Leveraging Generative AI for Cross-Regional Small Object Detection in Satellite Imagery

Zheyang Qin[a], Stanislav Panev[a], Celso De Melo[b], Shayok Chakraborty[c], Jessica Hodgins[a], and Fernando De la Torre[a]

[a]The Robotics Institute, Carnegie Mellon University, USA
[b]DEVCOM Army Research Lab, USA
[c]Department of Computer Science, Florida State University, USA

ABSTRACT

Detecting small objects like cars in satellite imagery is challenging, especially when labeled data is unavailable for a target region. This study introduces a novel approach using generative AI to address this issue by fine-tuning a Stable Diffusion model on images from one geographical region, the source (Selwyn, New Zealand), and a second region of interest, the target (Utah, USA) environment. Our framework employs state-of-the-art cross and self-attention mechanisms alongside the CLIPSeg image segmentation method to generate high-quality synthetic datasets with minimal supervision. Our empirical results show a significant improvement in detection performance, improving the accuracy by over 20% when testing the target dataset compared to the performance of a baseline model trained only on source data.

Keywords: Diffusion Model, Generative AI, Domain Adaptation, Satellite Imagery

1. INTRODUCTION

Satellite imagery plays a pivotal role in a wide array of applications, ranging from environmental monitoring to military surveillance.[1,2] Within this domain, detecting small objects such as vehicles is particularly important for tasks including threat assessment and strategic planning.[3,4] However, one of the key challenges in applying object detection models to satellite data is the limited availability of annotated imagery, particularly in novel or high-priority regions. This lack of labeled data hinders the ability of pre-trained models to generalize effectively across different geographical domains.

Different geographical regions can be either image-sparse or annotation-sparse for several reasons. Areas with restricted airspace, adverse weather conditions, or extensive cloud cover are more likely to lack available imagery. Even when images are abundant, some regions may still suffer from insufficient granularity of the annotations. Additionally, satellite imagery comes with specific challenges that can hinder the performance of pre-trained models on new datasets without fine-tuning. For instance, sensor characteristics, such as *ground sampling distance* (GSD), are usually consistent within a single dataset, which means that pre-training datasets must include a variety of parameters. Moreover, satellite data platforms like *Planet Labs*[5] *NASA Worldview*,[6] and *Sentinel Copernicus*,[7] provide images including various spectral bands ranging from visible light to ultraviolet and infrared. This multispectral nature can additionally complicate zero-shot detection. Lastly, various geographical regions often exhibit different data distributions, including factors like vegetation, lighting conditions, and types of urbanization. These differences can also lead to poor detection performance.

This is not only a problem for object detectors in deployment but also makes it difficult for automatic annotation pipelines, which rely on pre-trained detection or segmentation models to create point or polygon

Further author information: (Send correspondence to F.T.)
Z.Q.: E-mail: zheyangqin.qzy@gmail.com
S.P.: E-mail: spanev@andrew.cmu.edu
C.M.: E-mail: celso.m.demelo.civ@army.mil
S.C.: E-mail: shayok@cs.fsu.edu
J.H.: E-mail: jkh@cs.cmu.edu
F.T.: E-mail: ftorre@cs.cmu.edu

annotations of objects. Suppose the region in question has not been adequately represented in the training dataset (along with the GSD, spectral bands, etc.). In that case, the annotation model will likely not perform well, leaving manual annotation as the only option for trustworthy annotations.

To address these challenges, we introduce a novel framework that leverages *Generative AI* to produce synthetic, annotated satellite imagery tailored to target regions. By fine-tuning a *Stable Diffusion* (SD)[8] model on both source *(e.g., Selwyn, New Zealand)* and target *(e.g., Utah, USA)* domains, our method generates realistic satellite images representative of the target distribution. We then apply an automatic annotation pipeline that combines cross-attention and self-attention[9] heatmaps with segmentation models to label the generated data with minimal human supervision.

Our contributions are summarized as follows:

1. We propose a generative framework for producing region-specific synthetic satellite imagery using fine-tuned diffusion models.
2. We introduce a dual-method annotation pipeline combining *DAAM*,[10] *DAAMi2i*,[11] *CLIPSeg*,[12] and *DatasetDM*[13] to generate high-precision labels for synthetic data.
3. We demonstrate significant improvements in cross-regional small object detection performance by incorporating the generated dataset into training.

This approach enables the deployment of robust object detection systems in underrepresented regions, improving readiness for various monitoring and surveillance applications.

2. RELATED WORK

The primary challenges in the cross-domain vehicle detection task from satellite imagery stem from restricted access to imagery from the target domain or insufficient available labels. Often, the data provided is minimal, and generating annotations is a lengthy process, frequently infeasible given the available satellite imagery. Traditional data augmentation techniques that rely on a moderate number of samples to perform augmentation may not work for a satellite imagery application, as the limited number of samples may not appropriately capture the true distribution of the data. Moreover, utilizing the same augmentation strategies across different datasets may produce sub-optimal results due to the variations in their underlying distributions. Two approaches can be implemented to overcome this challenge: generating synthetic data and annotating existing datasets.

2.1 Generating Synthetic Datasets

In the realm of synthetic imagery generation, the two popular methods are *Generative Adversarial Networks* (GAN)[14] and *Diffusion-based Models* (DM).[15] At the basic level, the GAN approaches use two networks: *the generator*, which generates the images, and *the discriminator*, which tries to classify whether the image is real or synthetic. GANs can produce high-quality images by optimizing both networks, and are pretty versatile. However, GANs can also suffer from mode collapse, which causes low variability in the generated data, which can be a problem with small datasets. On the other hand, diffusion-based models exploit a many-step framework that has a forward process of adding noise to an image until it becomes pure noise and a backward process of reconstructing the image from the pure noise state. These models are stable and can generate imagery with fine detail and complexity, but they cost time and computational power. DMs have recently demonstrated improved performance over GAN models for synthetic imagery generation.[8]

2.2 Dataset Labeling

We review two main types of dataset labeling techniques: detection and segmentation. In the realm of detection, the primary method *Open Vocabulary Object Detection* (OVOD).[16] This method uses a two-stage framework, starting with an abundant set of image-caption pairs for an unbounded vocabulary and then learning object detection downstream using annotations only. Newer methods of OVOD also use visual language models instead of image-caption pairs,[17] and many studies have been into advancing Open Vocabulary Detection.

However, detection methods may provide insufficient annotation information; instead, segmentation would provide more information within its annotations. Research has also been advancing in *Open Vocabulary Segmentation* (OVS), the same technique applied to panoptic segmentation. Several open-vocabulary segmentation models exist, including the *Segment Anything Model* (SAM).[18] The SAM model is a zero-shot segmentation model that allows for annotations of various granularities to be generated for any image. The SAM model can also deal with ambiguity in masks, which is helpful in pipelines that produce annotations on images without prompts. Another approach is *DatasetDM*[13]—a diffusion-based method with paired annotation masks for generated synthetic imagery. This approach requires training using the target dataset but only requires a small amount of data to be manually labeled to create a large annotated dataset under the same distribution.

DAAM[10] is yet another method that uses pre-trained diffusion models to produce pixel-level annotations and synthetic imagery from the given prompt, using word-pixel cross-attention maps, which can provide more accurate segmentation annotations. Finally, *CLIPSeg*[12] is a model that builds upon the *CLIP*[19] model with a transformer decoder that enables dense predictions. This approach generates synthetic imagery from either a text or image prompt and provides segmentation annotations.

3. REAL DATASETS

In this work, we use two overhead view aerial image datasets from two distinct geographical locations: *Selwyn, New Zealand* and *Utah, USA*. The visual data (georeferenced image tiles) were obtained from public online platforms such as *LINZ*[20] and *UGRC*[21] (hence their names). Our team manually annotated the location (object centers) of all small vehicles in these images. Both datasets have a GSD of 12.5 cm per px. The sample size used in our experiments is 112 × 112 px. The image count statistics of both datasets are presented in Table 1.

Table 1: Real Datasets Image Quantities

Data splits	Property	Real LINZ Dataset	Real UGRC Dataset
Training	Total image count	1,451,144	2,142,849
	Images containing cars	19,564	15,631
	Images without cars	1,431,580	2,127,218
Validation	Total image count	188,744	271,252
	Images containing cars	2,108	3,912
	Images without cars	186,636	267,340
Test	Total image count	438,189	270,557
	Images containing cars	2,629	1,510
	Images without cars	435,560	269,047

We generated captions for all images in both datasets using *BLIP*,[22] which are necessary for fine-tuning the Stable Diffusion generative model. LINZ images differ significantly from UGRC regarding image texture, color, and landform, as demonstrated in Figure 1. In this project, we treat UGRC as the target environment with no annotated data, while LINZ is the source environment with all images annotated.

4. METHOD

We propose a framework for automatically generating annotated satellite/aerial view datasets to enhance the performance of downstream object detectors. Our approach begins by fine-tuning a *Stable Diffusion* model with images from both source and target domains. Thus, it is adapted to generate high-quality satellite imagery resembling the target domain. To annotate the synthetic images, we developed two methods—heuristics-based and learning-based. The first one combines three methods—*DAAM*,[10] *DAAM-i2i*,[11] and *CLIPSeg*[12]—to produce heatmaps that highlight regions of an image most influenced by specific words, such as "car", from the text prompt used during image generation. After post-processing, we apply a segmentation technique to each heatmap, generating annotations that indicate the locations of objects like cars. Only images with consistent annotations

(a) LINZ dataset images

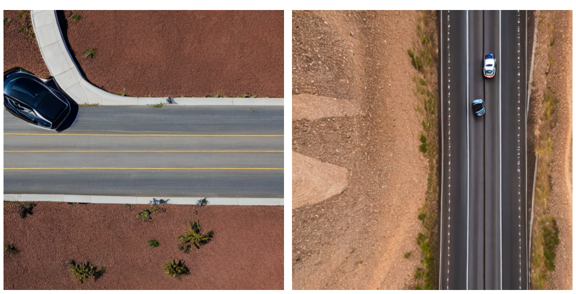
(b) UGRC dataset images

Figure 1: Visualization of the source LINZ *(left)* and the target UGRC *(right)* datasets used in this study.

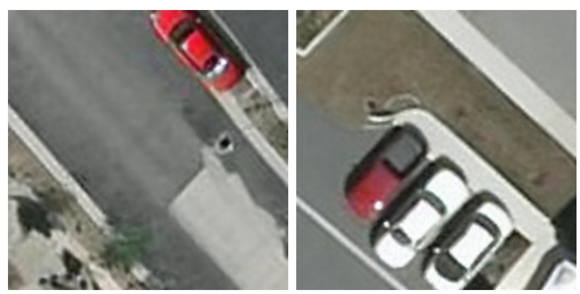
across all three heatmaps are retained to form the final synthetic dataset. The second annotations generation approach is based on the *DatasetDM*.[13] Like the method above, it relies on cross-attention information from Stable Diffusion's denoising U-Net, but it trains a small network to predict the final annotations. The training requires a fully annotated image dataset, and we use Real LINZ for this purpose.

4.1 Region-specific Synthetic Data Generation using Fine-tuned Stable Diffusion

Our empirical experience with the off-the-shelf Stable Diffusion model shows that it struggles to generate high-quality satellite/aerial view images (Figure 2a). That suggests that this view and scale were not well presented in the training set of this foundational latent diffusion model. To this end, we employ the *DreamBooth*[23] technique to adapt *Stable Diffusion 1.4* by fine-tuning it on data from our source and target environments. Note that this process does not require any object annotations and relies only on images and the textual captions generated by BLIP. It enables the model to embed new concepts while retaining its broader capabilities. To let the model learn how to discriminate between both environment styles, we append a unique token to the captions from each dataset (e.g., "in LINZ" for LINZ images and "in Utah" for UGRC images). Thus, the model is guided to learn region-specific visual features (Figure 2b).



(a) Images generated by the original SD



(b) Images generated by fine-tuned SD

Figure 2: Visualization of images synthesized from the base SD model *(left)* and the fine-tuned SD model *(right)*

Furthermore, to reduce the VRAM usage during training, we use mixed precision that stores non-trainable parameters in fp16. We also implement gradient checkpointing, which minimizes memory requirements by saving essential checkpoints during backpropagation.

4.2 Producing Automatic Annotations

4.2.1 Heuristics-based approach

The annotation module aims to generate reliable heatmaps and use the segmentation method to create accurate annotations from heatmaps. We achieve this by using cross-attention and self-attention layers in Stable Diffusion. Specifically, we utilize DAAM, DAAMi2i, CLIPSeg, and a proposed mixed-heatmap.

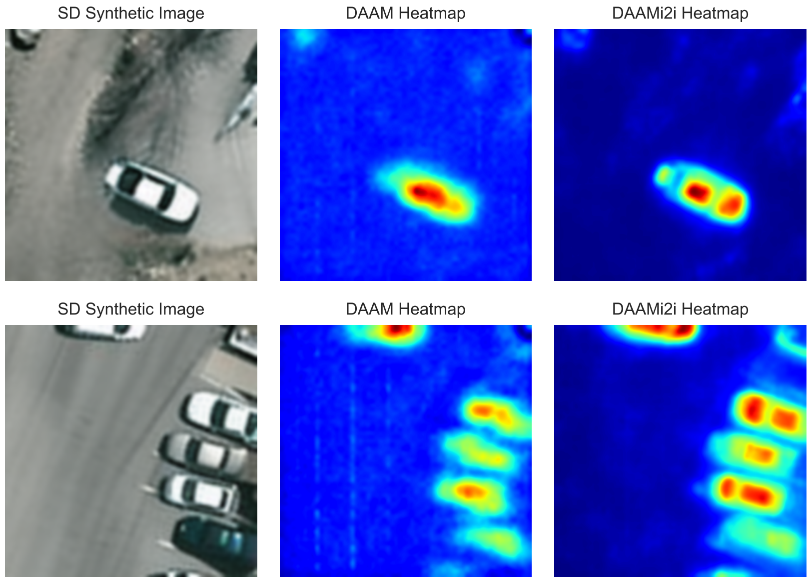


Figure 3: Examples of DAAM *(center)* and DAAMi2i *(right)* cross-attention heatmaps. The image locations representing cars have "hotter" pixels in the heatmaps, suggesting that they were influenced by the word "car" during generation.

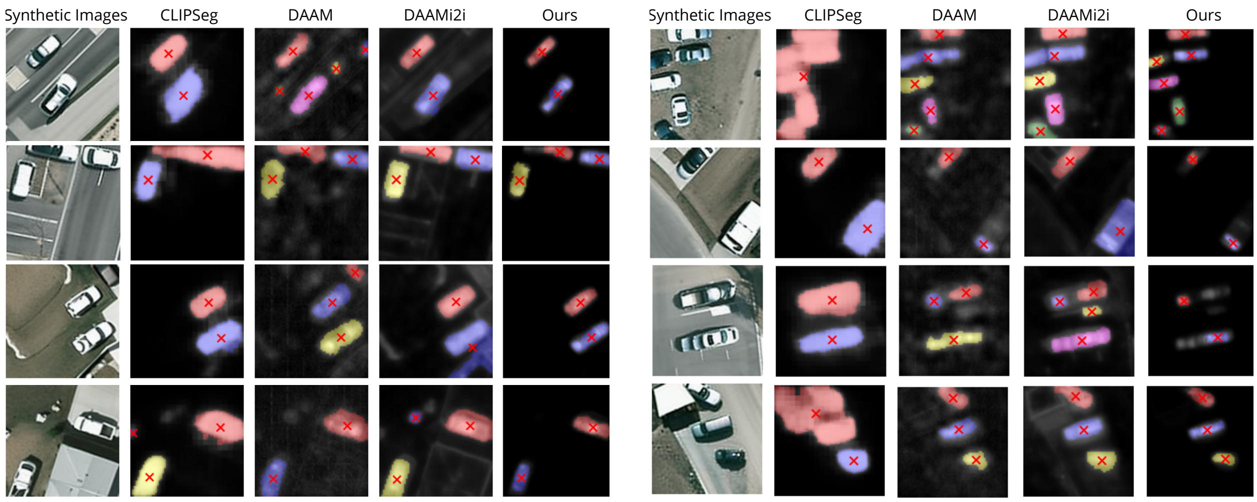
Diffusion Attentive Attribution Maps[10] (DAAM) identify regions in an image influenced by specific words (tokens) in the text input (Figure 3-*center*). This is achieved by aggregating cross-attention scores in the denoising sub-network of the diffusion model. It can provide a clear mapping of textual tokens to image features. In general, the heatmaps obtained from DAAM tend to be noisy. Therefore, we used DAAMi2i,[11] which extends DAAM by incorporating self-attention information. While DAAM focuses on text-to-image relations, DAAMi2i captures relationships within the image, making object boundaries more distinct (Figure 3-*right*). Heatmaps from an additional zero-shot image segmentation model, CLIPSeg,[12] are also utilized to make the pipeline even more robust.

The quality of individual heatmaps is often unreliable. For instance, the CLIPSeg heatmap (H_{CLIPSeg}) effectively identifies cars but lacks precise boundary definitions. Conversely, the heatmaps generated by DAAM (H_{DAAM}) and DAAM-image2image (H_{DAAMi2i}) provide clearer boundaries but occasionally fail to detect cars. To address these limitations, a combined heatmap approach is proposed. The mixed heatmap is computed as:

$$H_{\mathrm{mix}} = H_{\mathrm{DAAM}} \odot H_{\mathrm{CLIPSeg}} \odot H^2_{\mathrm{DAAMi2i}}. \tag{1}$$

The DAAMi2i heatmap is squared to emphasize and enhance the boundary clarity of high-temperature regions. Annotations are valid only when all four heatmaps yield a consistent segmentation mask center. Figure 4a illustrates the differences between the individual heatmaps and the resulting mixed heatmap, highlighting the improved boundary clarity and detection robustness achieved through this combination.

Despite these techniques' strengths, several challenges arose during implementation. For instance, when an image contained multiple cars, the keyword "car" often failed to map accurately to all instances. Large vehicles could appear as multiple heat sources, and overlapping cars or unfocused heatmaps frequently introduced noise. To address these issues, we developed a set of post-processing rules, which rely on defining thresholds for high-temperature and fuzzy pixels. High-temperature pixels have values greater than 0.8, while fuzzy pixels exceed 0.4. Images were filtered out under specific conditions. For instance, an image was rejected if high-temperature pixels comprised less than 15% of the fuzzy region, indicating a loss of focus. If high-temperature pixels covered more than 20% of the total image area, which suggests numerous cars, or if more than 40% of pixels fell within the fuzzy region, revealing excessive noise, the image was excluded. These measures reduce the likelihood of making errors when subsequent segmentation is performed.



(a) Mixed heatmap showing improved labeling (b) Example of Rejected Heatmap

Figure 4: Heatmap processing for automatic synthetic data annotations.

We apply the *watershed algorithm*[24,25] to generate precise annotations of objects from all four heatmaps generated. The watershed algorithm treats each heatmap as a topographical surface, where the intensity of the pixels represents elevation and "floods" the areas of the markers, creating boundaries between neighboring regions as they meet. This approach is relatively effective for delineating small, closely spaced objects like cars in satellite imagery. In addition to ensuring segmentation consistency, the proposed method addresses two additional failure conditions. The first condition arises when the generated cars exhibit abnormal or unrealistic shapes, compromising their visual plausibility. The second condition occurs when multiple cars overlap within the heatmap, leading to ambiguities in object representation and a loss of structural clarity. By applying these three conditions, the approach enhances both annotation accuracy and the image generative quality, ensuring a robust and realistic result. An illustrative example of these failure conditions is provided in Figure 4b, demonstrating the challenges posed by inconsistent heatmaps, abnormal shapes, and overlapping car instances.

4.2.2 Learning-based Approach

We developed a second image annotation generation approach based on the DatasetDM method,[13] which complements the previous section's method. DatasetDM could not generate bounding boxes or object location points as it is. Therefore, we employ its capability to work with semantic segmentation maps. We compose pseudo-segmentation maps by drawing small squares at each car location from the Real LINZ dataset. Thus, we train DatasetDM's P-Decoder network using the entire Real LINZ train set. During inference, it utilizes the cross-attention features generated by Stable Diffusion to predict each synthetic image's pseudo-segmentation maps. In the post-processing stage, each predicted square region from the segmentation maps is converted into an object location, constituting the final image annotation.

4.3 Object Detection

Based on the discussions in the preceding sections, the proposed method demonstrates the capability to generate synthetic datasets and perform autonomous annotations. This approach is a robust foundation for addressing the cross-regional object detection problem. To evaluate its effectiveness, the proposed method utilizes *YOLOv5*[26] as a baseline model. YOLOv5 is selected for its optimal balance between speed and accuracy, which makes it particularly well-suited for real-time object detection tasks. Moreover, its capability to handle objects of varying scales addresses the specific challenges posed by satellite imagery, where vehicles often appear in significantly different sizes depending on their spatial resolution and distance from the camera.

As described above, the annotations of our datasets (real and synthetic) contain only object locations, i.e., x and y coordinates of the objects' centers in the image. Since most off-the-shelf object detectors rely on bounding box-based annotations, we generate pseudo bounding box annotations by placing a square box at each object's center. The length of the side of the boxes is set to 42.36 px, which defines an approximate circular decision boundary with a radius of 12 px around each ground truth annotation. This boundary determines whether a prediction is positive or negative when using a bounding box IoU threshold of 0.5.

5. EXPERIMENTAL RESULTS

5.1 Quality of Generated Synthetic Images

After fine-tuning the Stable Diffusion model, it can produce data more closely resembling real-world Utah dataset images. As shown in Figure 2, the same BLIP-generated text caption is used as an input for both the original Stable Diffusion model and the fine-tuned one. We observed that the original Stable Diffusion 1.4 struggles to generate images with the correct style and viewpoint based on the text caption; even the shapes of cars, buildings, and roads appeared inconsistent. In contrast, the fine-tuned model generated correct views and styles that closely resembled the satellite images of the UGRC dataset.

In addition to visual comparisons, the *Fréchet Inception Distance* (FID) score[27] was computed between the real-world UGRC dataset and the synthetic images generated by the original and fine-tuned SD 1.4, as shown in Table 2. The FID score assesses the diversity and similarity of synthetic images relative to the reference dataset. In this experiment, $3,000$ images are generated for analysis for both the original and fine-tuned models. The fine-tuned model achieved a substantially lower FID score than the original Stable Diffusion model, indicating improved fidelity to the reference dataset.

Table 2: FID-based similarity comparison between two synthetic Stable Diffusion datasets—with and without fine-tuning on target images. The scores are calculated using the Real UGRC dataset as a reference.

Dataset Name	**FID Score ↓**
Synthetic UGRC (Original SD)	168.86
Synthetic UGRC (Fine-tuned SD)	**34.53**

5.2 Detection Performance

5.2.1 Quantitative Evaluation

The YOLOv5 model was trained under several configurations to investigate the utility of the proposed synthetic datasets. The baseline model was trained only on the Real LINZ dataset and tested on the entire Real UGRC dataset to quantify the performance when no synthetic data is used for training. Subsequently, three additional models were trained using a mixture of the Real LINZ dataset and three synthetic datasets generated by the fine-tuned SD and annotated by the methods described in Section 4.2. For the first model, we use a Synthetic UGRC dataset annotated using the method described in Section 4.2.1 *(heuristics-based automatic annotations using DAAM)*. For the second experiment, we train the object detector using a Synthetic UGRC dataset annotated by the method described in Section 4.2.2 *(learning-based automatic annotations using DatasetDM)*. The third experiment combines the datasets produced by the previous two methods.

The results of these experiments are presented in Table 3. We use *Average Precision* (AP) as the evaluation metric. Each of the automatic annotation generation techniques delivers a similar level of performance gain—about 16% AP advantage over the baseline model. Notably, the combination of both annotation techniques yields an additional performance boost, reaching 68.4% AP (21.1% above the baseline).

In addition to the previous set of experiments, we also assessed the performance of manually labeled synthetic data. Therefore, we produced manual annotations for all previously generated synthetic datasets above and trained the detector again using them. The results are presented in Table 4. Even though the automatic labels deliver a significant performance boost over the baseline model (Table 3), we observe a considerable performance gap between the automatic and manual labels (Table 4). The manual labels deliver 8-14% better AP than the automatic ones. We speculate that this results from inconsistencies in SD's cross-attention mechanism caused by

Table 3: Baseline vs. synthetic data with automatic labels performance comparison.

Training Data	Test AP_{50} ↑	AP Gain ↑
Real LINZ *(baseline)*	47.3%	—
Real LINZ + Synthetic UGRC (DAAM annot.)	63.8%	+16.5%
Real LINZ + Synthetic UGRC (DatasetDM annot.)	63.7%	+16.4%
Real LINZ + Synthetic UGRC (DAAM+DatasetDM annot.)	**68.4%**	**+21.1%**

its relatively low spatial resolution—32×32, 16×16, and 8×8 elements. Another factor is the distribution gap between the LINZ and UGRC datasets, which may interfere with the annotation generation when DatasetDM is used, because it has been trained on LINZ data only.

Table 4: Detection performance when only manual labels are used.

Training Data	Test AP_{50} ↑	AP Gain ↑
Real LINZ *(baseline)*	47.3%	—
Real LINZ + Synthehtic UGRC (DAAM annot.)	71.7%	+24.4%
Real LINZ + Synthehtic UGRC (DatasetDM annot.)	77.8%	+30.5%
Real LINZ + Synthehtic UGRC (DAAM+DatasetDM annot.)	**79.2%**	**+31.9%**

5.2.2 Qualitative Analysis

Figure 5 illustrates four examples of test predictions by the YOLOv5 trained on Real LINZ and Synthetic UGRC datasets. The predicted object locations align closely with the *ground truth* (GT) annotations. These results demonstrate that the proposed approach effectively addresses the challenge of cross-regional small object detection by leveraging a generalized base model (YOLOv5) trained on the synthetic dataset with automatic annotations.

Figure 5: YOLOv5 Detection Visualization, where the orange X is the ground truth annotation, and the blue dot with the bounding box is the prediction result.

6. DISCUSSION

6.1 General Discussion

Overall, we find that the proposed framework furnishes encouraging results for the challenging problem of detecting small objects in satellite or overhead aerial imagery. The idea of synthetic image generation as a data augmentation technique is not new, but our approach effectively meets specific requirements. We found that the realm of satellite imagery posed new challenges that may or may not hold for other kinds of imagery. For example, the trends of region-specific distributions may also extend to other imagery types like medical imagery (subject-specific distributions). In these realms of distribution-specific imagery, the proposed approach may be a practical solution when dealing with image-sparse or annotation-sparse distributions of interest, corroborating its usefulness in other domains.

6.2 Synthetic Image Generation

From these experiments, we have found that some of the current foundational generation models (i.e., Stable Diffusion) have difficulties generating satellite imagery, which we believe is due to the lack of satellite imagery in their training corpus. As shown earlier, the off-the-shelf Stable Diffusion model provided subpar and out-of-distribution imagery before we fine-tuned it on our sparse target dataset.

While the SD model, before fine-tuning, may be able to output images that can be understood to be satellite imagery, it does not seem to have a sense of consistency concerning the sensor parameters and geographical location. For example, the model may create a set of images from the same prompt where all images may have varying GSD values (one image may cover 100 m of ground while another could cover 200 m of ground in the same image size) or different lighting/road conditions (one image may have asphalt roads while another may have concrete). This is an unacceptable inconsistency when creating a target dataset because these attributes of the images should be the same as they would be if they were all taken with the same sensor in the same region. We also found that the original SD model creates objects in synthetic images that are not of the proper shape, which is also unacceptable, especially when those objects are the target of the detection algorithm. For these reasons, we believe our pipeline is an improvement over the prior work.

6.3 Automatic Annotation Creation

Combining various techniques with individual strengths has merit. In our annotation module, we utilized existing methods from the literature and proposed a fusion technique using the best components from each method. For example, the CLIPSeg method effectively identified cars, while the DAAM and DAAMi2i methods were better at detecting the boundaries between close objects. By using a fusion of these to generate a new heatmap and an ensemble to verify accurate annotations, we developed a new approach that combined the advantages of each approach.

6.4 Limitations and Future Works

While the pipeline performs well in creating in-distribution synthetic data and corresponding labeling, we notice some limitations. There is a reasonably high image rejection rate when performing automatic image annotation using the DAAM-based method. Around 90% of the generated images are rejected; thus, while the quality of the accepted images is high, many others are of low quality. This throughput may burden compute time and resources, depending on the application in which this pipeline is being used. Another limitation is that this high rejection rate may lead to a loss in diversity of the output images, where the synthetic dataset may only comprise a portion of the actual data distribution. This could lead to imbalances between portions of the data distribution supported by the synthetic imagery and those not. Addressing this limitation is a promising direction for future research.

6.5 Conclusion

This work addresses the challenging problem of detecting small objects like cars in cross-domain satellite or aerial imagery without annotated data in the target location. We propose a framework that generates and automatically annotates images and maintains high quality by retaining only the best synthetic imagery. We demonstrate that the generation component of the pipeline can generate images that are within the distribution of the target dataset and are more realistic than those generated by other off-the-shelf generation methods. We also demonstrate that the annotation component of the pipeline is robust in generating accurate labels and outperforms each sub-component independently. Finally, we also demonstrate that the synthetic dataset created by our pipeline can help boost the performance of an off-the-shelf object detector model by significant margins.

7. ACKNOWLEDGMENT

The DEVCOM Army Research Lab has funded this work.

REFERENCES

[1] Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine* **5**(4), 8–36 (2017).

[2] Robinson, C., Hou, L., Malkin, K., Soobitsky, R., Czawlytko, J., Dilkina, B., and Jojic, N., "Large scale high-resolution land cover mapping with multi-resolution data," in [*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 12718–12727 (2019).

[3] Yang, X., Sun, H., Fu, K., Yang, J., Sun, X., Yan, M., and Guo, Z., "Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sensing* **10**(1) (2018).

[4] Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L., "Dota: A large-scale dataset for object detection in aerial images," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (June 2018).

[5] Labs, P., "Planet Labs." https://www.planet.com/ (2025). Accessed: 24th March, 2025.

[6] NASA, "Worldview." https://worldview.earthdata.nasa.gov/ (2025). Accessed: 24th March, 2025.

[7] Copernicus, "Sentinel." https://sentinels.copernicus.eu/ (2025). Accessed: 24th March, 2025.

[8] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B., "High-resolution image synthesis with latent diffusion models," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 10684–10695 (June 2022).

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I., "Attention is all you need," in [*Advances in Neural Information Processing Systems*], Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., eds., **30**, Curran Associates, Inc. (2017).

[10] Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., and Ture, F., "What the daam: Interpreting stable diffusion using cross attention," in [*The 61st Annual Meeting Of The Association For Computational Linguistics*], (2023).

[11] Chowdhury, R. D., "Daam-image2image: Extension of daam for image self-attention in diffusion models," (2022).

[12] Lüddecke, T. and Ecker, A., "Image segmentation using text and image prompts," in [*2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 7076–7086 (2022).

[13] Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M. Z., and Shen, C., "Datasetdm: Synthesizing data with perception annotations using diffusion models," in [*Advances in Neural Information Processing Systems*], Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., eds., **36**, 54683–54695, Curran Associates, Inc. (2023).

[14] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," in [*Advances in Neural Information Processing Systems*], Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., eds., **27**, Curran Associates, Inc. (2014).

[15] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S., "Deep unsupervised learning using nonequilibrium thermodynamics," in [*Proceedings of the 32nd International Conference on Machine Learning*], Bach, F. and Blei, D., eds., *Proceedings of Machine Learning Research* **37**, 2256–2265, PMLR, Lille, France (07–09 Jul 2015).

[16] Zareian, A., Rosa, K. D., Hu, D. H., and Chang, S.-F., "Open-vocabulary object detection using captions," in [*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 14388–14397 (2021).

[17] Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., and Angelova, A., "Open-vocabulary object detection upon frozen vision and language models," in [*The Eleventh International Conference on Learning Representations*], (2023).

[18] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., and Girshick, R., "Segment anything," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*], 4015–4026 (October 2023).

[19] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I., "Learning transferable visual models from natural language supervision," in [*Proceedings of the 38th International Conference on Machine Learning*], Meila, M. and Zhang, T., eds., *Proceedings of Machine Learning Research* **139**, 8748–8763, PMLR (18–24 Jul 2021).

[20] (LINZ), L. I. N. Z., "Selwyn 0.125m Urban Aerial Photos (2012-2013), Islington, NZ." https://data.linz.govt.nz/layer/51926-selwyn-0125m-urban-aerial-photos-2012-2013/ (2012). Accessed: 25th March, 2025.

[21] (UGRC), U. G. R. C., "High Resolution Orthophotography (HRO) - 2012." https://gis.utah.gov/products/sgid/aerial-photography/high-resolution-orthophotography/#2012-hro (2012). Accessed: 25th March, 2025.

[22] Li, J., Li, D., Savarese, S., and Hoi, S., "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in [*International conference on machine learning*], 19730–19742, PMLR (2023).

[23] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K., "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 22500–22510 (2023).

[24] Vincent, L. and Soille, P., "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(6), 583–598 (1991).

[25] Kornilov, A. S. and Safonov, I. V., "An overview of watershed algorithm implementations in open source libraries," *Journal of Imaging* **4**(10) (2018).

[26] Ultralytics, "ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation." https://github.com/ultralytics/yolov5 (2022). Accessed: 24th March, 2025.

[27] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S., "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in [*Advances in Neural Information Processing Systems*], Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., eds., **30**, Curran Associates, Inc. (2017).