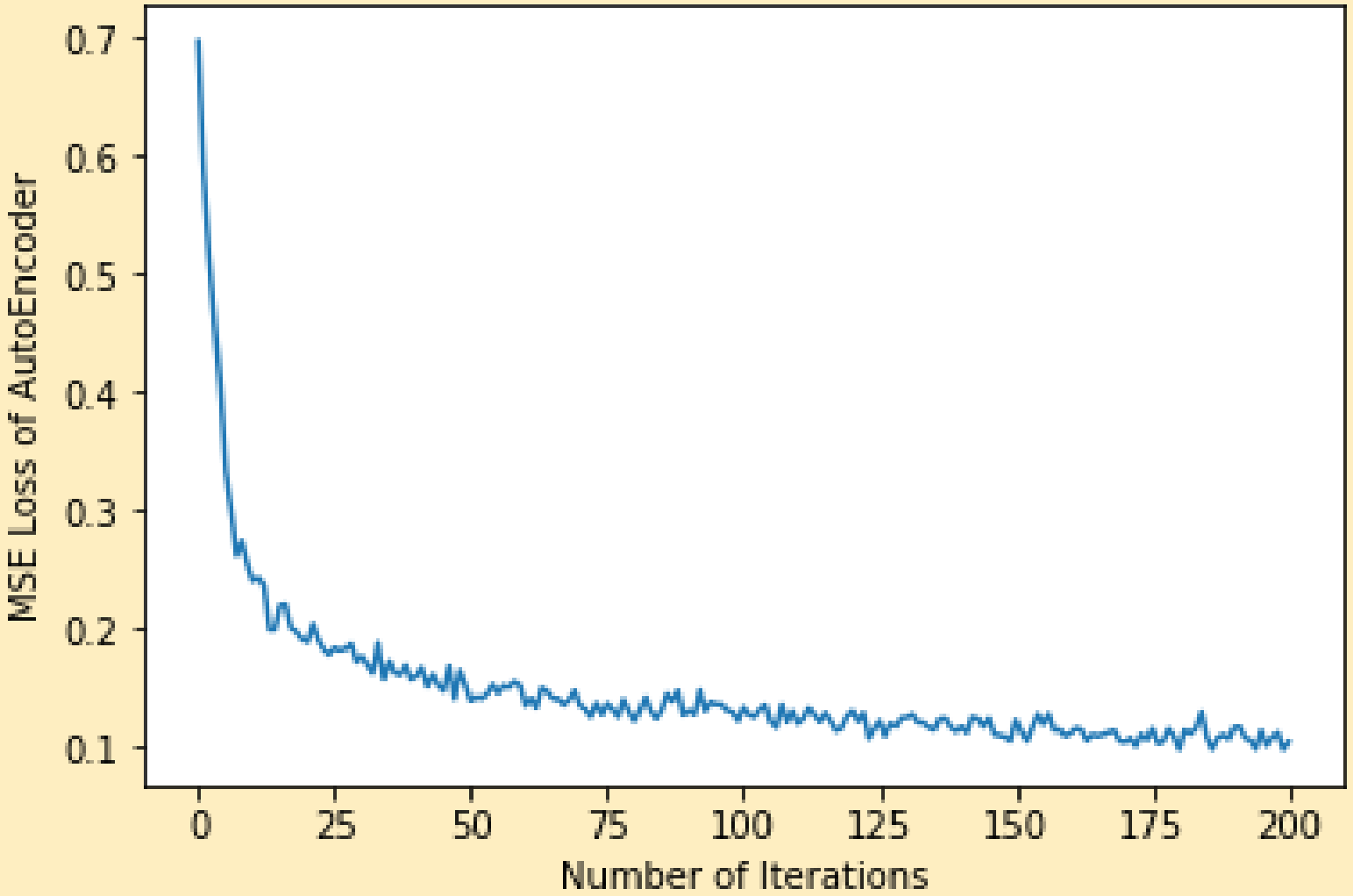

Figure 6: Training of the AutoEncoder Network

Convergence of Total Loss of LSGAN

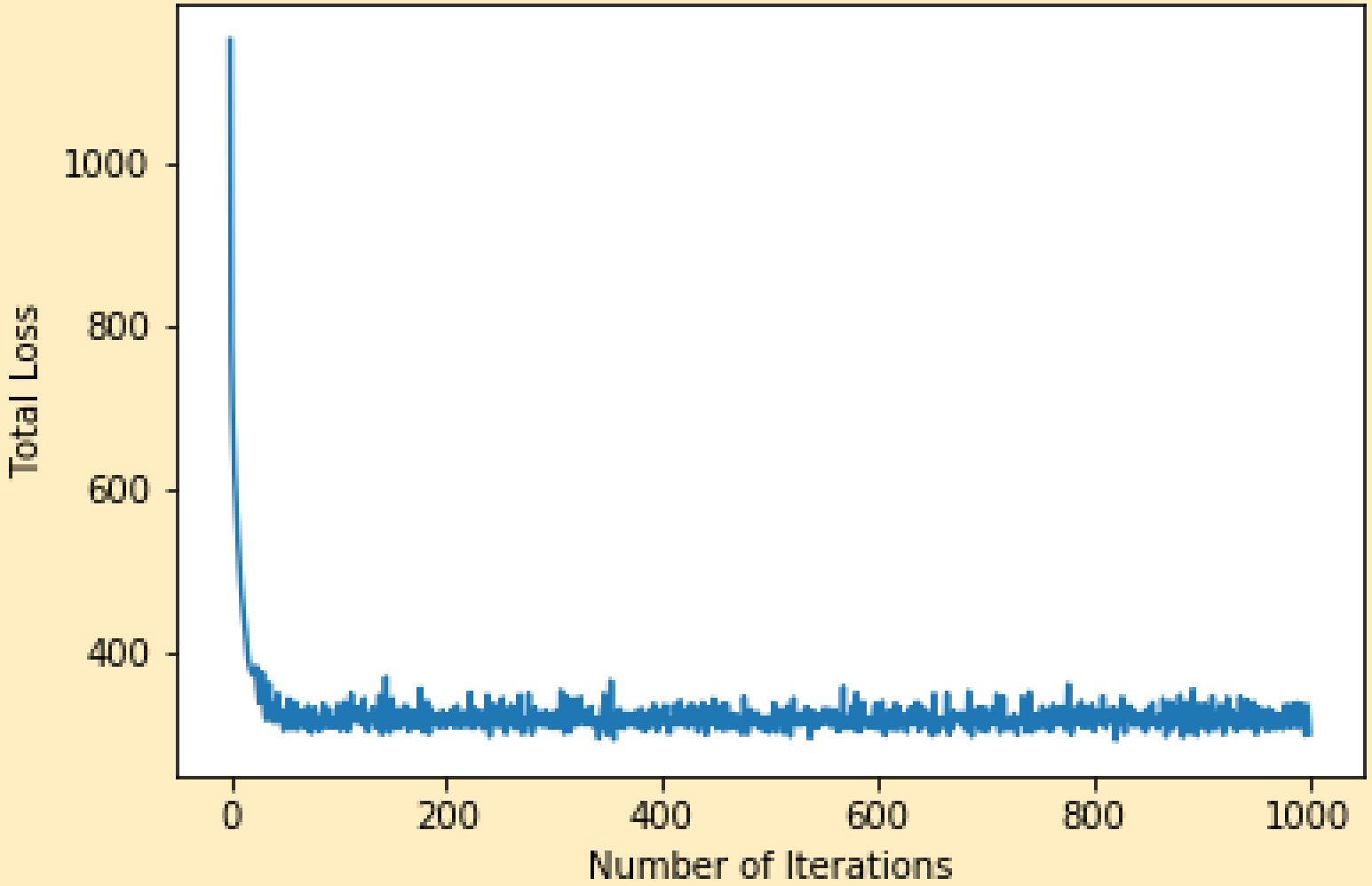

Figure 7: Convergence of Total Loss of LSGAN to find the most similar image to the given image

Results

We get the final result through the joint model of LSGAN and AutoEncoder Network which separates our work from the existing techniques.

Inspection of Metrics		
Model	**SSIM**	**PSNR**
Context Encoders	0.872	22.85
Pluralistic Inpainting	0.856	21.79
LSGAN	0.817	19.30
LSGAN+AutoEncoder	**0.883**	**22.30**

Figure 8: Comparison of our LSGAN+AutoEncoder model with other established and existing methods

Results

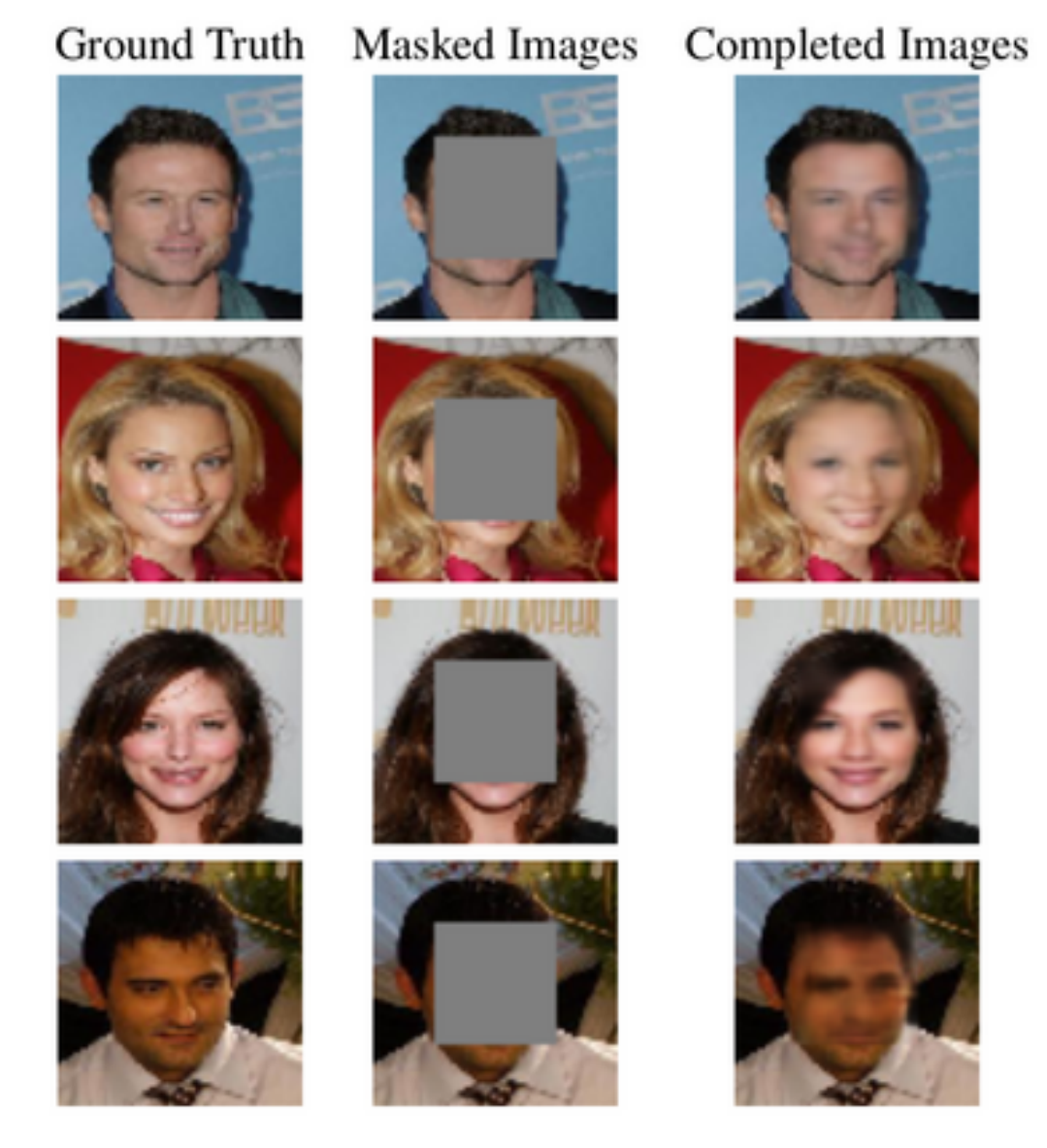

Figure 9: Results of our Model on our Test set

Results

- This noise produced by LSGAN has an effect on the structural similarity of the generated images which is around 0.80 - 0.82 along with a PSNR in the range of 0.20-0.22.
- When this is passed to the trained AutoEncoder model, we see a spike in the SSIM as well as PSNR which is around 0.87-0.89 and 0.21-0.23 respectively.
- The Existing Industry Standards[5] provide the Structural Similarity 0.85 and 0.87 and a PSNR near 0.21 and 0.22 [17]. So, this model performs at par with the current industry standard and in some cases crosses the set bar.
- we saw an improvement in the image quality by around 1.5% - 3% when compared to Pluralistic inpainting and Context Encoders.

Conclusion

- In our thesis, we present LSGAN + AutoEncoder architecture which is an end to end network model, for the purpose of completion of images by generative inpainting.
- The potency of LSGAN can be demonstrated from the decrease in the error rates [16] and an overall increase in the accuracy.
- The use of AutoEncoder enables the encoding of spatial and temporal features of the input images which in turn helps to enhance the quality of generated images.

Conclusion

- LSGAN solves the problem of vanishing gradients encountered and provides a guarantee of better quality image generation and stable training and decrease in loss.
- Conditional dependence is given by AutoEncoder network which corrects if there are some outlying samples generated by the network.

Future Scope

- We intend to extend our existing model not only for completion of faces but for different domains of images present.
- We plan to implement the model on a dataset with more extensive range of facial features to ensure even better results .
- The input images used for training and testing have been taken in a consistent environment. We intend to implement the model on a more noisy environment and check the accuracy.

References

1. Bank, D., Koenigstein, N. and Giryes, R.: 2020, Autoencoders.
2. Bao, J., Chen, D., Wen, F., Li, H. and Hua, G.: 2017, Cvae-gan: Fine-grained image generation through asymmetric training, Proceedings of the IEEE Interna- tional Conference on Computer Vision (ICCV).
3. Barnett, S. A.: 2018, Convergence problems with generative adversarial networks (gans), CoRR abs/1806.11382. URL: http://arxiv.org/abs/1806.11382

References

4. Chen, Y. and Hu, H.: 2018, An improved method for semantic image inpainting with gans: Progressive inpainting, Neural Processing Letters 49, 1355â1367.
5. Dong, H. and Yang, Y.: 2019, Towards a deeper understanding of adversarial losses, CoRR abs/1901.08753. URL: http://arxiv.org/abs/1901.08753
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: 2014, Generative adversarial nets, in Z. Ghahra- mani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger (eds), Ad- vances in Neural Information Processing Systems 27, Curran Associates, Inc., pp. 2672â2680.
7. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A.: 2017, Improved training of wasserstein gans.

References

8. Haoyu Ren, J. L. and El-khamy, M.: 2018, Dn-resnet: Efficient deep residual network for image denoising.
9. He, K. and Sun, J.: 2014, Image completion approaches using the statistics of similar patches, Pattern Analysis and Machine Intelligence, IEEE Transactions on 36, 2423â2435.
10. Hore, A. and Ziou, D.: 2010, Image quality metrics: Psnr vs ssim, International Conference on Pattern Recognition .
11. Huang, J., Kang, S., Ahuja, N. and Kopf, J.: 2014, Image completion using planar structure guidance.

References

12. Ji, J. and Yang, G.: 2020, Image completion with large or edge-missing areas, Algorithms 13(1). URL: https://www.mdpi.com/1999-4893/13/1/14
13. Jiang, J., Kasem, H. M. and Hung, K.: 2019, Robust image completion via deep feature transformations, IEEE Access 7, 113916â113930.
14. Kim, J., Song, S. and Yu, S.: 2017, Denoising auto-encoder based image enhance- ment for high resolution sonar image, 2017 IEEE Underwater Technology (UT), pp. 1â5.

References

15. Kingma, D. P. and Ba, J.: 2017, Adam: A method for stochastic optimization.
16. Liu, G., Yang, R., Li, S., Shi, Y. and Jin, X.: 2018, Painting completion with generative translation models, Springer
17. Liu, Z., Luo, P., Wang, X. and Tang, X.: 2015, Deep learning face attributes in the wild, Proceedings of International Conference on Computer Vision (ICCV).
18. Mao, X., Li, Q., Xie, H., Lau, R. Y. K. and Wang, Z.: 2017, Least squares gener- ative adversarial networks.

References

19. Pathak, D. and Krahenbuhl, P.: 2016, Context encoders: feature learning by in- painting.
20. Salakhutdinov, R. and Hinton, G.: 2009, Deep boltzmann machines, Vol. 5 of Pro- ceedings of Machine Learning Research, PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, pp. 448â455. URL: http://proceedings.mlr.press/v5/salakhutdinov09a.html
21. Thompson, R.: 1 April, 2019, Action detection using deep neural networks: Prob- lems and solutions. URL: https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529

References

22. Yi, K., Guo, Y., Fan, Y., Hamann, J. and Wang, Y. G.: 2020, Cosmovae: Varia- tional autoencoder for cmb image inpainting.
23. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X. and Huang, T. S.: 2018, Generative image inpainting with contextual attention, CoRR abs/1801.07892. URL: http://arxiv.org/abs/1801.07892
24. Zarif, S., Faye, I. and Awang Rambli, D.: 2014, Image completion: Survey and comparative study, International Journal of Pattern Recognition and Artificial In- telligence 29, 1554001.

References

25. Zhang, Z., Li, M. and Yu, J.: 2018, On the convergence and mode collapse of gan, pp. 1â4.
26. Zhao, G., Liu, J., Jiang, J. and Wang, W.: 2017, A deep cascade of neural net- works for image inpainting, deblurring and denoising, Multimedia Tools and Ap- plications 77.
27. Zheng, C., Cham, T.-J. and Cai, J.: 2019, Pluralistic image completion, Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1438â1447.

Thank You!