A Comprehensive Study On Satellite Image Super-Resolution Using Diffusion and GAN Based Model

Abhishek Pandey
HERE Technologies
Navi Mumbai, MH, India
Abhishek.pandey@here.com

Anto Felix Immanuel
HERE Technologies
Navi Mumbai, MH, India
antofelix.immanuel@here.com

Deekshant Saxena
HERE Technologies
Navi Mumbai, MH, India
Deekshant.saxena@here.com

Karan Sahu
HERE Technologies
Navi Mumbai, MH, India
Karan.sahu@here.com

Nitin Mukesh
HERE Technologies
Navi Mumbai, MH, India
Nitin.mukesh@here.com

***Abstract*— Object detection and feature extraction from satellite images is a crucial step while using satellite images for purposes like navigation, urban planning, weather monitoring, etc. While deep learning approaches are too common for object detection task, but when the satellite images are of low quality, the small objects are missed by detection model due to their size and visibility issue. In this paper we propose a study of two broad areas of Generative AI models namely GANs and Diffusion model and their ability to handle the low-resolution images to improve overall detection problem. We train SRGAN and Diffusion based super-resolution model on custom real-time datasets and present a comprehensive performance evaluation and analysis. We found that Diffusion model increased the object detection rate by almost 130% when compared with Raw image object detection.**

Index Terms*—Diffusion model, GAN, Satellite images, object detection, YOLO, GeoAI

I. INTRODUCTION

Satellite images are used these days for various purposes from creating detailed maps of the Earth's surface to other areas such as navigation, urban planning, and environmental management.

For all these purposes, certain features/objects are required to be detected from satellite images. Deep learning-based object detections models are very useful these days for detecting certain features from images. Since satellite images are captured from a distance from earth surface, objects are usually very small, and sometimes certain objects are not visible due to low resolution of the image. The deep learning-based models are not able to detect features from satellite images in such cases.

In recent years, researchers have proposed different Convolutional Neural Networks (CNN)-based architectures such as Faster Region CNN (R-CNN), You Only Look Once (YOLO), and Retina Net for object detection in satellite images [1][2]. If the satellite images are of low resolution, the model is no good in detecting objects.

Recent emergence of generative AI models has introduced the concept of synthetically enhanced datasets to increase model performance while decreasing training data requirement [3]. Generative AI models suffer from the problem of hallucination, a phenomenon in which the model outputs are so different from reality that the results are nonsensical.

While low-resolution satellite images do have limitations in object detection, one approach can be to generate the high-resolution image from low-resolution using Generative AI models.

In this paper, to address these challenges, we propose a study of two class of model GANs and Diffusion Models in addressing low-resolution images and their impact on overall object - detection task.

A. What is GAN?

GAN stands for Generative Adversarial Network and is a machine learning model that can create new data instances that resemble training data. GANs are made up of two neural networks, the generator and the discriminator, that compete with each other to improve predictions.

The generator takes a random number and generates an image, which is then inputted to the discriminator. The discriminator contains both real and fake images and tries to predict the labels with the identification of real and fake images. As an output, it returns probabilities of a number between 0 and 1, where 0 represents a prediction of fake and 1 represents authenticity.

Generative adversarial networks (GANs) have been extensively studied in the past few years. Arguably their most significant impact has been in the area of computer vision where great advances have been made in many problems such as image generation, image-to-image translation, facial attribute manipulation and similar domains [4].

Super-Resolution Generative Adversarial Network (SRGAN) leverages a deep residual network (ResNet) with skip-connections, departing from the sole use of Mean Squared Error (MSE) as the optimization target [5]. Unlike prior works, the authors define a novel perceptual loss function. This loss function incorporates high-level feature maps extracted from a pre-trained VGG network [6, 7, 8]. Additionally, a discriminator network is employed to steecr the solution towards images perceptually indistinguishable from high-resolution (HR) reference images.

B. What are Diffusion Models?

Diffusion Models (DMs) are based on the diffusion concept in Physics wherein an image has Gaussian noise added to it

iteratively and a neural network is trained to learn the reverse process.

In the forward process, Gaussian noise is added to the image in *T* time steps until the initial distribution ends up resembling random noise. The trained neural network learns to predict the reverse process in each time step, effectively taking random Gaussian noise and transforming it across *T* time steps back into the initial data distribution [9i].

Since predicting the initial data distribution from Gaussian noise is a long iterative process which requires high amount of computational power, latent diffusion models which operate on the compressed *'latent space'* representation of the data were introduced [10i].

Diffusion models can also be conditioned on image/text embeddings to steer the generation process to a desirable output [11i].

Conditional latent diffusion models have been used to successfully generate photo-realistic images based on text prompts in such tools as DALL-E, Midjourney and Stable Diffusion. The success of image generation using conditional diffusion led to the use of these models for image restoration tasks such as super resolution, in-paining, etc.

C. *Diffusion model for Super resolution*

Single image super-resolution (SISR) aims to reconstruct high-resolution (HR) images from given low-resolution (LR) images. It is an ill-posed problem because one LR image corresponds to multiple HR images [12].

Recently, learning based SISR methods have greatly outperformed traditional methods. However, PSNR-oriented, GAN-driven and flow-based methods suffer from over-smoothing, mode collapse and large model footprint issues, respectively.

Many works have already been done for SISR using diffusion models and its variation like [12]. Some models have utilized the remote sensing images for SISR purpose [12].

In this paper, we will be presenting the comparative study between a Diffusion Model namely, DiffIR, Efficient diffusion model for image restoration and SRGAN for SISR on satellite-based images. We will also show how the super resolution results by these two models affects the objects visible in the image by inferencing on these images using a custom trained object detection model YoloV8.

Additionally, we also present some other metrics to quantify the results of super resolution using diffusion model and SRGAN and present the conclusion on which is better for object detection and feature extraction task.

We use the same model same model to make inference on SR image of both the model and present the exact count and improvements in respect to each other.

II. Literature Review

Diffusion-based models are good in presenting better results, but often the tradeoff comes from optimizing computational complexity in which SRGAN and better and light models.

Here we discuss in detail about these two models and Yolov8 which has been used for object detection task.

A. *DiffIR*

Traditional diffusion models (DMs) require many iterations, computational resources, and model parameters to generate accurate and realistic images or latent feature map.

Although DMs achieve impressive performance in generating images from scratch (image synthesis), it is a waste of computational resources to directly apply the DM paradigm of image synthesis to Image Restoration task [13].

Since most pixels and information in IR are given, performing diffusion model on whole images or feature maps not only spends a lot of iterations and computation but also is easy to generate more artifacts. Overall, DMs have strong data estimation ability, but applying the existing DM paradigm in image synthesis to IR is inefficient. To address the issue, the DiffIR which adopts DM to estimate a compact IPR to guide the network to restore images is used. Since the IR prior representation (IPR) is quite light, the model size and iteration of DiffIR can be largely reduced to generate more accurate estimations compared with traditional DM [11].

B. *SRGAN*

SRGAN is a GAN-based network designed to generate super-resolution images from low-resolution. SRGAN architecture consists of convolution, batch-normalization, and Parametrized Relu activation layer which act as feature extraction layer. This layer is stacked up in such a way that it creates a residual network. [13]

Further, this extracted feature is used to upsample the image using shuffle pixel which rearranges elements in a tensor of shape (*, $C \times r^2$, H, W) to a tensor of shape (∗, C, $H \times r$, $W \times r$), where r is an upscale factor[14].

SRGAN when used with Satellite images, has shown good results. A 54% increase was observed in SR of different satellite sources. [15]

C. *Yolov8*

YOLOv8 represents a state-of-the-art (SOTA) model that expands upon the achievements of earlier YOLO iterations while introducing novel features and enhancements. Its combination of speed and accuracy renders it an outstanding option for object detection tasks [16].

Yolo v8 provides pre-trained model which can be easily finetuned for custom use-cases. Hence, it is a suitable choice for object detection from Satellite Images.

III. Methodology

We are using DiffiR model for DM based super resolution model and SRGAN for GANs based super resolution model.

The DM model is trained with original setting to replicate the similar results on satellite image and same is done for SRGAN models also. The satellite images used in the current study are of 2 types i.e. 0.6 meter spatial resolution image classified as low resolution (LR) and 15 cm spatial resolution image classified as high resolution (HR). Image pairs were

downloaded for the same with resolution being 1024*1024 and 4096*4096 respectively. These images were used to create super-resolution (SR) image with a output resolution of 15 cm.

A. Approach for Diffusion model:

Since a satellite image can have large area cover in a single image, we first make patches of size 256 x 256 from 1024 x 1024 images for LR. We intend to go to 4x resolution of low-resolution image, and all the results are compared based on 4x resolution images. Similarly, we are creating 1024*1024 sized patches for 4K resolution image.

When the super-resolution is performed, each 256x256 image results into 1024x1024 resolution and combining all these resultant images, gives us an overall resolution of 4096x4096.

B. Approach for SRGAN model

Input for SRGAN model is 64x64 patches constructed from same input 1024x1024 (LR) and 256*256 patches constructed from 4K image (HR). The final resolution for both model is same i.e. 4096x4096.

After super resolution we ingest the image to our custom trained Yolov8 model and there we detect the certain class of objects from the images.

For comparison, the inference is on Yolov8 is done for three types of images, original 1024x1024 image, super resolution image from DM, and super resolution image from SRGAN model.

We also show comparative study of two types of satellite images, captured with 15cm spatial resolution and 60 cm spatial resolution.

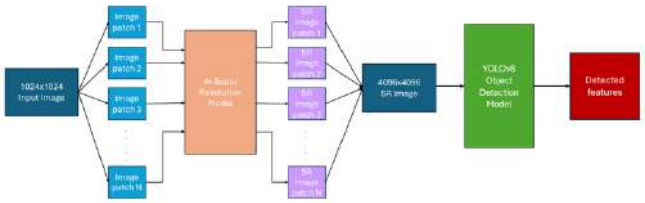

Fig. 1. Super-Resolution to object detection

All the models used in this study has been trained on custom dataset to convey the results on real-time dataset.

C. YoloV8

YoloV8 nano model was finetuned on 8 classes of road attributes. The following classes were used:

Direction Arrow, Stop Line, Bicycle Sign, Crosswalk, Stripes Crosswalk, Yield Indication Triangle, Bus Sign, Yield Indication, Lane Boundary

IV. RESULTS

TABLE I. EVALUATION METRIC

Model	PSNR	SSIM
SRGAN	27.72	0.75
DiffIR	34.44	0.90

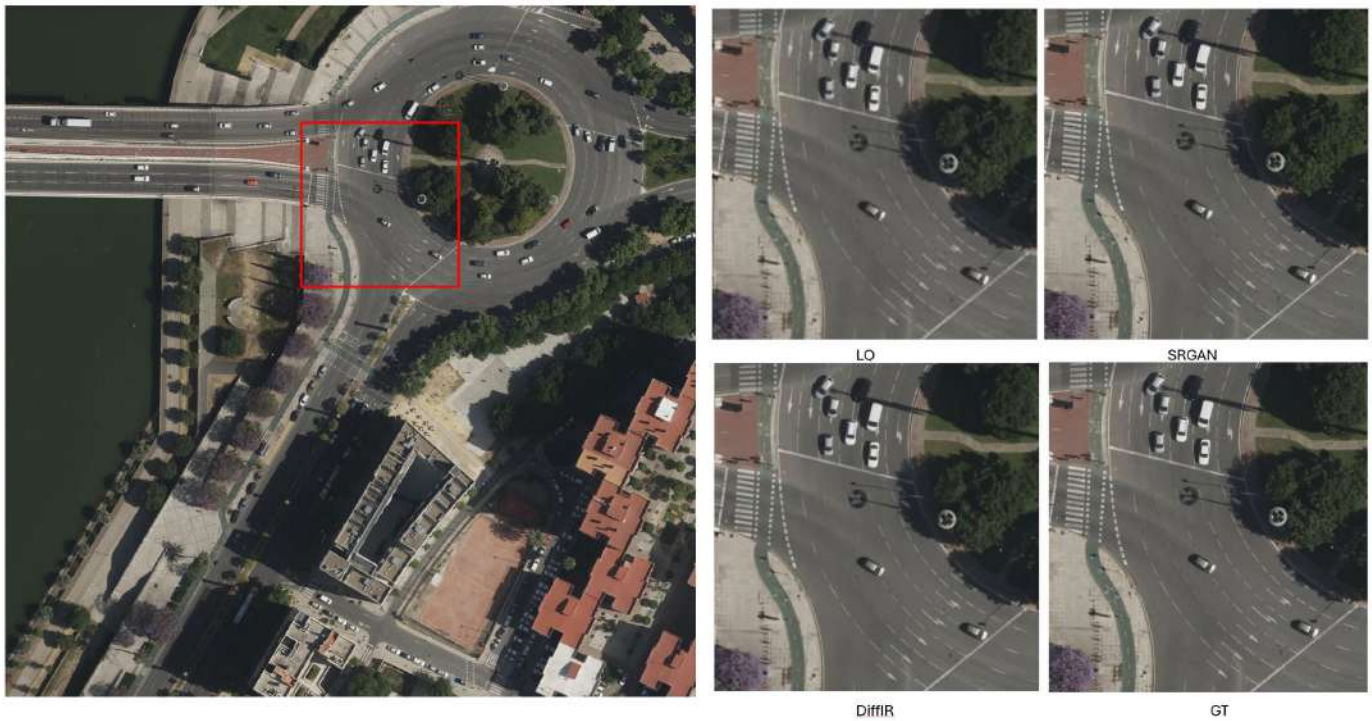

Fig. 2. The original low quality satellite image (LQ) from a source with a spatial resolution of 15cm compared to the output of SRGAN, DiffIR and the ground truth image (GT). Zoom in for a better comparison

Based on experimentation, the PSNR obtained during model training using the SRGAN model was 27.72 while its SSIM was 0.75. The diffusion model outperformed the SRGAN model by a good margin in terms of perceptual quality metrics with a PSNR of 34.44 and SSIM of 0.90 (refer Table 1). The visual differences between the two are compared with the input low quality image and ground truth high quality image in Fig 2.

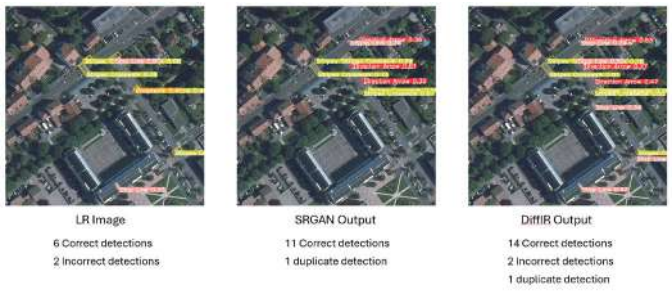

Fig. 3. Comparison of features detected from the low resolution (LR) image with a spatial resolution of 50cm, the output of SRGAN and DiffIR

Further, on applying our finetuned Yolov8 model for detection of different road features we observed that on raw LR image only 22% of the road attributes were detected while on the HR image we had a detection rate of 78%. SR of SRGAN model trained on our dataset showed a detection rate of 40% while SR of Diffusion model showed a detection rate of 52% .

V. CONCLUSION

The SRGAN and diffusion model were used to create output images of 15 cm from 60 cm input images. We observed that the Object detection model detected more features in the SR of the Diffusion model. The Super-resolution images not only had high perceptual quality but also very high SSIM and PSNR scores. This resulted in 136% more detection when compared with low-resolution images. Furthermore, we believe that in regions where we do not have high resolution images, the current model can help in producing high-quality images in a much more cost and time-saving manner.

As a next step we can leverage more dataset of various geographies to eliminate any bias that currently might exist, along with further development in loss function of L1 and VGG to further improve the quality of generated image.

REFERENCES

[1] Li, Z.; Wang, Y.; Zhang, N.; Zhang, Y.; Zhao, Z.; Xu, D.; Ben, G.; Gao, Y. Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey. Remote Sens. 2022, 14, 2385. [Google Scholar

[2] Karim, S.; Zhang, Y.; Yin, S.; Bibi, I.; Brohi, A.A. A brief review and challenges of object detection in optical remote sensing imagery. Multiagent Grid Syst. 2020, 16, 227-243. [Google Scholar]

[3] Burlina PM, Joshi N, Pacheco KD, Liu TA, Bressler NM. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. JAMA ophthalmology. 2019 Mar 1;137(3):258-64

[4] Zhengwei Wang, Qi She, and Tomás E. Ward. 2021. Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy. ACM Comput. Surv. 54, 2, Article 37 (March 2022), 38 pages. https://doi.org/10.1145/3439723

[5] C. Ledig, et al., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017 pp. 105-114. doi: 10.1109/CVPR.2017.19

[6] J. Bruna, P. Sprechmann, and Y. LeCun. Super-resolution with deep convolutional sufficient statistics. In International Conference on Learning Representations (ICLR), 2016. 2, 3, 5

[7] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super- resolution. In European Conference on Computer Vision (ECCV), pages 694–711. Springer, 2016. 2, 3, 4, 5, 7

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR)), 2015. 2, 3, 4, 5

[9] Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. ArXiv. /abs/1503.03585

[10] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. ArXiv. /abs/2006.11239

[11] Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., & Van Gool, L. (2023). DiffIR: Efficient Diffusion Model for Image Restoration. ArXiv. /abs/2303.09472

[12] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng,Zhihai Xu, Qi Li, Yueting Chen,SRDiff: Single image super-resolution with diffusion probabilistic models, Neurocomputing,Volume 479,2022,Pages 47-59,ISSN 0925-2312,https://doi.org/10.1016/j.neucom.2022.01.029.

[13] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2016). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. ArXiv. /abs/1609.04802

[14] Pixelshuffle¶ (no date) PixelShuffle - PyTorch 2.2 documentation. Available at: https://pytorch.org/docs/stable/generated/torch.nn.PixelShuffle.html (Accessed: 10 April 2024).

[15] Puri, J. S. and Kotze, A.: Evaluation of SRGAN Algorithm for Superresolution of Satellite Imagery on Different Sensors, AGILE GIScience Ser., 3, 57, https://doi.org/10.5194/agile-giss-3-57-2022, 2022.

[16] Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics