Enhancing Object Detection through Generative AI-Based Data Augmentation Techniques

Omkar Jois[#1], Ananya Jha[*2], Bhoomi Bhat[#3] ,Kunjam Nanavaty[#4]

[#]*Computer Science Department, PES University*
Ring Road Campus, 100 feet road, BSK 3rd Stage, Bangalore 560085

Abstract— **Object detection is a crucial task in computer vision with applications ranging from autonomous driving to surveillance systems. However, the performance of object detection models heavily relies on the quality and diversity of the training data. Data augmentation techniques have been widely adopted to address the challenges posed by limited and imbalanced datasets. In this paper, we propose leveraging generative AI methods for data augmentation in object detection tasks. We explore various generative models and augmentation strategies to enhance the performance and robustness of object detection systems.**

Keywords— **Object detection, Generative AI, Data augmentation, Computer vision, Deep learning.**

I. Introduction

Generative AI techniques, particularly those involving adversarial and diffusion models, have shown promising advancements in enhancing object detection systems. Traditionally, object detection frameworks rely heavily on large, well-annotated datasets to train robust models. However, the cost, time, and effort required to compile these extensive datasets pose significant challenges, especially under constraints like limited data scenarios or rare object occurrences[1].

Recent developments have shifted focus towards innovative data augmentation methods that can synthetically expand dataset size and diversity without extensive manual annotation. Techniques such as Generative Adversarial Networks (GANs) and diffusion models have been pivotal. These methods not only augment data but do so in a way that enhances the model's exposure to varied, yet realistic, training samples[2].

This paper discusses the impact of generative AI-based data augmentation on object detection accuracy. It explores various generative techniques, focusing on their ability to produce high-quality, diverse images that maintain the integrity of object characteristics. This includes a critical examination of diffusion-based data augmentation where controllable aspects of image generation—like style, lighting, and bounding box precision—are manipulated to yield better model performance across standard datasets such as MSCOCO and PASCAL VOC. The subsequent sections will delve into specific methodologies, results from empirical evaluations, and comparisons with traditional data augmentation techniques.

II. Related Work

The landscape of object detection has been dynamically transformed by the advent of deep learning techniques, particularly through the integration of Generative Adversarial Networks (GANs) and diffusion models. These generative techniques are increasingly recognized for their potential to synthesize realistic training data, which is particularly beneficial for enhancing the performance of object detection models under constrained data conditions.

Lee, Kang, and Chung (2023) [1] emphasized the role of GANs in generating high-quality synthetic images that effectively mimic the statistical distribution of real datasets. This approach is particularly advantageous for domains where data collection is challenging, such as medical imaging or rare event detection. Their work demonstrates how GAN-generated images can be used to train object detection models, resulting in improved detection accuracy without the need for extensive real-world data collection.

In parallel, Fang et al. (2023) [2] introduced a novel approach using controllable diffusion models that adaptively manage the generation of synthetic data by manipulating visual priors such as textures and boundaries. This method allows for the synthetic generation of object images with precise bounding box annotations, which are crucial for training accurate detection models. The flexibility in controlling the synthetic image characteristics ensures that the augmented data align with the variations encountered in real-world scenarios, thus enhancing the model's generalization capabilities.

Hojun Lee, Minhee Kang, Jaein Song, Keeyeon Hwang (2022) [3] describe the process of collecting and preprocessing data for training a Pix2Pix model to generate black ice images. The data was collected from Google and YouTube, and video frames were extracted to obtain images. The images were cropped to 256(w)×256(h) px size, and data selection was performed to select images containing black ice. Data labeling was conducted to label black ice and roads in the images, and out-label data and label data were constructed. The out-label data was used as ground truth, and the label data was used for learning the characteristics of each object. Data combining was conducted to match the format of the data input to Pix2Pix, and it was divided into train data and test data through data split. The Pix2Pix model was designed with a generator and discriminator, and the generator was designed as an encoder-decoder structure with skip connections to minimize feature loss. The discriminator was designed to determine the authenticity of the generated image using the PatchGAN structure. The model was trained for 200 epochs.

Justin Bunkera , Georgios M. Hadjidemetrioua, Alix Marie d'Avigneaua, Mark Girolam (2023) [4] discuss that proposed methodology for pothole detection examines the impact of data transformations on algorithm performance. The dataset is divided into two subsets for search and evaluation, with multiple models used for testing, including Faster R-CNN, Mask R-CNN, and YOLOv8. Transformations like Affine operations, RandomContrast, RandomBrightness, MedianBlur, Solarize, and RandomShadow are compared to enhance model performance. The search procedure involves training and evaluating 100 configurations for each model and using fANOVA to determine the importance of each configuration choice. The results provide insights for improving pothole detection algorithms and can be applied to other infrastructure asset monitoring.

Disentangled representation learning is an unsupervised learning technique. Its goal is to find a disentangled representation that affects only one aspect of the data, while leaving others untouched.To find a disentangled representations, InfoGAN [5] was proposed, which is a variation of GAN that finds interpretable disentangled representations instead of unknown noise. InfoGAN allows the model to learn a disentangled representation by employing constraints during representation learning. Moreover, it divides the input into incompressible noise and latent code, and maximizes the mutual information between the latent code and generator distribution. That is, the latent code information is retained during the generation process.

The Image-to-Image (I2I) translation technique maps images of one domain to another. Although this task may seem similar to style transfer, they have a key difference. Style transfer aims to translate images such that they have the style of one target image while maintaining the contents of the image. In contrast, I2I translation aims to create a map between groups of images [6].

Pix2Pix [7] was the first supervised I2I conditional GAN-based model used for learning mappings between two paired image groups. However, because Pix2Pix has an objective function based on the L1 loss between translated and real images, unpaired datasets cannot be used for training. Unsupervised I2I translation models have been proposed to solve this problem. Cycle-consistent adversarial network (CycleGAN) [8] is one of the best-known unsupervised I2I translation models. It contains two pairs of generators and discriminators. Each generator and discriminator pair learns to map an image onto the opposite domain. Additionally, cycle-consistency loss have been proposed, which is defined using the L1 distance between the original image and that recovered from the image translated into another domain. Cycle-consistency loss can alleviate the problem caused by the absence of a paired dataset [9]. Contrastive learning for unpaired image-to-image translation (CUT) [10] is an unsupervised I2I translation model based on contrastive learning. Its goal is to ensure that a patch of translated image contains the content of the input image. CUT achieves this goal by maximizing mutual information through contrastive loss. Contrastive loss maximizes the similarity of patches for the same location in both the input and output images and minimizes patch similarity at different locations.

III. Methodology

A. Prelude

The research aims to address the limitations in object detection accuracy within vehicle datasets when trained on conventional models such as YOLOv8. Recognizing the challenges in acquiring diverse and sufficient training data, we propose leveraging generative AI-based data augmentation techniques to enrich our dataset. This approach centers on the use of image-to-image translation facilitated by the Stable Diffusion model, which generates multiple semantically similar yet visually diverse images from a single input. This methodology is designed to enhance the robustness and performance of the object detection model by introducing varied visual representations without altering the essential annotations.

B. Solution Pipeline

Our system incorporates a comprehensive pipeline designed to optimize object detection through generative AI-based data augmentation. Initially, a pre-trained YOLOv8 model processes the original vehicle dataset to establish a baseline for object detection performance. Following this, our data augmentation pipeline is initiated, wherein each image from the dataset is fed into the Stable Diffusion model along with a contextual prompt designed to generate three semantically similar yet visually distinct images. These images are created without altering the positions of key objects, thus allowing for the direct transfer of existing labels and bounding boxes to the newly generated images.

The augmented dataset, now enriched with these additional images, undergoes a re-training phase on the YOLOv8 model. This re-training is aimed at exposing the detection model to a broader range of visual scenarios, thereby enhancing its ability to generalize from the training data to real-world applications. The entire process is iterated with various prompts to maximize diversity and dataset robustness. Finally, the

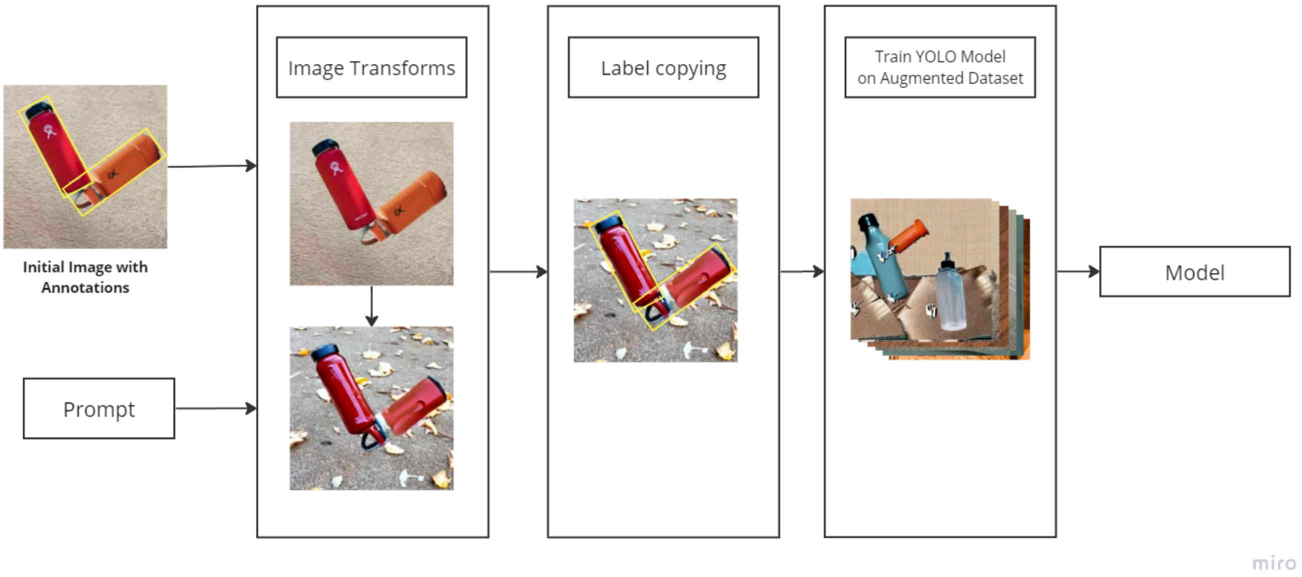
Initial Image with Annotations
Prompt
Image Transforms
Label copying
Train YOLO Model on Augmented Dataset
Model
miro

Fig 3.1 Diagram of the Solution pipeline

performance of the object detection model is evaluated using rigorous metrics such as accuracy, precision, recall, and F1-score, ensuring that the enhancements from the generative augmentation translate into tangible improvements in model performance. This pipeline not only leverages advanced AI techniques for dataset enhancement but also maintains a seamless flow that ensures data consistency and optimal learning conditions for the object detection model.

C. *Data collection*

We used 2 main datasets
Our objective is to compile a comprehensive dataset that accurately represents the variety of vehicles which the object detection system is expected to identify.

Source images from public datasets or through the deployment of cameras in various environments like urban streets, highways, and parking lots to ensure diversity in vehicle types and lighting conditions.

Annotate images with bounding boxes using CVAT, marking the precise location and category of each vehicle within the images.

D. *Generative Model Selection*

The goal is to select a generative model that can augment our vehicle dataset with high-quality images, enriching the diversity and aiding in the generalization capability of our object detection model.

1)*Fidelity and Diversity*: Stable Diffusion models are known for their ability to produce highly realistic images with a wide range of variations, which is crucial for representing the many possible appearances of vehicles under different conditions.

2)*Controllability*: The model accepts text prompts, allowing us to generate specific scenarios such as varying weather conditions or times of day that may not be sufficiently represented in the base dataset.

3)*Efficiency*: Compared to some GANs, Stable Diffusion can be more efficient in terms of training and generation time, which is beneficial when augmenting large datasets.

4)*Community and Support*: Stable Diffusion has an active community and ongoing support, providing a wealth of pre-trained models and resources that can accelerate our training and augmentation process.

E. *Algorithm and Mathematical Formulae*

The primary algorithm behind Stable Diffusion is a type of denoising autoencoder that iteratively refines images from a noise distribution. It combines concepts from diffusion models, which model the data distribution as a gradual diffusion process, with denoising techniques to generate clear, high-resolution images.
The core algorithm of the Stable Diffusion model is based on the concept of reverse diffusion, which can be mathematically represented by a Markov chain that starts with a distribution of pure noise and sequentially learns to remove noise to reconstruct the data. This can be described as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (1)$$

In this equation $\mu\theta$ and $\Sigma\theta$ are parameterized by the neural network with parameters with θ which are learned during training. xt represents the image at a particular time step in the reverse diffusion process, and *N* the normal distribution from which we sample to perform the reverse diffusion.

F. Stable Diffusion Model

1) Objective: To adapt the Stable Diffusion model to our vehicle dataset to generate realistic and diverse synthetic images that will augment our training data for object detection.

2) Training Protocol:

- *Model Initialization:* Initialize the Stable Diffusion model with weights from a version pre-trained on a comprehensive image dataset. Ensure the model is configured to generate high-resolution images that match the quality and dimensions of the base vehicle dataset.
- *Training Setup:* Prepare the vehicle dataset, selecting a subset of images that showcase the variety of vehicles and scenarios present in the dataset. Formulate text prompts that correspond to each image, designed to instruct the model to produce realistic variations of the original images.
- *Fine-Tuning Process:* Fine-tune the Stable Diffusion model on the vehicle dataset for a total of 10 epochs. This duration is chosen to allow the model sufficient time to learn the specific features and variations of vehicles without overfitting to the training data. Utilize a learning rate that decreases progressively across epochs, which helps the model to converge smoothly to a stable set of weights. Implement checkpoints at the end of each epoch to save model weights, allowing for recovery and continuation of training if necessary.
- *Monitoring and Validation:* Employ both qualitative and quantitative assessments after each epoch to monitor the model's performance and the quality of the generated images. Use a validation subset of the vehicle dataset to evaluate how well the generated images align with the distribution and characteristics of real-world data.
- *Training Execution:* Execute the training process on a suitable hardware setup, our system uses Nvidia GeForce RTX 3060 GPU, to manage the computational load efficiently. Employ data loaders and batching techniques to optimize memory usage and ensure stable training performance.
- *Expected Training Dynamics:* As the epochs progress, the generated images should increasingly reflect the diversity and complexity of the vehicle dataset. By the end of the 10 epochs, the Stable Diffusion model is expected to produce synthetic images that, when labeled and used in conjunction with the original data, enhance the object detection model's accuracy, providing improved performance on previously challenging scenarios

G. Training Custom YOLOv8 Model

IV. Results and Discussions

After training the YOLO models with the Augmented Dataset and with the normal Dataset, we found the following map50 and map 50-95 accuracies for the models.

Class	**map 50**	**map 50-95**
Ambulance	0.785	0.667
Bus	0.62	0.435
Car	0.405	0.282
Motorcycle	0.645	0.393
Truck	0.281	0.211

Table 4.1 Accuracies for the model without Augmentation

Class	**map 50**	**map 50-95**
Ambulance	0.823	0.674
Bus	0.722	0.562
Car	0.479	0.327
Motorcycle	0.497	0.287
Truck	0.351	0.254

Table 4.2 Accuracies for the model after Augmentation

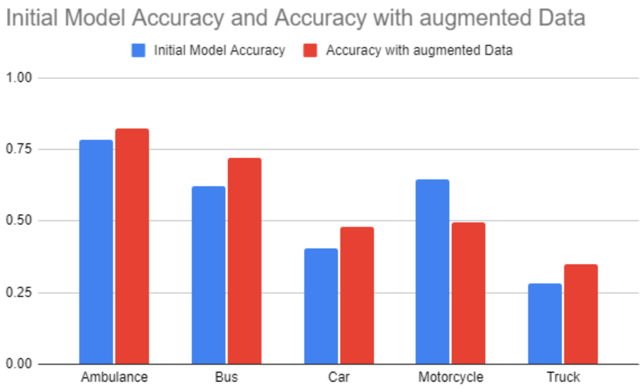

Fig 4.1 A chart displaying the comparison between the model accuracies before and after augmentation

Here we can see that in all but one case the accuracy of the model increases significantly.
In the case of the motorcycle the images contain very small motorcycles compared to the image size, which causes the motorcycle image generation to be faulty.

V. Conclusions

We investigated the potential of generative AI-based data augmentation techniques for enhancing object detection performance. We employed Stable Diffusion, a powerful diffusion model, to generate synthetic images that augmented our original dataset. YOLOv8, a state-of-the-art object detection model, was then trained on the augmented dataset. Our results demonstrate that incorporating Stable Diffusion-generated data augmentation significantly improved YOLOv8's object detection accuracy compared to training on the original dataset alone. This approach proved particularly effective in scenarios with limited training data.

VI. Future Work

Future work in this domain could focus on refining generative AI methods specifically tailored for object detection tasks, addressing the challenge of retaining very small objects in augmented images. This could involve developing novel techniques to ensure that generative models accurately preserve and enhance small objects during data augmentation processes. Moreover, investigating the impact of different generative models and augmentation strategies on the detection of small objects could provide valuable insights for optimizing the augmentation pipeline

Acknowledgment

I would like to express my sincere gratitude to Dr. Uma D and PES University for their invaluable guidance and support throughout this project.

I extend my thanks to Dr. Mamtha H.R. the Head Of Computer Science Department, for her leadership and for providing a supportive learning environment.

I would like to acknowledge the Dean of Faculty, the Vice-Chancellor, the CEO & Pro-chancellor, and the Chairman & Chancellor of PES University for their continuous support and for providing the resources necessary for my success.

References

[1] .Lee, H., Kang, S., & Chung, K. (2023). Robust Data Augmentation Generative Adversarial Network for Object Detection. Sensors, 23(157).

[2] Fang, H., Han, B., Zhang, S., Zhou, S., Hu, C., & Ye, W.-M. (2023). Data Augmentation for Object Detection via Controllable Diffusion Models. AWS AI.

[3] Lee, HoJun and Kang, Minhee and Song, Jaein and Hwang, Keeyeon, Pix2pix-Based Data Augmentation Method for Building an Image Dataset of Black Ice.

[4] Justin Bunker, Georgios M. Hadjidemetriou, Alix Marie d'Avigneau, Mark Girolami,On the performance of pothole detection algorithms enhanced via data augmentation,TransportationResearchProcedia, Volume78,2024

[5] Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.

[6] Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal Unsupervised Image-to-image Translation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018

[7] Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.

[8] Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.

[9] Pang, Y.; Lin, J.; Qin, T.; Chen, Z. Image-to-Image Translation: Methods and Applications. IEEE Trans. Multimed. 2022, 24, 3859–3881

[10] Park, T.; Efros, A.A.; Zhang, R.; Zhu, J.Y. Contrastive Learning for Unpaired Image-to-Image Translation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020