NATIONAL UNIVERSITY OF COMPUTER & EMERGING SCIENCES
DEPARTMENT OF COMPUTER SCIENCE

IMAGE RESTORATION USING TRANSFORMERS

MOIZ FAROOQUI 23K-8024

MS (DS)

2025

SUPERVISOR: Dr. Maria Siddiqua

1. Abstract

Sand Image restoration has become a crucial task in the fields of deep learning and computer vision . During this era of technology where conventional cars are becoming outdated and we are moving into the era of Autonomous vehicles with major technological advancements, Image restoration becomes a fundamental task as self-driving vehicles relies on the images to gather information and make decisions. Image restoration in also vital in Satellites which relies on images for various applications in Environmental Planning, Google Maps, and Disaster Response. These applications make Image Restoration a very pivotal task. Traditional image restoration techniques that use CNNs and Generative Adversarial Networks (GANs), have shown state of the art performance but still struggle to capture global context and long-range dependencies . In this thesis we have implemented a Pix2Pix GAN based baseline model for the restoration of sand effected images . This research delves into the application of transformer-based models for image restoration, using their self-attention mechanism to enhance feature learning and improve image quality. In this research we proposed the implementation of vision transformer models for image restorations and evaluate their performance against the baseline deep learning approaches. A synthetic dataset is prepared by adding sand effect to the images to simulate real world sand degradation. A baseline model is implemented on the dataset to recover and restore clean images from the degraded images . After that a vision transformer-based model will be applied to the dataset and the performance will be compared and evaluated with the baseline model . This report presents a brief literature review, research methodology(dataset preparation and implementation of baseline model) and the proposal for the future work to implement vision transformer-based model for image restoration tasks.

Table of Contents

Table of Figures

Table of Tables

Chapter 1

Introduction

1.1. Brief Introduction:

With the rapid advancement of autonomous vehicles and satellite imaging the need for image restoration is significantly increasing. In the real-world scenarios, images are often degraded due to several environmental issues . In high wind areas, sand and dust storms often degrade images which are very critical in the application of self-autonomous cars and satellite imaging as the decisions are based on these images. This degraded images effects the performance of autonomous systems, therefore image restoration is a very pivotal task to maintain continuous and reliable operation. Image Restoration means recovering or rebuilding high quality images from the degraded ones (e.g., noise, blur, rain drops). Due to the ill-posed nature, it is a highly challenging problem that usually requires strong image priors for effective restoration [1]. We aim to design and implement an efficient model that can recover clean real-world images from the sand degraded images .

Current research on image dedusting mainly includes methods based on image processing and methods based on atmospheric scattering models. Although certain results have been achieved in image dedusting, there are still some problems [7] .Recent advancements have been made and models based on Generative Adversarial Networks (GANs), have shown great results and opened new doors in the advancement of image restoration . Generative adversarial networks (GANs) were originally proposed by Goodfellow et al. [8] and have achieved great success in the field of image generation. However, such networks still require paired datasets for training. Recently, Zhu et al. [9] proposed a cycle-consistent generative adversarial network (CycleGAN) for image translation problems. GANs can learn complex mappings but they are unable to learn the long-range dependencies and often work on paired datasets.

1.2. Problem Statement:

In the real-world scenarios, images are often degraded due to several environmental issues . In high wind areas, sand and dust storms often degrade images which are very critical in the application of self-autonomous cars and satellite imaging as the decisions are based on these images. This degraded images effects the performance of autonomous systems, therefore image restoration is a very pivotal task to maintain continuous and reliable operation.

1.3. Motivation of Work:

These limitations have opened the doors to the exploration of Transformer based models for image restoration tasks . Due to their self-attention mechanism and capturing long range dependencies they are currently the talk of the town . Especially vision transformers for image restoration tasks are gaining popularity. The ViT is a deep neural network architecture for image recognition tasks based on the Transformer architecture initially developed for natural language processing tasks. The main idea behind ViT is to treat an image as a sequence of tokens (typically, image patches). Then, the transformer architecture is used to process this sequence. The transformer architecture, on which ViT is based, has been adapted to many tasks and is effective in many of them, such as image restoration and object detection [2]. [10] introduced an effective hybrid architecture that takes advantage of CNN and recent ViT for sand image restoration.

The motivation for this research is driven from the increasing demand for high-quality image restoration in real-world scenarios, particularly in autonomous vehicle systems. A major application of image restoration is in self-driven cars. for safe and accurate navigation, Real-time, high-quality visual data is very important. To detect objects, obstacles, turns and road signs self-driving cars are based on cameras and sensors. Weather conditions such as rain, fog, motion blur, or darkened environments can degrade image quality, which leads to inaccurate decision-making. With the help of transformer-based image restoration approaches into autonomous driving systems, we can refine the clarity of visual data, improve object detection accuracy, and increase safety while driving. One of the major benefits of this research is that this model can not only be applied in self-autonomous vehicles but in several other domains, including satellite imaging, drone images, and security surveillance.

1.4. Scope of Work:

This research will be carried out in three stages. In the first stage synthetic datasets are prepared by adding sand like noise to make real world images as sand degraded images to simulate the real-world degradation . In the second stage a baseline model consisting of GANs (Pix2Pix and CycleGAN) based architecture. Then in the third stage we will implement the vision transformer-based image restoration model to address the limitations of GANs.

Chapter 2

Related Work

Image Restoration is a very important task in the field of computer vision. The fundamental objective is to recover high quality images from the input images that are compromised by noise, blur and low resolution. Traditional methods for Image restoration rely on convolutional neural networks (CNNs) and GANs, but transformer architectures with their ability to capture long-range dependencies and contextual information put them above all.

In 2024, An improved pix2pix generative adversarial networks for sand-dust image enhancement was proposed. It enhances the original Pix2Pix model. These improvements help the model better capture fine details and preserve the overall semantics of restored images, making it more effective for sand-dust image enhancement [3].

The CycleGAN framework [10], and its use in Unsupervised Image Dedusting [5], addresses the issue of unpaired training data. It introduces cycle consistency and identity loss to maintain color and structure, making it valuable in real-world scenarios where clean reference images are unavailable.

USIR-Net [9] advances unsupervised methods further by incorporating attention mechanisms and multi-scale feature fusion. It focuses on key structural details while adapting to various levels of sand-dust interference, delivering strong results on real-world datasets.

To improve both local and global feature learning, SANDFORMER [4] combines CNNs with Vision Transformers using a gated fusion mechanism. This hybrid model leverages the strengths of CNNs for texture recovery and Transformers for global context, achieving better performance across different sandstorm intensities.

In the recent time Vision Transformers have gained popularity and emerged as an effective alternative. In 2023, Ali et al. introduced vision transformers for the task of image restoration. The ViT is a deep neural network architecture for image recognition tasks based on the Transformer architecture initially developed for natural language processing tasks. The main idea behind ViT is to treat an image as a sequence of tokens (typically, image patches). Then, the transformer architecture is used to process this sequence [2]. The Author also compares the performance of Vision Transformer with the baseline models and achieved SOTA performance.

It can be clearly stated that in the image restoration tasks, Transformer based models have outperformed traditional techniques like CNN and diffusion models. Traditional methods for Image restoration rely on convolutional neural networks (CNNs) and Diffusion models, but transformer architectures with their ability to capture long-range dependencies and contextual information put them above all. Transformer models can further be optimized for their high computational complexity. This research will focus on capturing real images from the sand like images degraded due to the harsh weather conditions like dust storms.

This research will play a significant role in the advancement of navigation and surveillance systems where decisions are based on real time images. Noise often degrades the image causing wrong decision making. By restoring such images, we can have a great impact and improve the overall performance.

Chapter 3

Literature Review

3.1. Image Restoration:

Image Restoration is one of the most fundamental tasks in deep learning and computer vision. The basis purpose of image restoration is to restore real images from the degraded images due to any cause.

Common types of degradation are as follows:

Degradation Type	Cause/Source	Example
Noise	Sensor limitations, transmission	Gaussian, salt-and-pepper
Blur	Camera shake, defocus	Motion blur, out-of-focus blur
Compression Artifacts	Lossy image formats (e.g. JPEG)	Blocking, ringing
Occlusion / Masking	Dust, rain, scratches	Text overlays, stains
Low Resolution	Down sampling, poor hardware	Pixelated images

Table 1. Types of Degradation

Following are the most common applications of Image restoration:

- **Medical Imaging**: Denoising MRI or CT scans.
- **Surveillance**: Enhancing low-light or degraded footage.
- **Photography**: Removing blurs, noise, or weather artifacts.
- **Autonomous Driving**: Enhancing visual clarity under harsh conditions (rain, dust, fog)**.**
- **Satellite/Remote Sensing**: Restoring clarity in Earth observation imagery.

3.1.1. Models for Image Restoration:

The most used models for Image restoration tasks are as follows:

Model Type	Description
Convolutional Neural Networks (CNNs)	Learn a mapping from degraded to clean images.
Autoencoders / U-Nets	Encoder-decoder structures with skip connections for restoration.
GANs (e.g., Pix2Pix, Cycle GAN)	Learn to generate realistic output using adversarial loss.
Transformers (e.g., SwinIR, Uformer)	Capture long-range dependencies, perform better on complex degradations.

Table 2. Common Models for Image Restoration

3.2. Generative Adversarial Networks (GANs):

Generative Adversarial Networks (GANs) were introduced by Ian Goodfellow and his colleagues in 2014. GANs are a class of neural networks that autonomously learn patterns in the input data to generate new examples resembling the original dataset [11].

GANs are based on two main blocks

- Generator (Try to create real type images with noisy inputs)
- Discriminator (Compares the restored image with a real clean image.)

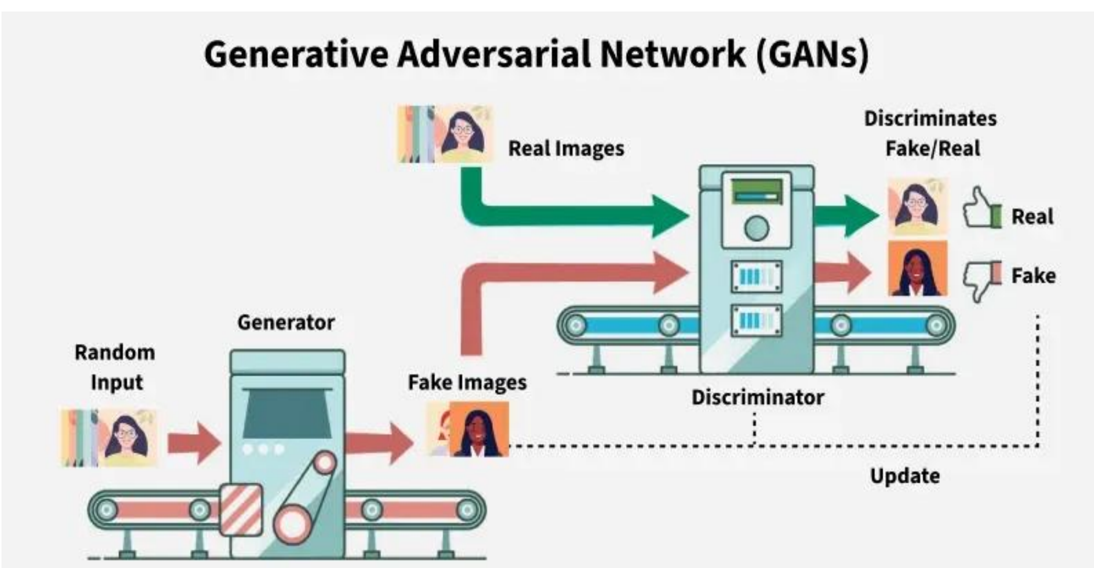


Figure 1. Architecture of GANs

Following are the major types of GANs:

AN Variant	Purpose in Image Restoration	Example Models	Key Feature
Pix2Pix	Paired image-to-image translation (needs clean & degraded image pairs)	Facade to building mapping	Uses conditional GAN and L1 loss
CycleGAN	Unpaired image translation (works without clean-degraded pairs)	Style transfer, domain adaptation	Cycle consistency loss
DeblurGAN	Motion blur removal	GoPro deblurring	Combines perceptual and adversarial losses
Denoising GAN	Removes noise (e.g., Gaussian, sensor)	DnCNN-GAN, Noise2Noise	Learns noise distribution
Inpainting GAN	Fills missing/occluded regions in images	Context Encoders, EdgeConnect	Context-aware generation
Super-Resolution GAN	Enhances low-resolution images	SRGAN, ESRGAN	Uses perceptual loss and deep residual blocks

Table 3. Types of GANs

3.2.1. Pix2Pix:

Pix2Pix is a conditional image-to-image translation architecture that uses a conditional GAN objective combined with a reconstruction loss. The conditional GAN objective for observed images x, output images y and the random noise vector z is [12].

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[log(1 - D(x, G(x, z))]$$

We augment this with a reconstruction term:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[||y - G(x, z)||_1]$$

and we get the final objective as:

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

3.2.2. Loss Functions:

- **Adversarial loss** in GANs measures the discrepancy between real and generated at distributions, guiding the generator to produce realistic samples by fooling the discriminator. Its purpose is to ensure that generated samples are indistinguishable from real ones, promoting the overall realism of the generated data.

- **Cycle consistency loss** in CycleGAN enforces that the translation from one domain to another and back again results in images like the original ones. This constraint aims to maintain semantic content during image translation, facilitating accurate and coherent cross-domain transformations.

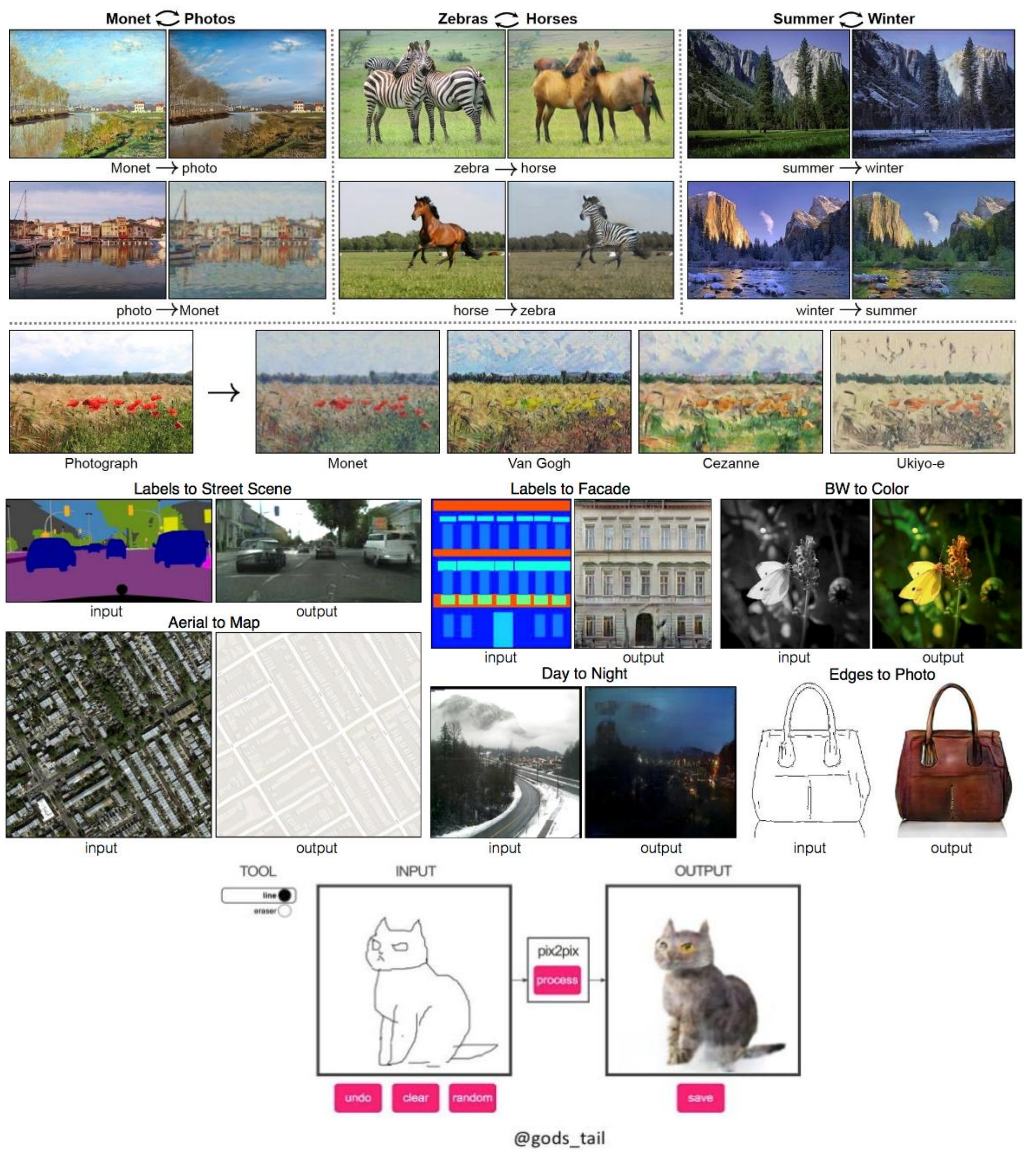


Figure 2. Applications of GANs

3.3. Diffusion Models:

Diffusion models are a class of generative models used in machine learning to create new data samples that resemble a given dataset. Unlike traditional models that generate data directly, diffusion models operate by gradually transforming a simple noise distribution into complex data through a series of steps [11].

- **Forward Process:** In forward process random noise is added at each step to make clean images noisy . This is being done to train how the noise can be removed.

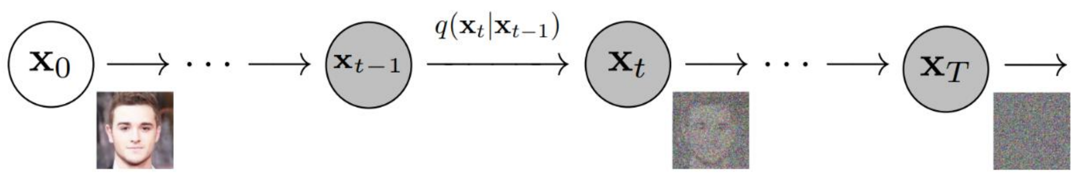


Figure 3. Forward Diffusion Process

- **Reverse Process:** In reverse process random noise is removed and the image is restored to the original form.

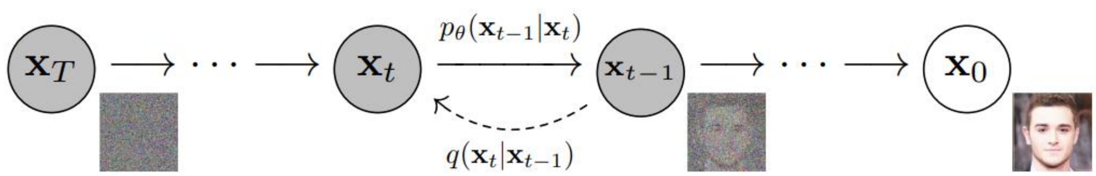


Figure 4. Reverse Diffusion Process

3.4. Transformers:

Transformer is a neural network architecture used for performing machine learning tasks particularly in natural language processing (NLP) and computer vision. In 2017 Vaswani et al. published a paper " Attention is All You Need" in which the transformers architecture was introduced. The article explores the architecture, workings and applications of transformers [11].

Figure 5. Transformer

Transformers consist of two major blocks:

- **Encoder:** gets the input and encodes it into a matrix representation of that input.

- **Decoder:** Decoder gets the output of the encoder as input decodes that matrix representation and gives the output as required

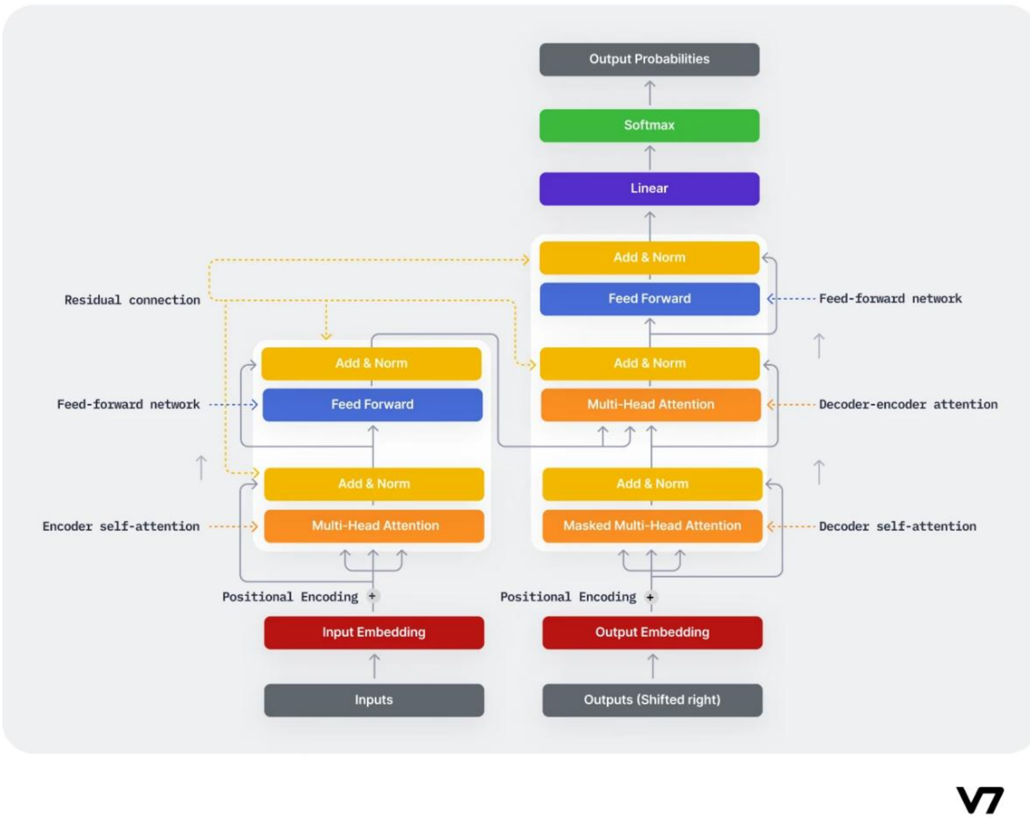


Figure 6. Transformer Architecture [14]

3.4.1. Vision Transformers:

In ViTs, images are represented as sequences, and class labels for the image are predicted, which allows models to learn image structure independently. Input images are treated as a sequence of patches where every patch is flattened into a single vector by concatenating the channels of all pixels in a patch and then linearly projecting it to the desired input dimension [14].

Let’s examine the vision transformer architecture step by step.

- Split an image into patches
- Flatten the patches
- Produce lower-dimensional linear embeddings from the flattened patches

- Add positional embeddings
- Feed the sequence as an input to a standard transformer encoder
- Pretrain the model with image labels (fully supervised on a huge dataset)
- Finetune on the downstream dataset for image classification

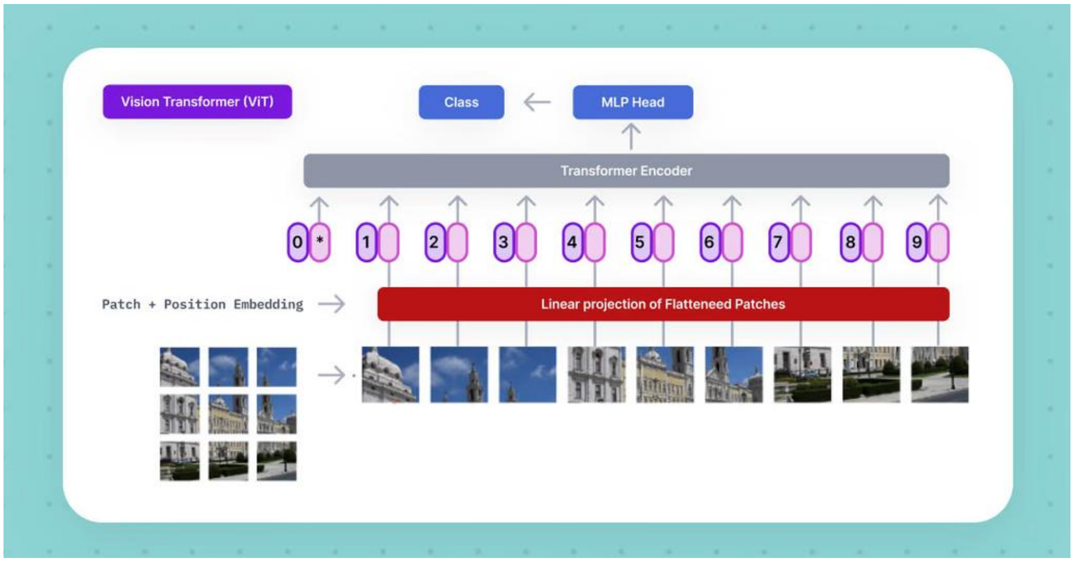


Figure 7. Architecture of Vision Transformer

Chapter 4

Methodology

Following Steps have been taken to implement the baseline model, and these steps will be followed for the implement of our model based on vision transformers

4.1. Data Preparation:

Google Streetview Dataset has been used for implementation of our models . It consists of around 23000 images of random streets and scenic images .

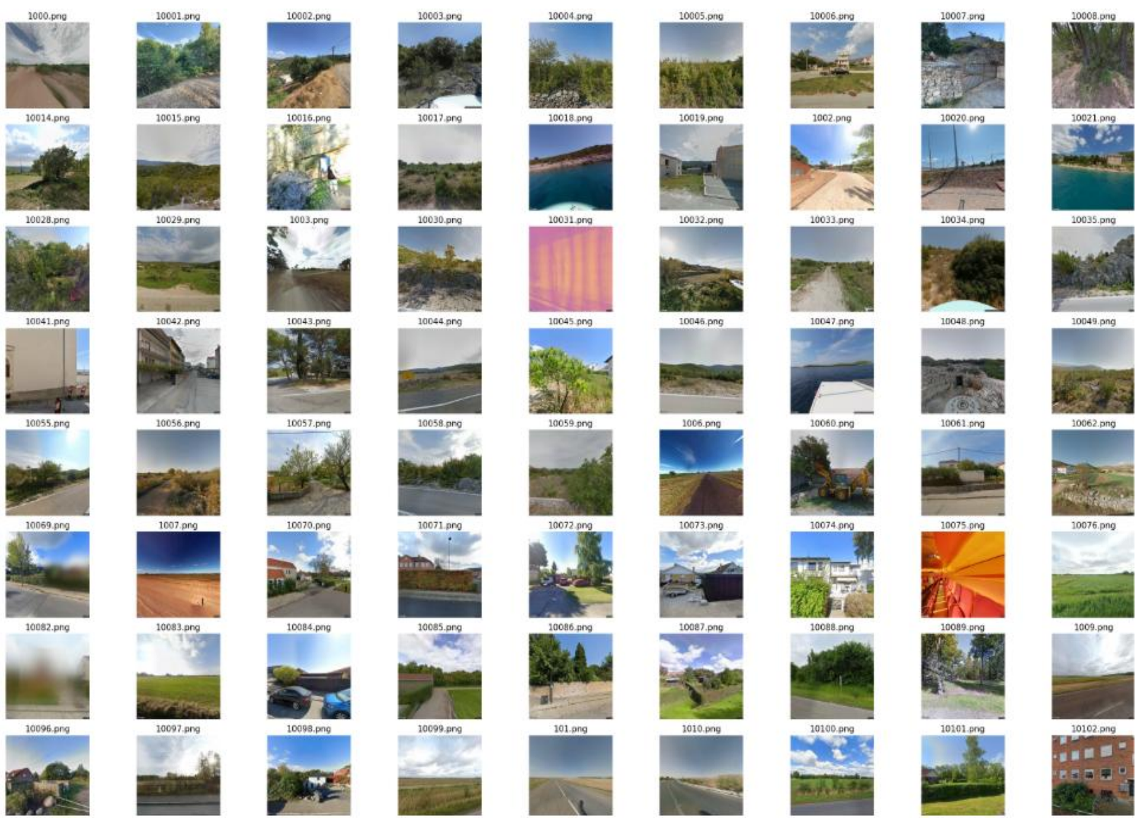


Figure 8. Street View Dataset

The data was divided into training(70%) and test(30%) dataset . The data has been synthesized by adding sand effect to make the images look like sand degraded images.

A sand like effect is added to the images by applying gaussian noise, gaussian blur and different shades of tints given in the below table.

S.no	RGB Array	HEX Code	Shade	Color
1	(204, 166, 122)	#CCA67A	Sand yellow / Light dusty sand	
2	(161, 74, 26)	#A14A1A	Rusty sand / Dusty reddish brown	
3	(198, 133, 61)	#C6853D	Dusty tan / Sandstone	
4	(242, 204, 102)	#F2CC66	Pale sand / Light sandy beige	
5	(217, 166, 64)	#D9A640	Golden sand / Rich sandy gold	
6	(166, 115, 26)	#A6731A	Earthy sand / Brownish dust	
7	(153, 102, 13)	#99660D	Dark sand / Dusty ochre	

Table 4. Tints for Sand Effect

The sand degraded images after adding noise are as follows:

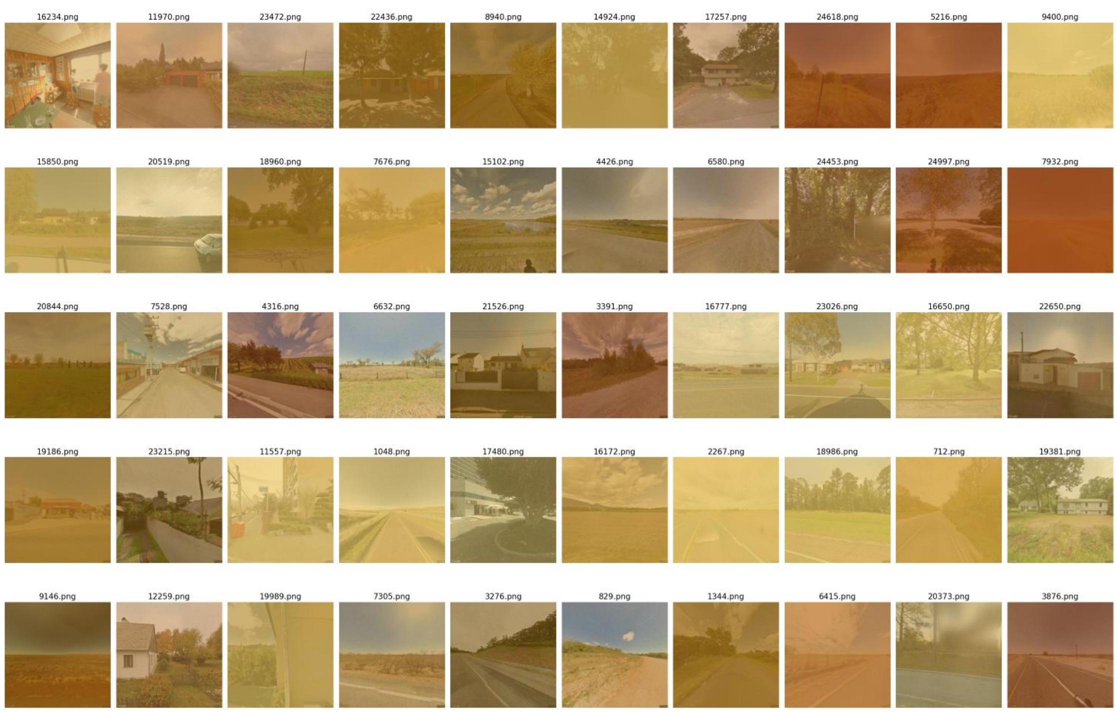


Figure 9. Sand Effected Images

The images are now merged together . same images (Clean and degraded ones) are placed together as our Pix2Pix model works on paired datasets.

Figure 10. Merged Images

This Dataset will now be used to train the baseline model and afterward test on test dataset to restore real world like images from the degraded ones.

4.2. Implementing Baseline Model (Pix2Pix):

In this model we have trained conditional generative adversarial network (cGAN) called pix2pix that learns a mapping from input images to output images, as described in [11] by Isola et al. (2017). pix2pix is not application specific—it can be applied to a wide range of tasks, including synthesizing photos from label maps, generating colorized photos from black and white images, turning Google Maps photos into aerial images, and even transforming sketches into photos.

As described in the pix2pix paper, you need to apply random jittering and mirroring to preprocess the training set.
Define several functions that:

- Resize each 256 x 256 image to a larger height and width—286 x 286.
- Randomly crop it back to 256 x 256.
- Randomly flip the image horizontally i.e., left to right (random mirroring).
- Normalize the images to the [-1, 1] range.

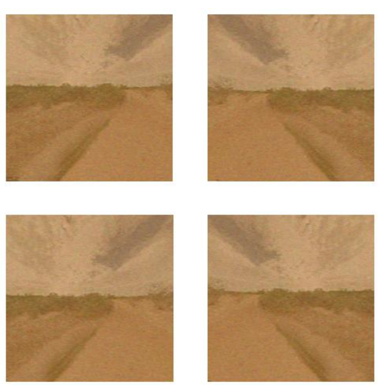


Figure 11. Picture after Data Augmentation

4.2.1. Build the Generator:

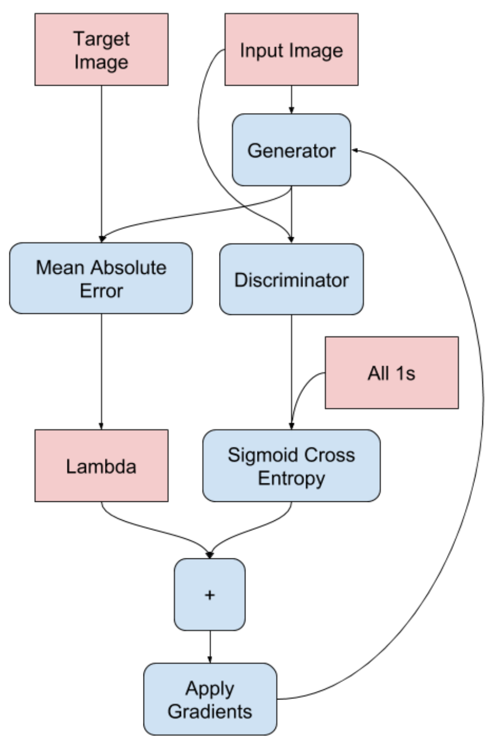
The generator of your pix2pix cGAN is a _modified_ [U-Net]. A U-Net consists of an encoder (downsampler) and decoder (upsampler)

- Each block in the encoder is: Convolution > Batch normalization > Leaky ReLU.
- Each block in the decoder is: Transposed convolution > Batch normalization -> Dropout (applied to the first 3 blocks) > ReLU
- There are skip connections between the encoder and decoder (as in the U-Net).

GANs learn a loss that adapts to the data, while cGANs learn a structured loss that penalizes a possible structure that differs from the network output and the target image, as described in the pix2pix paper.

- The generator loss is a sigmoid cross-entropy loss of the generated images and an array of ones.
- The pix2pix paper also mentions the L1 loss, which is a MAE (mean absolute error) between the generated image and the target image.
- This allows the generated image to become structurally like the target image.
- The formula to calculate the total generator loss is gan_loss + LAMBDA * l1_loss, where LAMBDA = 100. This value was decided by the authors of the paper.

The training procedure for the generator is as follows is as follows:



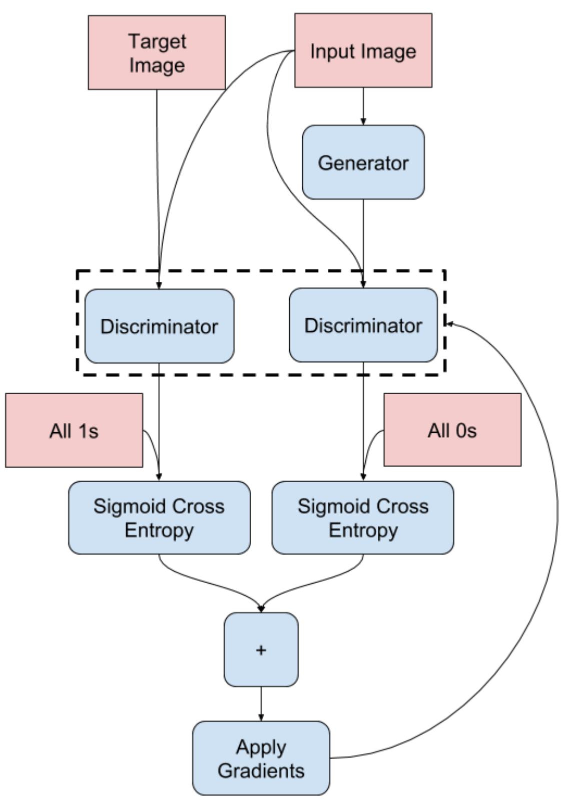
4.2.2. Build the Discriminator:

The discriminator in the pix2pix cGAN is a convolutional PatchGAN classifier—it tries to classify if each image *patch* is real or not real, as described in the pix2pix paper.

- Each block in the discriminator is: Convolution -> Batch normalization -> Leaky ReLU.
- The shape of the output after the last layer is (batch_size, 30, 30, 1).
- Each 30 x 30 image patch of the output classifies a 70 x 70 portion of the input image.
- The discriminator receives 2 inputs:
    - The input image and the target image, which should classify as real.
    - The input image and the generated image (the output of the generator), which it should classify as fake.
    - Use tf.concat([inp, tar], axis=-1) to concatenate these 2 inputs together.

The discriminator_loss function takes 2 inputs: real images and generated images.

- real_loss is a sigmoid cross-entropy loss of the real images and an array of ones(since these are the real images).
- generated_loss is a sigmoid cross-entropy loss of the generated images and an array of zeros (since these are the fake images).
- The total_loss is the sum of real_loss and generated_loss.



4.2.3. Generate Images:

Write a function to plot some images during training.

- Pass images from the test set to the generator.
- The generator will then translate the input image into the output.
- The last step is to plot predictions.

Figure 12. Testing the function

4.2.4. Training:

- For each example input generates an output.
- The discriminator receives the input_image and the generated image as the first input. The second input is the input_image and the target_image.
- Next, calculate the generator and the discriminator loss.
- Then, calculate the gradients of loss with respect to both the generator and the discriminator variables(inputs) and apply those to the optimizer.

Chapter 5

Proposed Work

While GANs based models have gained popularity but there are some limitations when learning the long range dependencies and the global context. Especially in sand effected image restoration tasks we require such models that can learn even minor details this is where transformer based models come in. Vision Transformers that can use self-attention mechanism to learn global context can be the solution.

5.1.1. Limitation of Baseline Model:

S. No.	Limitation	Description
1	Blurring and Artifacts	Visual distortions appear in areas with severe sand-dust, reducing image clarity.
2	Loss of Structural Details	Fine structures and edges are not preserved due to small receptive fields of convolutional layers.
3	Inconsistent Texture Generation	Distant or small objects often have unnatural or mismatched textures.

Table 5. Limitations of Baseline Model

The working and the literature review for the implementation of the Vision based model will be carried out in thesis 2.

Chapter 6

Results

6.1. Results for Baseline Model:

Following performance was noticed on the test dataset:

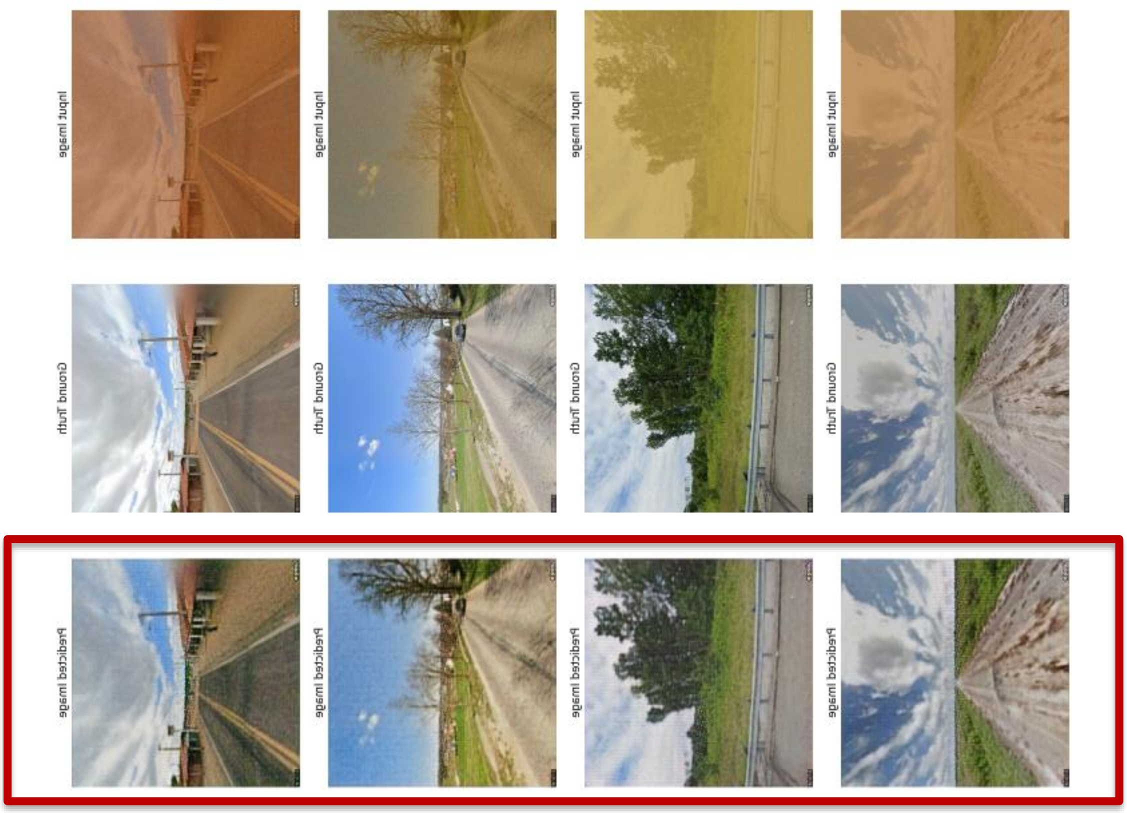


Figure 13. Restored Images

6.2. Evaluation Metrices:

Following Evaluation Metrices were analysed

- **PSNR(Peak Signal to Noise Ratio) :** It calculates the ratio of noise between the restored image and the real image.

- **SSIM(Structural Similarity Index Metric)** It considers luminance, contrast, and structure when comparing restored image and the real image.

Metric	Value	Quality Level	Implication
PSNR	54.55 dB	Excellent	Very low noise and high pixel accuracy
SSIM	0.9756	Near-perfect match	Strong preservation of structure and details

Table 6. Evaluation Metrices

Chapter 7

Conclusion

It has been seen that GANs are effective in image restoration tasks and when we applied the baseline model to our sand effected images it restores and were efficient on our test dataset and gave us promising results but still have some limitations . We look forward to implementing transformer based models comprising vision transformers to see the difference of restored images

Chapter 8

References

[1] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient Transformer for High-Resolution Image Restoration," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5728-5739.

[2] M. Ali, B. Benjdira, A. Koubaa, W. El-Shafai, Z. Khan, and W. Boulila, "Vision Transformers in Image Restoration: A Survey," *Sensors*, vol. 23, no. 5, p. 2385, 2023. [Online]. Available: https://doi.org/10.3390/s23052385.

[3] Y. Zhang, S. Liu, and H. Li, "An improved pix2pix generative adversarial network for sand-dust image enhancement," *Journal of Image and Graphics*, vol. 10, no. 4, pp. 324-330, 2022.

[4] X. Chen, Z. Wang, and L. Zhang, "SANDFORMER: CNN and transformer under gated fusion for sand-dust image restoration," *IEEE Transactions on Image Processing*, vol. 30, no. 1, pp. 1221-1234, 2021.

[5] L. Wu, J. He, and Z. Tang, "Unsupervised image dedusting via a cycle-consistent generative adversarial network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 2, pp. 1458-1466, 2020.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Far, S. Ozai, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[7] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proceedings of the IEEE*

International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017, pp. 2223-2232.

[8] J. Zhao, F. Li, and M. Zhang, "AOSR-Net: All-in-One sandstorm removal network," *Journal of Visual Communication and Image Representation*, vol. 72, pp. 102859, 2020.

[9] M. Li, Y. Zhang, and H. Li, "USIR-Net: Sand-dust image restoration based on unsupervised learning," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5899-5909, 2019.

[10] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 22-29 October 2017, pp. 2223-2232.

[11] GeeksforGeeks, “GeeksforGeeks,” [Online]. Available: https://www.geeksforgeeks.org/. [Accessed: 14-May-2025].

[12] Papers with Code, “Image Restoration - Papers with Code,” [Online]. Available: https://paperswithcode.com/task/image-restoration. [Accessed: 14-May-2025].

[13] DataCamp, “DataCamp: Learn Data Science Online,” [Online]. Available: https://www.datacamp.com/. [Accessed: 14-May-2025].

[14] V7 Labs, “Vision Transformer (ViT): A Guide,” [Online]. Available: https://www.v7labs.com/blog/vision-transformer-guide. [Accessed: 14-May-2025].

Chapter 9

Appendices