Augmenting Satellite Imagery for Airplane Detection Using Stable Diffusion Inpainting

Shawn Oyer
Drexel University, CS 614

7/26/2024

Abstract

The increasing volume and complexity of satellite imagery data necessitate advanced techniques for automated object detection and segmentation. This project explores the application of Stable Diffusion, specifically Inpainting models for augmenting satellite imagery data, particularly focusing on enhancing object detection of airplanes. Traditional data augmentation methods often fall short in scenarios involving specific object localization and diverse backgrounds. Leveraging Stable Diffusion's Inpainting capabilities, this project aims to generate synthetic images by retaining the position of objects-of-interest while varying backgrounds or objects within predefined masks. The methodology includes using pretrained models, generating synthetic images through various prompts, and fine-tuning with Textual Inversion. Frechet Inception Distance (FID) and Structural Similarity Index (SSIM), were employed as evaluation metrics to assess the quality of the generated images. Results indicate that while fine-tuning improved background variation, the quality of the generated airplanes could still be enhanced. The study demonstrates that Stable Diffusion Inpainting can significantly contribute to augmenting satellite imagery for object detection tasks, although further refinement in creating more realistic objects is required.

1 Introduction

The increasing amount of data collected from satellite imagery poses challenges for imagery analysts and Earth scientists, especially in extracting meaningful intelligence. Automated target recognition (ATR) is crucial for various applications, such as site monitoring, discovery and tracking of objects, analyzing land management, and humanitarian aid and disaster relief [6]. Traditional data augmentation methods include rotation, flipping, and color changes, but they are limited in complex scenarios. With the advent of text-to-image models, it has become possible to perform more complicated data-augmentations than just simple image-level transformations [11]. This project uses Stable Diffusion Inpainting to augment satellite imagery for better object detection.

2 Background

Satellite imagery-based computer vision applications face challenges due to low pixel resolution and the detection of small objects in large-scale images. In recent years, different problems have been solved related to object detection in satellite imagery using ML and DL techniques. There are many available pipelines (SSD (Single Shot Detector), U-Net, Masked Image Generators, Stable Diffusion, YOLO (You Only Look Once) just to name a few) that are all feasible frameworks to use for different types of object augmentation and segmentation [5].

Object segmentation is a vital step in satellite imagery-based computer vision applications and presents a very complicated task due to various reasons including challenges in identifying class variations, multiple objects in close space, high variances in object size, differences in illumination and dense backgrounds [2]. Data scarcity is also a major issue while using satellite imagery as there is limited augmented data sufficient to train machine learning (ML) and deep learning (DL) models [1].

This project employs Stable Diffusion Inpainting to overcome these challenges by generating synthetic images with varied backgrounds and objects.

3 Methodology

Specifically, this project aims at two kinds of augmentations:

- Retain an object and generate different backgrounds
- Retain the background and generate different objects inside the mask

3.1 Data Source

The data can be downloaded here: data download
The data consists of:

- Real Training Images: 5825 images in PNG format
- Real Test Images: 2710 images in PNG format
- Synthetic Training Images: 2000 images in PNG format
- Real and Synthetic Masks in JSON format

3.2 Model and Data Justification

This project will use a Stable Diffusion pipeline which belongs to a class of deep learning models called diffusion models. They are generative models, meaning they are designed to generate new data similar to what they have seen in training. It generates artificial intelligence (AI) images from text as a text-to-image-model [13]. The original intent of Stable Diffusion was to generate new images from pure noise. However, one modification that has been gaining attention is the process of Inpainting [4].

Stable Diffusion Inpainting models are very well suited to achieve the goal of retaining the bounding-box/mask while generating diverse images. They can be used to remove defects and artifacts, or replace an image area with something entirely new [7]. They involve automatically filling in the missing regions of an image with plausible content, such that the filled-in areas appear visually coherent and seamless with the surrounding areas. By masking certain areas of an image, Stable Diffusion can be directed to modify only the masked areas by iteratively re-noising and then denoising them based on a textual prompt [4]

The Stable Diffusion model, specifically Inpainting, was chosen for its ability to retain the position of objects-of-interest while varying other elements of the image. This model is known for its stability and smooth results, making it suitable for images with intricate structures like airplanes [12]. This specific dataset was selected as it contained sufficient training and testing satellite images in addition to containing masks for all of the airplanes in all of the images.

Textual Inversion will be used for the fine-tuning of the model. Textual Inversion is a training technique

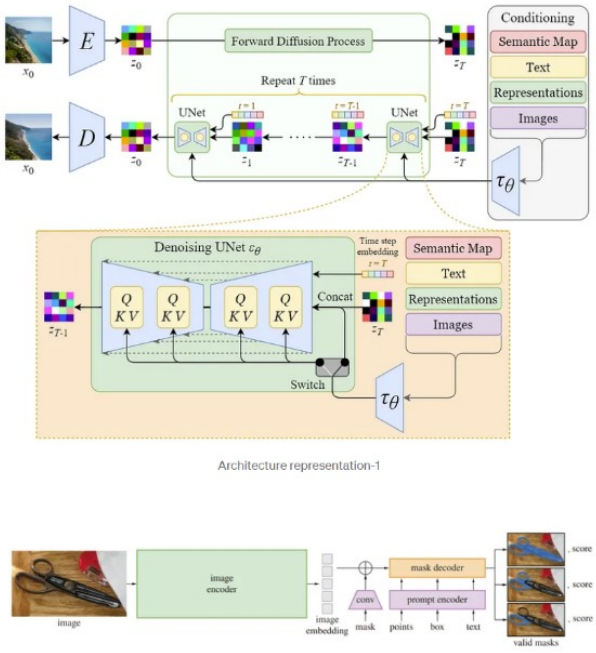

Figure 1: Stable Diffusion Inpainting Model [9]

for personalizing image generation models by only providing a few example images of what the model will learn. The model learns and updates text embeddings to match the example images provided [8].

The goal will be to generate 1000 new synthetic images by creating custom prompts using random foreground and background adjectives/nouns such as (big, small, long, red, blue, green, desert, forest, snow, etc) to allow varying backgrounds as well as varying features of the airplane (See Figure 2).

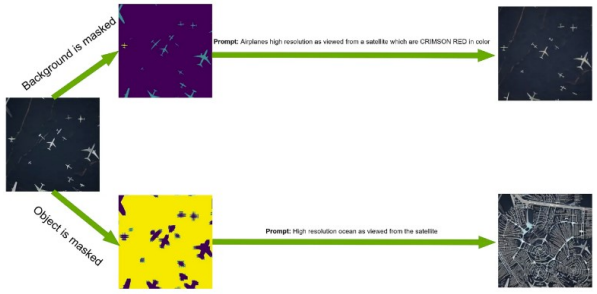

Figure 2: Procedural Model [11]

3.3 Procedure

1. Load images and masks.

2. Use the pre-trained Stable Diffusion Inpainting pipeline provided by Hugging Face.

3. Generate 1000 synthetic images using custom and random prompts.

4. Fine-tune the model with Textual Inversion to improve output.

5. Evaluate using FID and SSIM metrics.

4 Experiments and Results

4.1 Model Output Examples

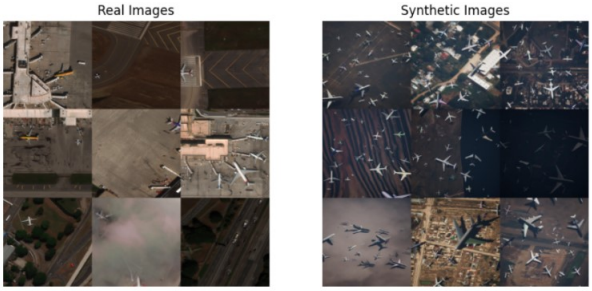

Figure 3: Real Training Images and Manually Processed Synthetic Images

The images were included in the original dataset that were used to train the model (see Figure 3).

Figure 4: Images after 1st model run with Stable Diffusion Inpainting without fine-tuning

After the model's first run, the masks were kept in place, however, the model had not been trained on this distribution of airplane images causing them to look almost like clouds as opposed to actual airplanes. The background did not change very much and doesn't match the random prompt selections such as desert, snow, forest, or sea (see Figure 4).

After the model's fourth run, the diffusion model did a much better job varying the backgrounds as there is a forest, urban, runway, ocean, desert, etc visually displayed as backgrounds. However, even though the masks stayed where they should, the airplanes still lack realistic effects and don't take on different colors and shapes as the prompts indicated (see Figure 5).

Figure 5: Images after last model run with Stable Diffusion Inpainting with fine-tuning

Figure 6: Masking process with original image, masked image, and inverted masked image

The masking process loads the annotations, initialize the mask image, aggregates the masks, ensures masks are binary, and shifts the masks around the nearby pixels. The Mask image is used for the foreground generation and the inverted mask image is used for the background generation (see Figure 6).

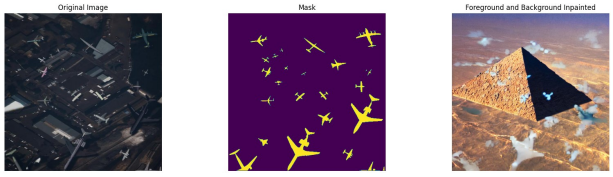

Figure 7: Testing Example

The testing example (see Figure 7) shows the original image, masked image, and foreground/background Inpainting output of a foreground and background prompt that the user inputs based on their preference. The foreground prompt used in this example was:"A Large airplane high resolution as viewed from a satellite with Blue wings." The background prompt used in this example was: "High resolution Pyramids as viewed from a satellite."

Model Runs	Model Type	Model Parameter Changes	# of synthetic images generated	FID	SSIM
0	Manual	N/A	2000	189.8	0.34
1	Stable Diffusion - Inpainting not fine-tuned	None	1000	274.67	0.23
2	Stable Diffusion - Inpainting fine-tuned	Used 10 textual inversion images with hyperparameters	1000	243.46	0.20
3	Stable Diffusion - Inpainting fine-tuned	Used 20 textual inversion images, added more prompt options	1000	219.59	0.19
4	Stable Diffusion - Inpainting fine-tuned	Used 20 textual inversion images, added more prompt options, refined mask handling	1000	221.41	0.19

Figure 8: Results Table

5 Results

The table shows the model was run four times, the model types (either not fine-tuned or fine-tuned), what changes occurred in the model parameters with each model run, the amount of synthetic images generated with each model, the FID and the SSIM score for the generated images.

5.1 Evaluation Metrics

The Frechet Inception Distance (FID) and the Structural Similarity Index (SSIM) were used in this project. The FID score is a metric that calculates the distance between vectors calculated for real and generated images. The score summarizes how similar the two groups are in terms of statistics. Lower scores indicate that two groups of images are identical and have been shown to correlate well with higher quality images [3]. The SSIM is a metric that quantifies image quality degradation caused by processing such as data compression or by losses in data transmission. It measures the similarity between two images with values ranging from -1 to 1 (1 means a perfect match, 0 means no structural similarity, -1 means images are structurally dissimilar) [10]. This combination of FID and SSIM provides a comprehensive evaluation of the quality of inpainted images .

In reviewing the average FID scores for the generated images, the best score (189.8) is from the manual synthetic images, which makes sense considering a human manually created the images based on real imagery and masks. In regards to the automated methods, the best score (219.59) is from the 3rd model run where 20 textual inversion images were used for testing as well as added more background prompt options such as urban, runway, parking apron.

In reviewing the average SSIM scores for the generated images, the best score (0.34) is from the manual synthetic images. In regards to the automated methods, the best score (0.23) is from the 1st model run with no fine-tuned parameters. The scores (0.19) were lowest from the 3rd and 4th model run with more textual inversion images, more prompt options and more refined mask handling. Even though the scores were low, a low SSIM is acceptable if the generated backgrounds and foregrounds are meant to be distinct or highly varied from the original dataset. Since the goal was to create a diverse set of images that do not necessarily need to resemble the original set closely, a lower SSIM would be more tolerable [10].

6 Discussion

Initial results without fine-tuning showed limited effectiveness in changing backgrounds and realistic rendering of airplanes. However, after fine-tuning the model with Textual Inversion, expanding prompt options, and refining mask handling, the background diversity improved significantly. Despite these advancements, the generated airplanes remained suboptimal, indicating the need for further refinement. The combination of FID and SSIM metrics provided a comprehensive assessment of image quality, highlighting the trade-offs between background diversity and object detection.

7 Conclusion

This project used Stable Diffusion Inpainting for augmenting satellite imagery, focusing on the detection of airplanes. The use of Stable Diffusion models, specifically inpainting, allowed for the generation of synthetic images by retaining object positions while varying backgrounds and object parameters. Overall, this project underscores the potential of generative models in addressing data scarcity issues in satellite imagery while also pointing out areas for further improvement in achieving better quality augmented data.

References

[1] Aayushibansal, M., et al. (2020). A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications. ACM. Retrieved from url

[2] Arsalan, T., et al. (2022). Automatic Target Detection from Satellite Imagery Using Machine Learning. NCBI. Retrieved from url

[3] Brownlee, J. (2019). How to Implement the Frechet Inception Distance (FID) for Evaluating GANs. Machine Learning Mastery. Retrieved from url

[4] Cheng, W. L. (2023). How to Build Your Own AI-Generated Image with ControlNet and Stable Diffusion. Datature. Retrieved from url

[5] Cole, R. M. (2024). Techniques for deep learning with satellite and aerial imagery. GitHub. Retrieved from url

[6] Groener, A., et al. (2020). A Comparison of Deep Learning Object Detection Models for Satellite Imagery. Arxiv. Retrieved from url

[7] Hugging Face. (2024). Inpainting. Retrieved from url

[8] Hugging Face. (2024). Textual Inversion. Retrieved from url

[9] İlaslan, D. (2023). Stable Diffusion-Inpainting Using Hugging Face Diffusers with Serving Gradio. Medium. Retrieved from url

[10] Imatest. (2024). SSIM: Structural Similarity Index. Retrieved from url

[11] Koneripalli, K. (2023). Satellite Image Data Augmentation using Stable Diffusion for Object detection and segmentation. Medium. Retrieved from url

[12] Kumari, P. (2023). Inpainting with Stable Diffusion: Step-by-Step Guide. Lancer Ninja. Retrieved from url

[13] Stable Diffusion Art. (2024). How does Stable Diffusion work? Retrieved from url