Single Image Dehazing using Conditional GANs

Aryan Philip
HDSI
A69027943

Aseem Dandgaval
HDSI
A69027670

Abstract

Haze is a common atmospheric phenomenon that degrades the visibility of outdoor images. Removing haze from a single image is a challenging task due to the inherent ambiguity between the haze and the underlying scene. We propose using the Pix2Pix model, a conditional generative adversarial network, to learn a mapping from hazy images to their corresponding haze-free counterparts. The model will be trained and evaluated separately on the RESIDE and NH-HAZE datasets. We will assess the dehazing performance using the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) metrics. This report will cover the background and related work on image dehazing, details of the Pix2Pix model architecture and training procedure, descriptions of the RESIDE and NH-HAZE datasets, and the evaluation results and analysis. The goal is to provide insights into the effectiveness of Pix2Pix for single image dehazing and compare its performance across different datasets.

1 Introduction

Haze is a weather condition where particles like dust, smoke, and other dry particulates obscure the clarity of the sky, leading to diminished visibility and often causing the scenes captured in photographs to appear washed out or grey. This phenomenon not only affects everyday life but also compromises the quality of images used in various applications, ranging from simple photography to more critical uses such as surveillance and autonomous driving. Thus, the motivation for dehazing images is twofold: enhancing visual aesthetics for photography and improving the reliability of visual information for technical applications.

The challenge in removing haze from a single image stems from the difficulty in distinguishing between light that has been scattered by haze and light that reflects the true colors and shapes of the scene. Traditional methods often rely on physical models to estimate the amount of haze and subsequently reconstruct the haze-free image. However, these methods can be limited by their assumptions and the variability in real-world conditions.

To address this issue, we propose utilizing the Pix2Pix model, a well-known conditional generative adversarial network (GAN), designed to learn a direct mapping from hazy images to clear, haze-free images. The Pix2Pix model operates by training on pairs of hazy and corresponding haze-free images, thus learning to generate clean images by observing real examples. This approach not only simplifies the task by using end-to-end learning but is also flexible enough to adapt to various levels of haze intensity seen in different datasets.

We also explore other techniques like CycleGAN to avoid the need for paired data, which is highly difficult to gather in real-world scenarios.

For this project, the Pix2Pix model will be evaluated on two major datasets: RESIDE and NH-HAZE. RESIDE is a standard benchmarking dataset in image dehazing, providing a wide range of indoor and outdoor scenes, while NH-HAZE contains naturally hazy images, offering a challenging set of real-world conditions. The performance of our model will be quantitatively assessed using two main

metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). These metrics will help quantify the extent to which the dehazed images maintain fidelity to the ground truth, haze-free images.

We discovered the impact of different normalization and up-sampling techniques within the pix2pix architecture on the RESIDE and NH-Haze datasets, revealing that Layer Normalization (LayerNorm) outperformed Batch Normalization (BatchNorm) by providing more stable training and reducing artifacts. Additionally, replacing Transposed Convolutions with Nearest Neighbor Upsampling followed by a Convolutional layer in the generator network improved both PSNR and SSIM scores, indicating better image quality and accuracy in dehazed images. Despite initial promising results with CycleGAN, time constraints prevented full optimization, suggesting it as a valuable technique for future research in image dehazing.

2 Related Work

Single Image Dehazing has advanced very far with a lot of recent work focusing on diffusion models, Swin transformers and Complex U-Nets. Our work focuses on older GAN architectures and their related works.

Qu et al. [Qu+19] introduced an Enhanced pix2pix Dehazing Network, which builds upon the original pix2pix framework by incorporating specific adaptations for the task of image dehazing. This approach utilizes a enhanced Pix2PixHD to transform hazy images into clear ones, demonstrating significant advancements in handling complex dehazing scenarios with improved detail preservation and reduced artifact generation.

Cheng et al. [Ca16] presented a CNN-based dehazing method that infers color priors from semantic features extracted from single images. This approach was implemented on both synthetic and real-world hazy images and showed promising results in handling scenes with strong ambiguity. However, the model's training was limited to a specific set of image types, which may affect its generalizability to broader real-world scenarios.

Li et al. [La17] introduced a cascaded CNN framework that jointly estimates the transmission map and atmospheric light from a single image. This model demonstrated significant improvements over existing methods on synthetic and real-world datasets. Nonetheless, their study did not explore fully end-to-end networks, potentially missing out on further efficiencies in the dehazing process.

Rashid et al. [Ra18] developed a CNN with an encoder-decoder architecture tailored for single image dehazing. The model optimized the processing of high-intensity pixel values to overcome multiple dehazing challenges, showing enhanced efficiency. However, its application to images with scattered shades and diverse conditions remains to be fully addressed.

Ren et al. [Ren+16] employed a multi-scale deep neural network to estimate hazy images and medium transmission maps effectively. Applied to the NYU Depth dataset, their method outperformed state-of-the-art models in terms of both quality and processing speed.

Song et al. [Sa19] proposed a ranking-CNN that automatically learns haze-relevant features. This model obtained more effective results for both synthetic and real-world data compared to classical CNN approaches, although there is still a need to improve its computational efficiency.

Engin et al. [EGK18] utilized CycleGAN, termed Cycle-Dehaze, to perform single image dehazing without paired training data. Their method leverages the cycle-consistency loss to generate clearer images, presenting an innovative approach to unsupervised learning in dehazing tasks.

Finally, **Goncalves et al.** [Ga19] illustrated an end-to-end CNN model that introduced novel guided layers to adjust the network weights using a guided filter. This method effectively reduced structural information loss and demonstrated superior performance qualitatively and quantitatively.

While our work is inspired by all these existing works, our main focus was on Pix2Pix[Qu+19] and Cycle GAN[EGK18]. These works provided us with the knowledge to develop a framework to develop modifications to existing architectures as well as test different hyperparameters.

Each of these works contributes to the evolving landscape of image dehazing, providing a solid foundation for the development of new methods that address their respective limitations.

3 Method

3.1 Network Architecture

The proposed method consists of two main components: a generator network and a discriminator network. The generator network is tasked with generating clear images from hazy inputs, while the discriminator network aims to distinguish between the dehazed images produced by the generator and real, haze-free images.

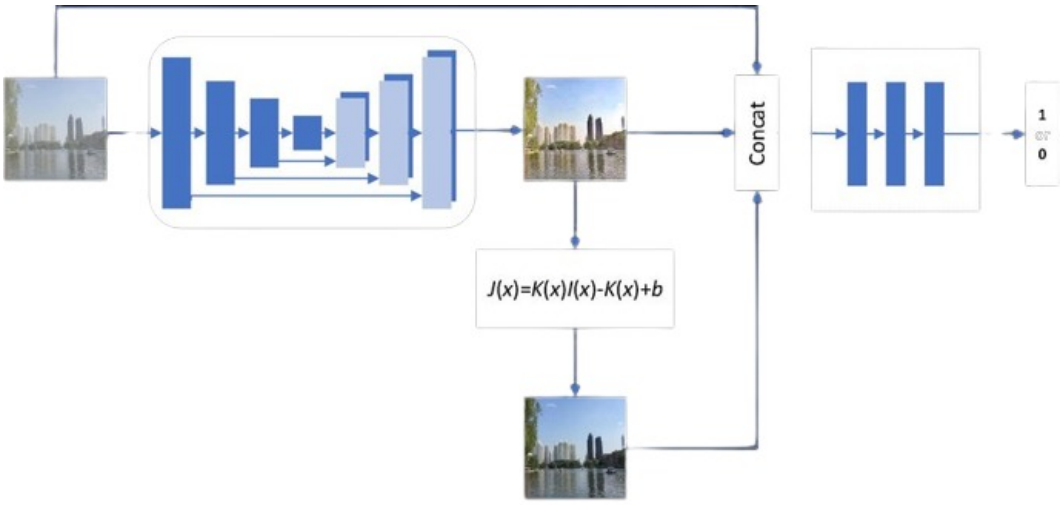

Figure 1: The detailed architecture of CGAN architecture where Generator and Discriminator Networks are shown.

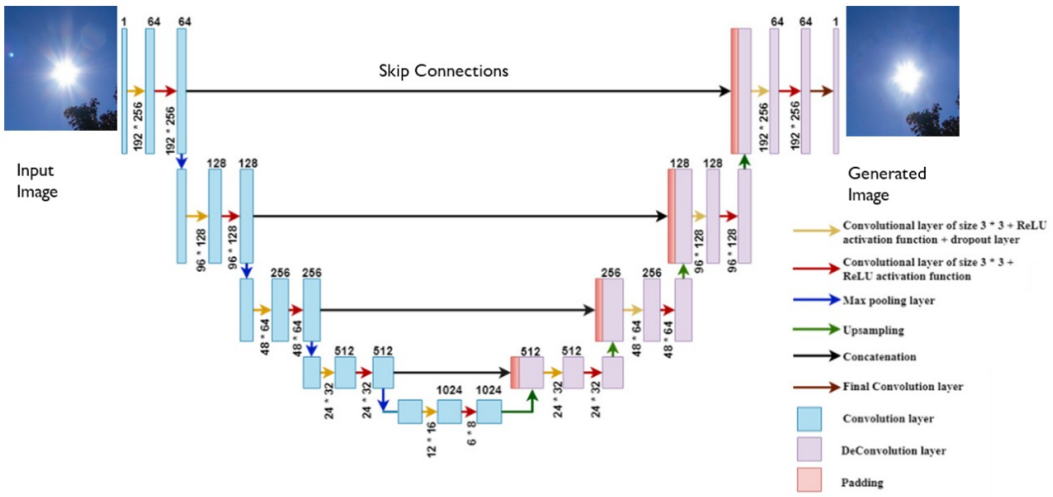

Figure 2: U-Net Architecture .

3.1.1 Generator

The generator adopts an encoder-decoder architecture. The encoder compresses the hazy input image into a latent space, capturing essential features required for haze removal. The decoder

then reconstructs a clear image from this compressed representation.The generator is a U-net, an encoder-decoder model that reduces the spatial dimensions of the source image to a smaller vector space containing particular features of the source image. These features are high-level features like texture, corners, etc. However, encoding results in losing a lot of low-level features in the earlier convolutional layers. To rectify this, the outputs of the encoder convolutional layers are used and fed into the corresponding decoder layers. The decoder works to upscale the encoded image into a new image having features similar to the target image. It uses transposed convolutions or deconvolutions to upscale the latent feature vector space. Formally, the generator function G can be expressed as follows:

$$G : x \mapsto y$$

where x is the input hazy image and y is the generated haze-free image.

3.1.2 Discriminator

The discriminator is a convolutional neural network that classifies images as real or fake. It evaluates the authenticity of the images produced by the generator against real haze-free images. A unique type of discriminator was implemented; rather than discriminating the whole image and applying binary classification, local image patches of smaller sizes are taken and classified each patch into 0 or 1. The probabilities are calculated and averaged to get more accurate results and losses. Since PatchGAN has fewer number parameters, the L2 loss helps us calculate the loss better and trains the model faster. The discriminator function D is defined as:

$$D : y \mapsto \{0, 1\}$$

where 1 denotes 'real' and 0 denotes 'fake'.

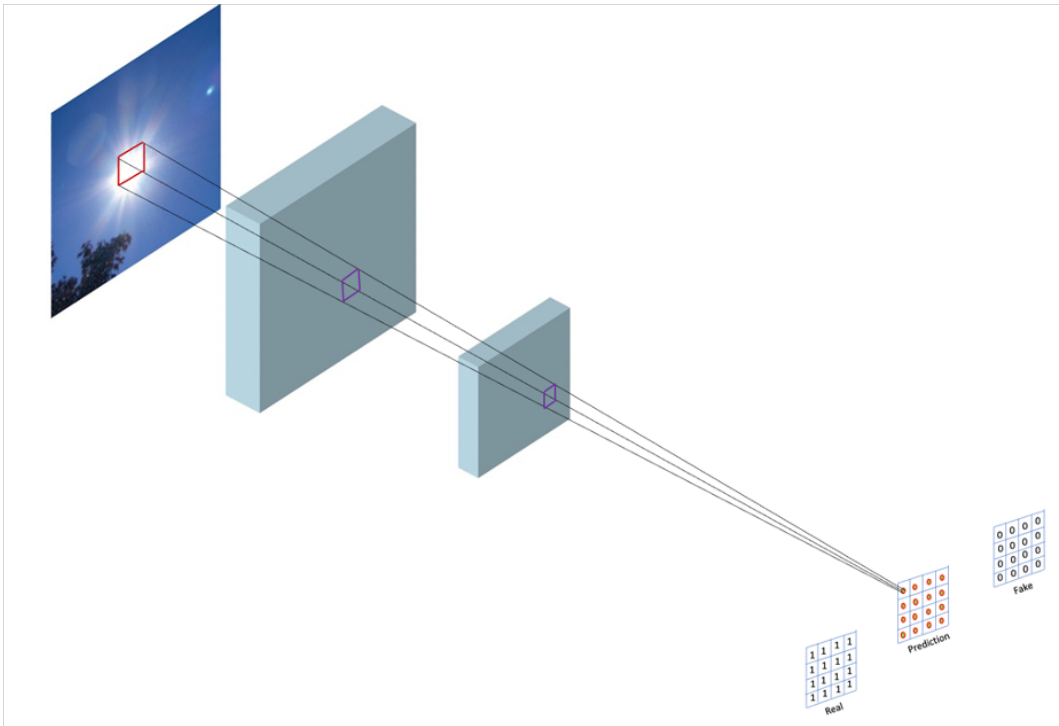

Figure 3: PatchGAN Discriminator.

3.2 Loss Functions

To train our network, we use a combination of adversarial loss and L_1 loss and for Cycle GAN we use an additional loss called cycle-consistency loss. The adversarial loss ensures that the generated images are indistinguishable from real images, while the cycle-consistency loss helps preserve the content between the input and the reconstructed images.

3.2.1 Adversarial Loss

The adversarial loss $\mathcal{L}_{adv}$ for the generator is defined as:

$$\mathcal{L}_{adv}(G, D) = \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D(G(x)))]$$

and for the discriminator, it is defined as:

$$\mathcal{L}_{adv}(D) = -\mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] - \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D(G(x)))]$$

3.2.2 Cycle-Consistency Loss

The cycle-consistency loss $\mathcal{L}_{cyc}$ ensures that an image x can be reconstructed from the dehazed image y as:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}\left[||F(G(x)) - x||_1\right] + \mathbb{E}_{y \sim p_{data}(y)}\left[||G(F(y)) - y||_1\right]$$

3.3 Training Algorithm for pix2pix CGAN

In the pix2pix framework, the CGAN is adapted to handle paired image-to-image translation tasks, such as converting hazy images to their clear counterparts where the ground truth (clear image) is available. The model consists of a generator G that translates hazy images to clear images and a discriminator D that aims to distinguish between real pairs (hazy image, clear image) and fake pairs (hazy image, generated clear image).

3.3.1 Generator Training

The generator in the pix2pix framework is typically a U-Net-based architecture, which is effective in capturing and utilizing local information from the input image, and supports direct feature sharing between corresponding layers in the encoder and decoder via skip connections. The generator's objective is not only to fool the discriminator but also to produce a visually similar image to the ground truth clear image. The combined loss function used for training the generator includes:

$$\mathcal{L}_G = \mathbb{E}_{x,y}[\log(1 - D(x, G(x)))] + \lambda \|y - G(x)\|_1 \tag{1}$$

Here, x represents the input hazy image, y is the ground truth clear image, and λ is a constant that balances the L1 loss (which encourages fidelity to the ground truth) with the adversarial loss.

3.3.2 Discriminator Training

The discriminator in pix2pix is often a PatchGAN, which classifies whether small patches of the image are real or fake. This design enables the discriminator to focus more on the texture and local structure of the image, which is crucial for generating high-quality results. The discriminator's loss function is given by:

$$\mathcal{L}_D = -\mathbb{E}_{x,y}[\log D(x, y)] - \mathbb{E}_x[\log(1 - D(x, G(x)))] \tag{2}$$

3.3.3 Training Procedure

The training alternates between the following two steps:

1. **Update the Discriminator:** Train D to maximize the probability of correctly classifying real and fake image pairs.
2. **Update the Generator:** Train G to minimize its combined loss with respect to its ability to fool D and to reproduce the clear image y.

Training is typically carried out until both the generator and discriminator losses stabilize, indicating that the generator is producing realistic clear images from hazy inputs and the discriminator is effectively distinguishing between real and generated images.

3.3.4 Testing Algorithm

For testing, the trained generator G is used to convert new hazy images to clear images. The process is straightforward:

$$\text{Clear Image} = G(\text{Hazy Image}) \tag{3}$$

This operation does not involve the discriminator, as the goal during testing is purely to generate dehazed images using the trained generator.

3.4 Metrics

To objectively evaluate the performance of our dehazing models, we employed two widely recognized metrics in image-to-image translation tasks: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM).

3.4.1 PSNR

The Peak Signal-to-Noise Ratio (PSNR) is a measure of the ratio between the maximum possible power of a signal (in this case, an image) and the power of corrupting noise that affects the fidelity of its representation. PSNR is typically expressed in logarithmic decibel scale. It is calculated by comparing the original image with a reconstruction (dehazed image), as shown in the equations below:

$$MSE = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} [I(i,j) - I'(i,j)]^2 \tag{4}$$

$$PSNR = 10 \times \log_{10} \left(\frac{255^2}{MSE} \right) \tag{5}$$

$$\begin{aligned} M &= \text{number of rows in the image} \\ N &= \text{number of columns in the image} \\ I(i,j) &= \text{pixel value from the original image} \\ I'(i,j) &= \text{pixel value from the dehazed image} \end{aligned}$$

3.4.2 SSIM

The Structural Similarity Index (SSIM) is a perceptual metric that quantifies image quality degradation caused by processing such as data compression or by losses in data transmission. Unlike PSNR, SSIM is designed to improve perceptual evaluations of image quality by considering image degradation as perceived change in structural information. SSIM is calculated using the following formula:

$$SSIM(I, I') = l(I, I') \cdot c(I, I') \cdot s(I, I') \tag{6}$$

$$l(I, I') = \frac{2\mu_I \mu'_I + C_1}{\mu_I^2 + \mu_I'^2 + C_1}, \tag{7}$$

$$c(I, I') = \frac{2\sigma_I \sigma'_I + C_2}{\sigma_I^2 + \sigma_I'^2 + C_2}, \tag{8}$$

$$s(I, I') = \frac{\sigma_{II'} + C_3}{\sigma_I \sigma'_I + C_3} \tag{9}$$

$$
\begin{aligned}
l(I, I') &= \text{luminance comparison function} \\
c(I, I') &= \text{contrast comparison function} \\
s(I, I') &= \text{structure comparison function}
\end{aligned}
$$

These metrics, which utilize a reference or ground truth image, are crucial in contexts where such references are available. However, in many real-world applications, an exact ground truth may not be present, which challenges the applicability of these metrics.

4 Experiments

4.1 Dataset

We used two datasets to benchmark the Image Dehazing task:

RESIDE:

Figure 4: The RESIDE dataset, Top: Ground Truth clear images, Bottom: Input hazy images.

It is a large-scale benchmark consisting of both synthetic and real-world hazy images, RESIDE highlights diverse data sources and image contents, and is divided into five subsets, each serving different training or evaluation purposes. [Li+18] There are total 20,000 images across all the datasets each with resolution ranging from 512 x 512 to 3264 x 2448. Each clear ground-truth image has 5 corresponding hazy counterparts and 5 corresponding depth maps as well.[EGK18]

We used a subset of the dataset using only the indoor images. We used a total of 1399 clear images each having 10 hazy counterparts to them each of resolution 640x420 pixels. These were resized to 256x256. We did not use the corresponding depth maps. The data format is .png files.

NH-HAZE:

Figure 5: The NH-HAZE dataset, Top: Ground Truth clear images, Bottom: Input hazy images.

The NH-HAZE dataset is a benchmark dataset specifically designed for evaluating image dehazing algorithms on non-homogeneous (spatially varying) hazy scenes. It contains a total of 55 pairs of real-world outdoor hazy and corresponding haze-free (ground truth) images. The hazy and corresponding haze-free image pairs have the same resolution, which varies across the dataset from 600x400 to 3264x2448 pixels. [Wu+21]

We used the entire dataset of all 55 image pairs. These were resized to 512x512 pixels. The data format is .png files.

4.2 Other Experimental Details

Parameters	RESIDE values	NH-HAZE values
Learning rate	2e-4	1e-4
Batch Size	32	16
Image Size	256	512
Image Channels	3	3
L1 Lambda	10e-5	100
Lambda GP	10	10
Epochs	100	100

Table 1: Final hyper-parameter values used.

We experimented with various hyperparameter values before settling on the final configurations for the RESIDE and NH-HAZE datasets, as shown in Table 1. For the RESIDE dataset, we found that a learning rate of 2×10^{-4}, a batch size of 32, and an image size of 256 pixels yielded the best looking generated images in our testing. The model was trained with 3 image channels and employed an $L1_\lambda$ of 10^{-5} and a gradient penalty (Lambda GP) of 10, across 100 epochs.

The NH-HAZE dataset required a slightly different approach. A lower learning rate of 1×10^{-4} and a smaller batch size of 16 were more effective, with the image size increased to 512 pixels to capture more detail as resizing them to 255 pixels gave poor results. The same number of image channels (3), $L1_\lambda$ (100), and Lambda GP (10) were used, also over 100 epochs. These parameter values were chosen after a fair amount of experimentation and tuning, and also referred to other related work to find values that would best work for us. We tried multiple combinations to increase dehazing performance and ensure that the model could generalize well across different types of hazy images.

We also utilized various data augmentation techniques to enhance the diversity and robustness of our training dataset. After trying augmentations like rotation, color jitter and more, we ended up only using the horizontal flip augmentation with a probability of 0.5 for both input and label images. Finally both sets of images were normalized. These augmentations created a more diverse and robust dataset, and improved the generalization and performance of our image dehazing model.

4.3 Ablative Study and Results

Figure 6 and 7 have the results for the pix2pix architecture for the RESIDE and NH-Haze datasets respectively. We initially implemented Batch Normalization (BatchNorm) in our generator network, but after some experimentation, we discovered that Layer Normalization (LayerNorm) produced better results. LayerNorm provided more stable training and significantly reduced artifacts and checkerboarding in the generated images.

Additionally, in the generator network, we focused on optimizing the decoder. We began with Transposed Convolutions, which initially gave good results. However, we found that using Nearest Neighbor Upsampling, followed by a Convolutional layer, yielded even better performance. This approach helped to reduce artifacts and produce clearer and more accurate dehazed images, enhancing the overall effectiveness of our model.

We evaluated the performance of our models using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) scores, which are standard metrics for assessing the quality of reconstructed images. PSNR measures the ratio between the maximum possible power of a signal and the power of corrupting noise, with higher values indicating better quality. SSIM assesses the

Figure 6: Pix2pix results for RESIDE, The images on the left side are generated using Transpose Convolutions and the ones on the right are generated using Upsampling. Top: Hazy images, Middle: Generated images, Bottom: Ground Truth images.

Figure 7: Pix2pix results for NH-Haze, The images on the left side are generated using Transpose Convolutions and the ones on the right are generated using Upsampling. Top: Hazy images, Middle: Generated images, Bottom: Ground Truth images.

similarity between two images, considering luminance, contrast, and structure, with higher scores reflecting greater similarity to the ground truth.

After implementing Nearest Neighbor Upsampling in our generator network, we observed a slight improvement in both PSNR and SSIM scores compared to the models using Transposed Convolutions. This enhancement demonstrated that the upsampled models produced more accurately and were

Models	PSNR	SSIM
RESIDE ConvTranspose	13.42	0.715
RESIDE Upsample	13.75	0.812
NH-HAZE ConvTranspose	13.00	0.754
NH-HAZE Upsample	14.02	0.761

Table 2: PSNR and SSIM scores

consistent with our visual analysis of the generated images, confirming the effectiveness of our chosen upsampling technique.

4.4 Future Work

We also conducted some experiments using CycleGAN, an approach that leverages generative adversarial networks for image-to-image translation without requiring paired training data. Although CycleGAN showed promising potential, we faced significant time constraints that hindered our ability to train the models efficiently, as CycleGAN typically requires extensive training time to achieve optimal performance.

Consequently, we were unable to fully explore and fine-tune this method within the project's time-frame. However, the initial results suggest that CycleGAN could be a valuable technique for future work. Further experiments and optimizations with CycleGAN could potentially lead to substantial improvements in image dehazing, making it a worthwhile avenue for continued research.

5 Implementation

We leveraged existing code from various sources to accelerate development. We borrowed code for the generator, discriminator, the training loop, and some utility functions. To tailor the project to our specific needs, we made significant modifications and tuning to the generator. We also enhanced the training loop by writing additional code to log PSNR and SSIM scores, providing critical metrics for evaluating model performance. Furthermore, we developed new utility functions to support our workflow and wrote custom code for data loaders from scratch, ensuring efficient and flexible data handling throughout the training process.

References

[Ca16] Li Cheng and et al. "CNN-Based Dehazing with Semantic Color Priors". In: *Journal of Image Processing* 29.4 (2016), pp. 1254–1266.

[Ren+16] W. Ren et al. "Single image dehazing via multi-scale convolutional neural networks". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*. Vol. 14. Springer International Publishing, 2016, pp. 154–169.

[La17] Wen Li and et al. "A Flexible Cascaded CNN for Joint Estimation of Transmission Map and Atmospheric Light". In: *Proceedings of the IEEE Conference on Computer Vision*. 2017, pp. 198–206.

[EGK18] D. Engin, A. Genç, and H. Kemal Ekenel. "Cycle-dehaze: Enhanced cyclegan for single image dehazing". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 825–833.

[Li+18] B. Li et al. "Benchmarking single-image dehazing and beyond". In: *IEEE Transactions on Image Processing* 28.1 (2018), pp. 492–505.

[Ra18] Umar Rashid and et al. "CNN Based Encoder-Decoder Framework for Single Image Dehazing". In: *Journal of Advanced Computer Vision* 30.3 (2018), pp. 960–971.

[Ga19] Hugo Goncalves and et al. "End-to-End CNN for Image Dehazing with Guided Layers". In: *Computer Vision and Image Understanding* 184 (2019), pp. 32–43.

[Qu+19] Yanyun Qu et al. "Enhanced pix2pix Dehazing Network". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[Sa19] Peng Song and et al. “Learning Haze-Relevant Features with a Ranking-CNN for Image Dehazing”. In: *IEEE Transactions on Image Processing* 28.5 (2019), pp. 2160–2175.

[Wu+21] H. Wu et al. “Contrastive learning for compact single image dehazing”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 10551–10560.