A WGAN-based Framework for Aerial Image Denoising under Mixed Noise Conditions

Vedansh Kumar
Computer Science and Engineering
Vellore Institute of Technology, Chennai
Chennai, India, 600127
vedansh.kumar2022@vitstudent.ac.in

Hrut Rajesh Shah
Computer Science and Engineering
Vellore Institute of Technology, Chennai
Chennai, India, 600127
hrutrajesh.shah2022@vitstudent.ac.in

Sarthak Chaudhary
Computer Science and Engineering
Vellore Institute of Technology, Chennai
Chennai, India, 600127
sarthak.chaudhary2022@vitstudent.ac.in

Dr. Geetha S
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai
Chennai, India, 600127
geetha.s@vit.ac.in

***Abstract*—Aerial imagery plays a crucial role in various applications around the world, yet its quality is frequently compromised by complex and mixed noise patterns, originating from a multitude of factors such as sensor limitations, transmission artifacts, and also environmental conditions. To address this issue, we developed a robust de-noising framework based on Wasserstein Generative Adversarial Networks (WGANs) that effectively removes random, multiplicative, and Gaussian noises from aerial images. The framework uses a generator-discriminator architecture, where the generator is optimized using an adversarial loss function, and a hybrid loss function comprising of pixel-wise L1 loss, perceptual loss and total variance loss. Thorough experimentation on benchmark datasets demonstrated that this framework can consistently outperform conventional and deep learning based denoising techniques in terms of SSIM, PSNR, and visual fidelity.**

***Index Terms*—Aerial Image Denoising, WGAN, Mixed Noise Patterns, Image Restoration, Hybrid loss function**

I. Introduction

High resolution aerial images have recently found their way into multiple applications, ranging from environmental monitoring, urban development, disaster response, to even defence surveillance. The quality of these images however, is often degraded due to the presence of various types of noise, thus reducing the utility of such images. These noise patterns can be introduced during image acquisition, transmission, or even storage. In particular, mixed noise—a non-trivial combination of Gaussian, impulse, and sensor-specific artefacts—poses significant challenges to denoising algorithms because of its unpredictable structure and spatial heterogeneity.

Classical denoising techniques such as total variation minimisation [1] and non-local means filtering [2] offer initial solutions by preserving edges and exploiting self-similarity. These methods however, often oversmooth the textures, or become ineffective when noise distorts critical structural information. Methods such as deep learning have also substantially advanced image restoration, with models like DnCNN [3] and FFDNet [4] leveraging convolutional networks to learn powerful feature mappings for noise removal, but despite their effectiveness under controlled synthetic noise, these models tend to underperform in real-world aerial settings, where noise levels vary spatially and temporally.

The advent of Generative Adversarial Networks (GANs) by Goodfellow [5] later on, further revolutionised image synthesis and restoration tasks. His model worked by taking advantage of adversarial learning, where both a generator and a discriminator constantly learned from each other. Later on, conditional GANs such as Pix2Pix [6] and unpaired models like CycleGAN [7] extended this potential. This architecture however, relies heavily on stable training and specially curated datasets, limiting its use in aerial imagery applications as the images are often unlabelled and exhibit high inter image variance.

In order to overcome certain limitations and suit domain-specific applications, researchers created GANs with slight modifications, such as the ADNet introduced by Yang et al. [8]. This model used visual cues and an attention mechanism to guide the image reconstruction, thus recovering fine details that may have been lost to salt and pepper noise and Gaussian noise. Another such instance is the Wasserstein GAN [9]. WGANs worked by redefining the loss calculation using the Earth Mover distance, increasing the model's reliability and capability in terms of the pictures that it could generate.

Furthermore, incorporating attention mechanisms [10] and feature level guidance [11] has been known to enhance the denoising precision as it focuses on noise sensitive regions, something that is incredibly helpful in high frequency

aerial images. Despite all these advances, most models are confined to single noise-based scenarios, or require extensive paired datasets. These factors reduce their practicality in unpredictable and uncontrollable environments.

To improve on the shortcomings of the existing models, our team worked on a WGAN based denoising framework that uses a hybrid loss function, created by fusing the perceptual, pixel-wise, and total variation losses. The perceptual loss component in our framework is extracted using VGG-19 feature maps, ensuring that the output preserves all meaningful content of the image in the presence of severe noise. The main objective of our framework was to tackle the issue of having to remove multiple noise types from a single image successfully, thus generalising the outputs and making it more robust and suitable for use in cases of high unpredictability, such as the real world. Extensive tests on benchmark datasets showed that our proposed framework performs significantly better than many traditional or existing methods in quantitative metrics, such as Peak Signal-to-Noise Ratio (PSNR), in Structural Similarity Index Measure (SSIM), and qualitative visual fidelity.

II. Related Works

The paper [12] used a convolutional network, CBDNet, to denoise images with noise types most similar to real-world photographs. The model consisted of a noise estimator subnetwork in conjunction with a non-blind denoising subnetwork. The model was trained on both real-world images in clean-noisy pairs and synthetically noise-generated images made using realistic camera noise models from datasets such as RENOIR. The proposed model achieved PSNR and SSIM scores of 30.78 dB and 0.801 dB, respectively, on the SIDD benchmark. The key limitations were performance degradation when applied to images with different noise characteristics and heavy reliance on the accuracy of the noise detection layer.

On the other hand, the paper [11] aimed to tackle the problem of denoising real-world photographs with complex noise patterns using a single-stage blind denoising network known as RIDNet. The model employed a "residual on the residual" design which allows for flow of low frequency helping in the network's ability to reconstruct the clean image along side a feature attention mechanism allowing the network to focus on the most informative features among different channels. RIDNet achieved PSNR improvements of 9.5 dB and 7.93 dB over FFDNet and CBDNet respectively. The effectiveness of the model however can diminish when applied to images with different noise characteristics as compared to the training data.

Another CNN based model [3] used a feed-forward deep architecture that integrated techniques like residual learning with batch normalization, accelerating the training convergence and improving the denoising quality. The model created by the authors (DnCNN) was designed for blind denoising, thus handling variable noise levels without prior specification. DnCNN implicitly removes the latent clean image in the hidden layers, allowing the authors to train a single DnCNN model to tackle with several general image denoising tasks, such as Gaussian denoising, single image super-resolution, and also JPEG image de blocking.

Approaches different than GAN such as the NLM algorithm [2] aimed to denoise the images by the use of redundancy found in natural images beyond their own local neighborhoods. It used the Non-Local Means (NLM) algorithm which predicts the denoised pixel values based on the weighted average of the pixel in the image based on patch similarity in contrast to spatial proximity. Textures and fine details were preserved with the non-local strategy more efficiently in comparison to traditional filters. The model received a PSNR score of 31.7 dB on barbara image with a standard deviation of 25 and hence outperforming bilateral filtering. The main drawbacks of the model were its high computation cost and the chance of over smoothing in uniform regions.

The paper [13] sheds light on how traditional CNNs are limited by the requirements of the training sample size. The authors here propose an aerial image denoising model with a multi scale residual learning approach. Rather than reconstructing the clean image directly, the model first learns the noise distribution itself, and then subtracts it from the noisy input to obtain the final restored output. This allows the model to be beneficial for small training datasets. Similarly, another CNN based model FFDNet [4], is a custom CNN tailored for practical image denoising tasks. FFDNet achieved this practicality as unlike other models that required separate training for every noise intensity, FFDNet introduces a noise level map as part of the input, enabling a single model to adapt to a continuous range of noise levels. This method enhances generalization for real-world scenarios, where noise levels vary spatially across an image. Hence the proposed model from the paper proved to be better than other conventional CNN based models.

Lastly, another [10] GAN based approach of MA-GAN [10], improves the spatial resolution of remote sensing images. The generator here was made up of two key modules: the Pyramidal Convolution in Residual-Dense Block (PCRDB) that combined a multi-scale convolution and channel attention to adaptively learn of the residuals and Attention-Based Upsample (AUP) block that does flexible upsampling by utilizing pixel attention. The MA-GAN, for a scale factor of 2, secured PSNR and SSIM scores of 31.98 dB and 0.9102 respectively. The model performed equally good on scale factors of 4 and 8. The models drawbacks included and complexity and the large computational demands needed for deployment in low resource environments.

III. Model Architecture

The model that we used is a WGAN with a U-Net generator and a 2D Patch GAN disciminator. This architecture was designed to generate clean images from the noisy image given as input.

Generator

The generator G architecture was inspired from the U-Net model. It comprised of symmetric Downsampling and Upsampling blocks along with skip connections. Skip connections play an important role in this model because as we perform downsampling and upsampling the low level spatial information of the image gets lost so these skip connections prevent this loss of information by providing low level spatial information while upsampling. The generator took the input noisy image of size 224x224 px.

The generator consist of the following parts:

- **Downsampling layers:** It consists of a convolution block which consist of two 3x3 convolution layer with normalization and ReLU activation. And after every convolution block maxpooling of 2x2 was performed. This whole procedure was done 4 times during the downsampling.
- **Bridge:** It consist of a simple convolution block with two 3x3 convolution layers and after every convolution normalization was done and ReLU activation was used.
- **Upsampling layers:** Each block of Upsampling layers consisted of concatenation of low level spatial information given by skip connection, convolution block which had 2 convolution layer with 3x3 kernel and after every convolution the normalization was performed and ReLU activation was used and finally upsampling was done with the help of transpose convolution with bilinear upsampling.
- **Final output layer:** This is used to give the clean image with RGB channels as output. It takes in 64 channel and gives out 3 channels.

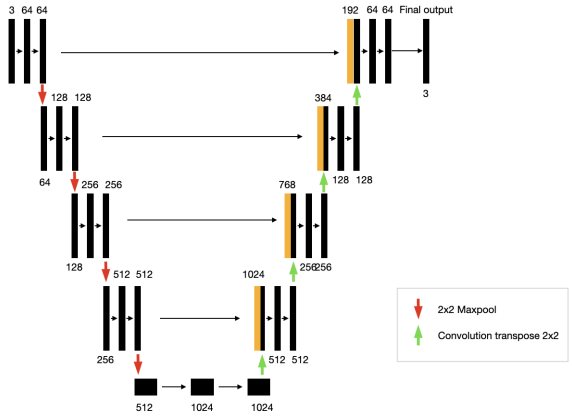

Fig. 1. Generator Architecture

Discriminator

The discriminator D is inspired by the PatchGAN architecture. The discriminator took both clean and denoised image as input. It is used to evaluate how similar is the denoised image to the clean image. To perform this operation, the model uses convolution block consisting of convolution layer with a 4x4 kernel, normalization and ReLU activation. Using these convolution blocks the discriminator compares the denoised image with the clean image.

IV. Loss function

During this research, our objective was to denoise the image, and in order to achieve that output, we used a complex loss function which consisted of adversarial loss, pixel-wise reconstruction loss and perceptual loss.

1) Adversarial Loss (with Wasserstain Gradient Penalty): We used this function to stably train the GAN.

$$\mathcal{L}_{adv} = \mathbb{E}_{x\sim P_r}[D(x)] - \mathbb{E}_{\hat{x}\sim P_g}[D(\hat{x})] \tag{1}$$

Lipschitz constraint which is a gradient penalty was also introduced.

$$\mathcal{L}_{gp} = 10 \cdot \mathbb{E}_{\tilde{x}\sim P_{\tilde{x}}}\left[(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2\right] \tag{2}$$

where $\tilde{x} = \epsilon x + (1-\epsilon)\hat{x}$ and $\epsilon \sim \mathcal{U}(0,1)$. The total discriminator loss becomes:

$$\mathcal{L}_D = \mathcal{L}_{adv} + \mathcal{L}_{gp} \tag{3}$$

A. Reconstruction Loss (L1 Loss)

To keep the pixel wise similarity that is to maintain the desired color for the output image we used L1 loss.

$$\mathcal{L}_{L1} = \mathbb{E}_{(x,y)}\left[\|G(x) - y\|_1\right] \tag{4}$$

B. Perceptual Loss using VGG19:

In order to get better and realistic output we used perceptual loss.

$$\mathcal{L}_{perc} = \sum_i \|\phi_i(G(x)) - \phi_i(y)\|_2^2 \tag{5}$$

where $\phi_i(\cdot)$ denotes that the feature at layer i of the pretrained VGG19 Network.

C. Final Generator Loss:

The final loss for generator is combination of equation (4), (5) and adversarial loss.

$$\mathcal{L}_G = 0.1 \cdot \mathcal{L}_{L1} + (-\mathbb{E}[D(G(x))]) + 100 \cdot \mathcal{L}_{perc} \tag{6}$$

V. Dataset

The dataset we used for training was built using 4 different datasets:

- UC-Merced Dataset [14]
- WHU-RS19 Dataset [15]
- RSSCN7 Dataset [16]
- AID Dataset [17]

Once the images of these datasets were combined we started introducing noise into the combined dataset. We introduced 10% to 15% noise in each image and there was only single type of noise in each image.

The noise models that we used were Gaussian noise, Multiplicative noise and Random valued noise. Upon performing these operations we got a dataset with a total of 13,700 images. The following is the **TABLE I** that describes our dataset.

TABLE I
DATASET DESCRIPTION

Noise Type	Number of images
Gaussian	4584
Multiplicative	4583
Random valued	4583

VI. Results

A. SSIM & PSNR comparision

Once the model was trained, it was tested on a test dataset the SSIM and PSNR value were noted down and compared with Convolutional Blind Denoising Network(CBDNet) [12], Proximal Neural Network (PNN) [18], Super-Resolution Using Deep Convolutional Network (SRCNN) [19] and methods like Weighted Nuclear Norm Minimization (WNNM) [20].

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	PSNR (dB)	SSIM
CBDNet	30.78	0.801
PNN	37.66	0.8796
SRCNN	**38.003**	0.9008
Proposed Model	33.04	**0.9172**
WNNM	37.661	0.9001

The values in the **TABLE II** are referenced from [13]. The proposed model has a PSNR value of 33.04dB and an SSIM of 0.9172.

PSNR is calculated with the help of equation (7).

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{\text{MSE}} \right) \tag{7}$$

PSNR is used to measure the ratio between the maximum value of the pixel and the mean squared error (MSE) between the denoised image and clean image. So our model performs better than CBDNet and a bit lower than the other model and methods like PNN, SRCNN, CBDNet, WNNM in terms of PSNR value.

However, if consider SSIM value our model perfoms slightly better than the other models. The SSIM score is calculated as shown in equation (8).

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{8}$$

where μ_x, μ_y are the means of the images, σ_x, σ_y are the variances, and σ_{xy} is the covariance between the two images.

Even though our model suggested does not have the highest PSNR value but our higher SSIM value indicates that the image reconstruction of our model is better compared to the other models.

B. Visual Analysis

As we can see **fig. 2** is devided into 3 sections clean image, noisy image and denoised image or image generated by generator.

Fig. 2. Qualitative comparison: (Left) Clean Ground Truth, (Middle) Noisy Input, (Right) Output generated by the proposed model.

The left most is the clean arial image which was used as ground truth during training and evaluation of our model. The image exibits clean, sharp details and visually we can differentiate the buildings, roads etc.

In the center we have the noisy image which is a degraded version of the clean image. This noise in the image can be introduced due to faulty sensor, noise during transmission of image or faulty transmitter etc. Due to this noise it becomes hard to get the meaning from the image, it becomes difficult to differentiate the buildings, roads etc. This takes away the clarity from the image and makes difficult to differentiate the features and textures in image. So to make the image more clear we try to denoise it using our model.

Upon examining the output image from our proposed model we found out that the image becomes more clearer and there is a strong preservation of structural elements. In this denoised image we can differentiate the features and textures compared to the noisy image.

The improved quality of image validates that our model is effectively denoising the ariel image and at the same time maintaining the spatial coherence of the image which makes it very effective for the real world application like ariel image denoising, satellite image denoising etc.

VII. Conclusion

This paper introduced a novel denoising framework built upon Wasserstein Generative Adversarial Networks (WGANs) to deal with the issue of aerial images being affected by various types of noise, i.e., random, multiplicative, and Gaussian distributions. The model took advantage of the WGAN training

procedure in conjunction with VGG16-based perceptual loss to significantly suppress noise while preserving structural details. Our dataset consisted of four publicly available aerial imagery datasets, namely UC Merced, WHU-RS19, RSSCN7, and AID. Each of the images was then augmented to contain 10-15% generated noise to produce degradations similar to those seen under real-world conditions with a total of 13,700 noise samples. The given approach outperformed various existing models with impressive PSNR and SSIM scores of 33.04 and 0.9172 respectively. This portrays the ability of the model to be used for practical remote sensing applications, and and work may involve extending the modelmultispectralral or hyperspectral data or optimizing the model for real-time processing on edge devices.

REFERENCES

[1] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1–2, pp. 89–97, 2004.

[2] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 60–65 vol. 2, 2005.

[3] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[4] K. Zhang, W. Zuo, and L. Zhang, "Ffdnet: Toward a fast and flexible solution for cnn-based image denoising," *IEEE Transactions on Image Processing*, vol. 27, p. 4608–4622, Sept. 2018.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2018.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2020.

[8] C. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, and H. Liu, "Attention-guided cnn for image denoising," *Neural Networks*, vol. 124, pp. 117–129, 2020.

[9] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017.

[10] M. Xu, Z. Wang, J. Zhu, X. Jia, and S. Jia, "Multi-attention generative adversarial network for remote sensing image super-resolution," 2021.

[11] S. Anwar and N. Barnes, "Real image denoising with feature attention," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3155–3164, 2019.

[12] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," 2019.

[13] C. Chen and Z. Xu, "Aerial-image denoising based on convolutional neural network with multi-scale residual learning approach," *Information*, vol. 9, no. 7, 2018.

[14] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279, ACM, 2010.

[15] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *ISPRS Symposium: 100 Years ISPRS - Advancing Remote Sensing Science*, (Vienna, Austria), 2010. Dataset available at: https://www.kaggle.com/datasets/sunray2333/whurs191.

[16] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2321–2325, 2015.

[17] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, p. 3965–3981, July 2017.

[18] H. T. V. Le, A. Repetti, and N. Pustelnik, "Unfolded proximal neural networks for robust image gaussian denoising," 2024.

[19] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," 2015.

[20] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2862–2869, 2014.