Semantic Scene Synthesis for Autonomous Vehicles

CSE 573 - Spring 2024

Team Autonomous

Pranav Polavarapu (50540640) Paige Shoemaker (50331415)

***Abstract*—In the realm of autonomous driving, the ability to train algorithms with a variety of realistic scenarios is crucial. This project , "Semantic Scene Synthesis for Autonomous Vehicles," leverages the power of Dual-Attention Generative Adversarial Networks (DAGAN) to generate synthetic urban scenes, enhancing the robustness and adaptability of autonomous driving systems. Initiated with a subset of the Cityscapes Dataset, the project successfully deployed a SPADE generator to synthesize scenes from semantic labels. Significant achievements include the refinement of image realism through advanced attention mechanisms and post-processing techniques, culminating in a retraining of the model with an expanded dataset of 300 images, significantly improving the diversity and quality of generated scenes. Furthermore, the development of a user-friendly interface using Streamlit has made it possible to visualize and interact with the synthetic images easily, facilitating rapid testing and iterations. This project not only advances the capabilities of scene synthesis but also sets a foundation for future enhancements that could lead to more sophisticated training environments for autonomous vehicles.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

A. Project Background and Motivation

The advancement of autonomous driving technologies necessitates robust training with realistic, varied scenarios to ensure safety and efficiency in urban environments. Traditional datasets, while extensive, often lack the variability needed to expose autonomous systems to the full spectrum of real-world conditions. This project addresses this limitation by synthesizing realistic urban scenes using advanced generative adversarial networks (GANs). The ability to generate diverse visual environments from semantic labels allows for scalable and controlled testing and training scenarios, which are crucial for improving the performance of autonomous driving algorithms under various conditions.

B. Objectives of the Project

The primary objective of this project is to enhance the capabilities of autonomous vehicle systems by providing a tool for generating high-quality synthetic images of urban environments. Specifically, the project aims to:

- Develop a model that can generate diverse and realistic scenes based on semantic inputs using the SPADE generator.
- Retrain the model on an expanded dataset to improve the quality and diversity of the generated images.
- Create a user-friendly interface using Streamlit to allow users to easily interact with the model, providing an effective tool for researchers and developers to generate and visualize synthetic scenes in real-time.

II. Problem Statement

A. Description of the Problem

Current autonomous driving research relies heavily on limited and unrealistic datasets, hindering the development of robust models for urban environments.

B. Importance

Dataset limitations lead to poor generalization and safety concerns in real-world deployments, necessitating the creation of more diverse and scalable training data.

C. Challenges Addressed

- **Variability**: Real-world datasets lack diversity, leading to overfitting.
- **Scalability**: Acquiring large-scale datasets is time-consuming and costly.
- **Realism**: Synthetic datasets lack the realism required for effective training.

Addressing these challenges is crucial for advancing autonomous driving technologies for real-world applications.

III. Methodology

A. Overview of the Approach and Pipeline

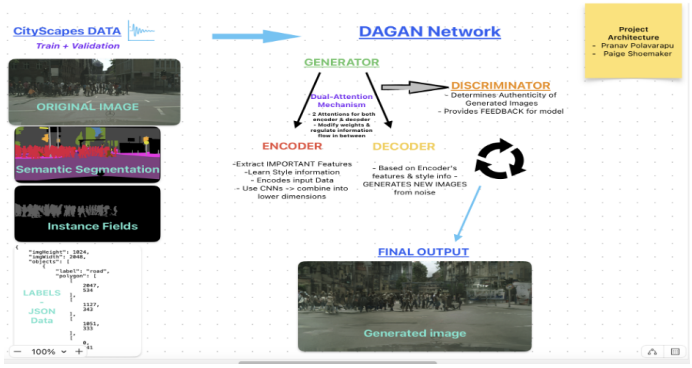
Our methodology encompasses a systematic approach to semantic scene synthesis, leveraging Dual Attention Generative Adversarial Networks (DAGAN) for generating realistic synthetic scenes. The pipeline involves processing input semantic maps through the DAGAN model, incorporating attention mechanisms, and applying post-processing techniques to enhance image realism.

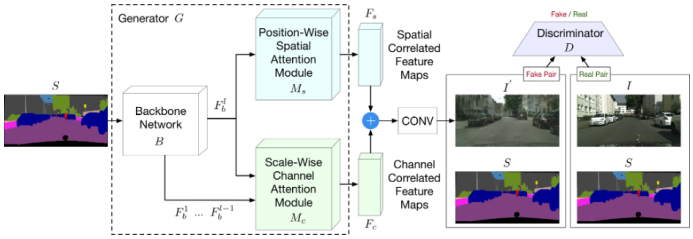
B. Initial Model Architecture

Overview of DAGAN Model:

DAGAN, based on the Generative Adversarial Network (GAN) framework, comprises a generator and a discriminator. The generator produces images while the discriminator evaluates them against real data. It incorporates attention mechanisms to refine details and maintain consistency by focusing on salient features. The model utilizes an encoder-decoder

architecture, employing convolutional layers to encode semantic information and deconvolutional layers to decode features into a coherent image. Through adversarial loss, it trains the generator to produce images indistinguishable from real ones, optimizing for photorealistic image generation. Additionally, DAGAN enables conditional image synthesis based on semantic input maps, facilitating the production of specific image types. This model is tailored to capture the complexity and variability of real-world imagery, making it ideal for tasks like training computer vision systems for autonomous vehicles.





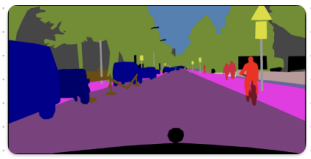
C. Architecture of our project

When the DAGAN model receives a semantic map as input, each layer in its architecture performs a specific role to ensure the final image is realistic. Here's how the layers work step by step for a given semantic map input:

1) Initial Understanding of the Scene:

- **Input Layer:** Takes a semantic map where each pixel color designates an object type, such as vehicles (blue), roads (gray), or trees (green).

Original Scene



SEMANTIC LABEL

2) Breaking Down the Map::

- **Encoder Layers:** Analyze the semantic map to detect features and outlines of the designated objects. Compress the spatial dimensions of the map while enhancing feature channels, which helps in capturing both the overall structure and the finer details.

3) Focusing on Uniformity and Precision:

- **Attention Layers: (DUAL-ATTENTION)**
  - **Spatial Attention**: Checks across the entire compressed feature set to maintain visual consistency. For example, it ensures that all pixels labeled as 'road' contribute to a continuous and consistent road surface in the image.
  - **Channel Attention**: Focuses within each feature channel to refine the details. It might enhance the texture on the 'road' channel to match the expected roughness or the reflectiveness of the 'vehicle' channel to give cars a metallic shine.

4) Integration and Refinement::

- **Bottleneck:** All the information processed so far converges here, where the model consolidates global features with local details, preparing for the reconstruction phase.

5) Constructing the Realistic Image:

- **Decoder Layers:** Start expanding the feature set back into a full-scale image, reintroducing spatial complexity while maintaining the richness of the features. Enhance the resolution progressively to develop a detailed image that retains the clarity of the identified objects and textures.

6) Final Image Synthesis: :

- **Output Layer:** The last layer compiles all the refined information to produce the final image. It ensures that the objects match their semantic labels, and the scene comes together cohesively, resembling a real-life urban environment as closely as possible.

D. Improvements and Modifications in Milestone 2

In this milestone, we improved our DAGAN model for semantic scene synthesis, focusing on tailored enhancements for autonomous vehicles. Here's a summary:

1) Subset Selection:

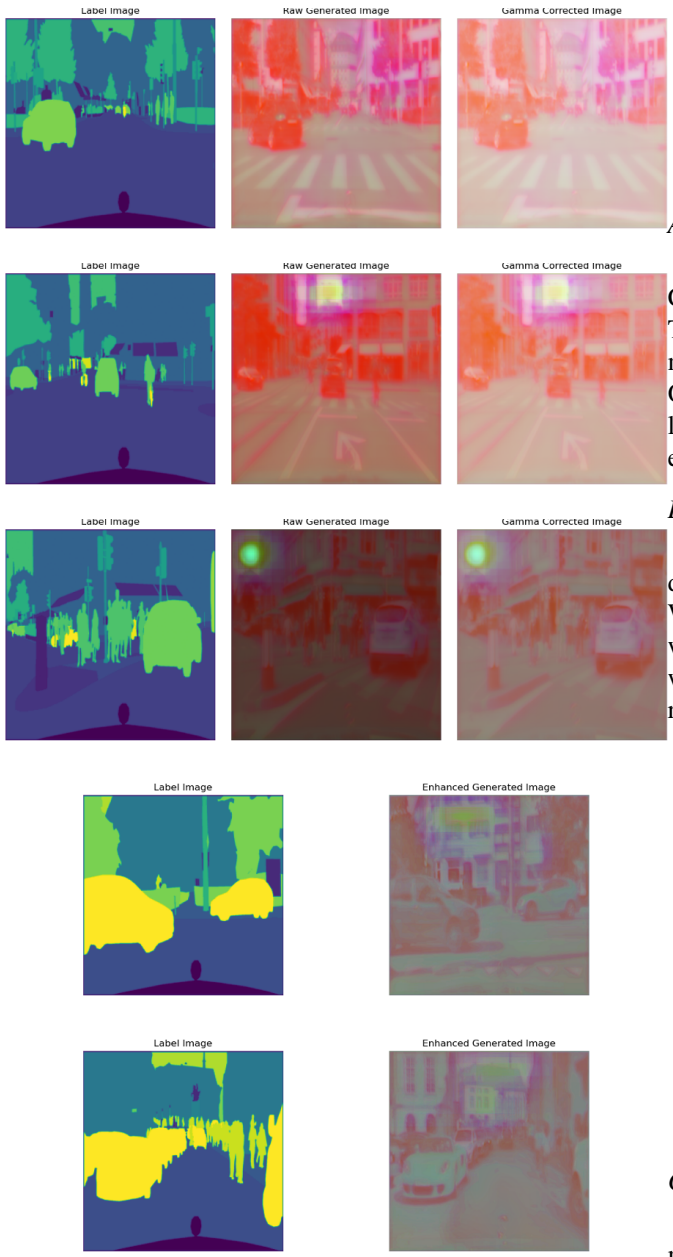
- **Curated Dataset**: We curated a set of 100 diverse images from the Zurich subset of the Cityscapes dataset, enriching the training data with varied urban features.
- **Challenge**: This diverse dataset challenges the DAGAN to maintain high synthesis quality across a broad range of complex urban scenes, enhancing model robustness.

2) Optimization of Training Parameters:

- Tuned key training parameters for optimal performance:
- **Learning Rate (0.001)**: Balanced convergence speed and accuracy, crucial for handling complex urban scenes.
- **Optimizer (Adam with beta1=0.5, beta2=0.999)**: Enhanced gradient responsiveness and stability during training.
- **Batch Size (5)**: Increased gradient noise for better generalization across urban environments while managing computational load efficiently.

3) Enhanced Image Corrections:

- **Gamma Correction**: Implemented gamma correction to adjust luminance to reflect real-world lighting conditions, vital for training vision systems in diverse lighting scenarios.
- **Sharpness Enhancement**: Improved edge and texture definition, enhancing the model's ability to generate crucial details for autonomous driving tasks.



E. Further Enhancements Made After Milestone 2

1) Increased Data Subset for Retraining:

- Expanded the training dataset from 100 to 300 images from the Zurich subset of Cityscapes. This larger dataset enhances model robustness and performance by providing more diverse urban scenes for training.

2) Modifications to Model's Architecture or Parameters:

- Implemented adjustments to the model parameters like decreasing the Learning Rate to 0.0005 to optimize performance.
- **Batch Size Modification**: Decreased batch size to optimize performance. These modifications aim to further enhance image synthesis quality and training efficiency.

3) Development of Streamlit UI:

- Created a user interface using Streamlit for easy visualization of generated images based on random input labels. This tool facilitates dynamic testing, iteration of model outputs, and user-friendly interaction with the model's results.

IV. Implementation

A. Dataset Description and Selection Criteria

For our Semantic Scene Synthesis project, we utilized the Cityscapes dataset, specifically focusing on the Zurich subset. The Zurich subset offers a diverse range of urban scenes, making it an ideal choice for training and testing our model. Our selection criteria prioritized images with varied semantic labels to provide a comprehensive test bed for evaluating the effectiveness of the DAGAN's attention mechanisms.

B. Technical Setup & Parameter Settings

Our implementation leveraged a combination of Python and deep learning frameworks, including TensorFlow and PyTorch. We employed a Dual Attention Generative Adversarial Network (DAGAN) architecture for image synthesis. The model was configured with specific parameters tailored to our project requirements:

- **Number of Generator Filters (ngf):** Set to 64 to enable detailed image reconstruction without excessive computational burden.
- **Latent Vector Dimension (z_dim):** Chosen as 256-dimensional to ensure sufficient variability in the generated images.
- **Semantic Number of Channels (semantic_nc):** Utilized 3 channels to correspond with RGB input, maintaining standard color space representation.
- **Number of Upsampling Layers (num_upsampling_layers):** Configured with the 'normal' setting to achieve a balanced upscaling, catering to both high-level detail and low-level precision.

C. Details of the Computational Environment

Our computational setup was designed to facilitate efficient model training and development:

- **Device:** Configured to run on a CPU to ensure compatibility with the hardware setup and accessibility for development and iteration.
- **Training Parameters:** Learning rate initialized at 0.001, suitable for steady convergence without missing global minima. Adam optimizer was employed for its adaptive learning rate capabilities, crucial for the varied data in

Cityscapes. We utilized a mix of adversarial loss, feature matching loss, and perceptual loss to encourage both global coherence and local detail fidelity.

- **Attention Modules:** Implemented Spatial Attention Module (SAM) to enhance spatial consistency across image regions sharing the same semantic labels, and Channel Attention Module (CAM) to strengthen the model's ability to perceive and emphasize relevant features across different scales.

Each aspect of our technical setup was carefully considered and selected to contribute to the overarching goal of synthesizing high-fidelity images for creating realistic training data for autonomous driving algorithms.

V. Results

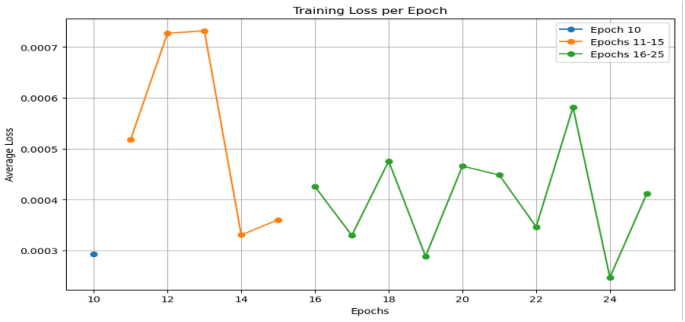
A. Milestone 1 Preliminary Results:

For the initial phase, the Cityscapes Dataset was selected, and a pre-trained DAGAN (Dual-Attention GAN) Model was implemented. This facilitated the creation of a model checkpoint and testing of the sample data against the latest checkpoint epoch. Below outlines the architecture and outcomes obtained.

B. Milestone 2 Results:

Output Comparisons and Technical Insights:

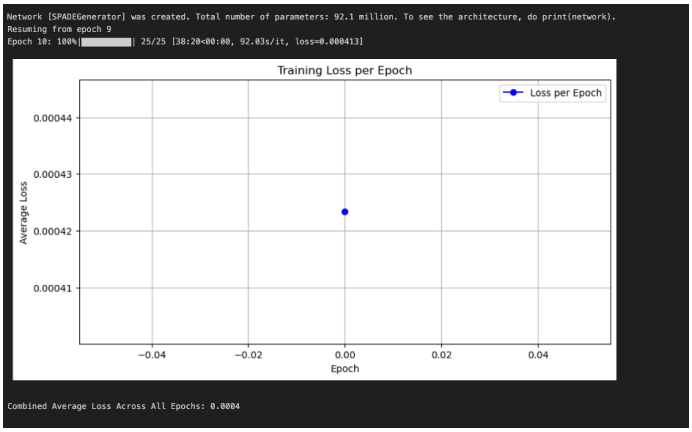
- **Gamma Correction:** Implemented to recalibrate the luminance of the images, resulting in a more realistic balance of light and shadow, thereby enhancing overall visual fidelity.
- **Sharpness Adjustment:** Employed to accentuate edges and textures within the images, leading to increased edge clarity, crucial for object detection systems.
- **CLAHE (Contrast Limited Adaptive Histogram Equalization):** Applied to enhance contrast in localized areas of the image, revealing richer textures and a depth of detail, particularly in buildings and road surfaces.
- **Saturation Enhancement:** Increased color intensity to create a more vibrant color palette, aiding in the differentiation of various elements within the scene.
- **Training Trajectory (Epoch vs. Loss) Analysis:** Initiating with a loss of 0.0003, the training trajectory encountered a peak loss of 0.0007 around epochs 11-15, stabilizing by the end of this phase. However, fluctuations were observed in later epochs, indicating potential overfitting concerns.
- **Optimal Loss and Epochs:** Targeting a loss near 0.0004, typically achieved between epochs 15-20, was identified as ideal. Post-20 epochs, there's a risk of overfitting, necessitating monitoring for increased loss and stabilization around the 0.0004 mark.



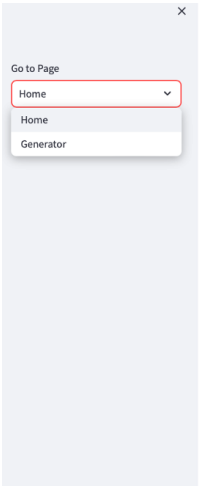
Enhanced the DAGAN model utilizing advanced attention mechanisms and post-processing techniques, resulting in lower loss values and increased consistency in image generation.

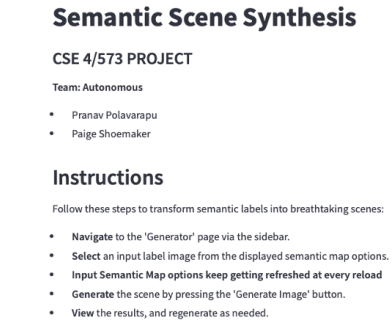
C. Milestone 3 Results:

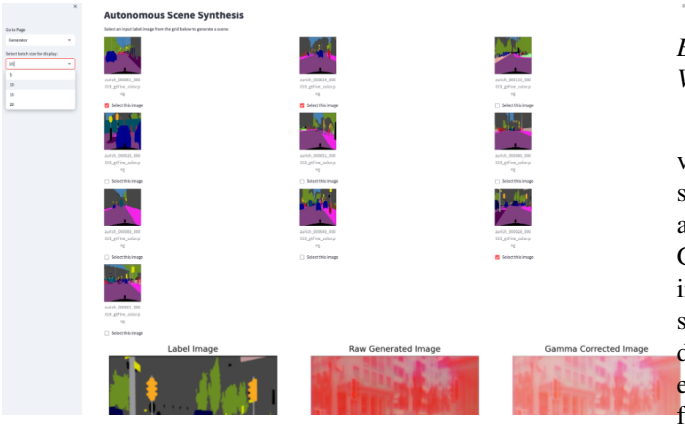
1) Model Retraining with Increased Subset:: Following the retraining phase with an expanded dataset of 300 images, significant progress was observed in the model's performance. The epoch-wise analysis revealed notable enhancements in image synthesis quality and consistency. At epoch 10, the model achieved a combined average loss of 0.0004, indicating improved convergence and fidelity in generating synthetic scenes. Further evaluation is underway to assess the model's capability to generalize across a broader range of urban environments and semantic elements.



2) Integration of Streamlit UI Application:: In parallel with model refinement, a user-friendly Streamlit UI application was developed to provide seamless visualization and interaction with the synthesized images. The interface offers intuitive controls for adjusting parameters and exploring variations in the generated scenes. Users can dynamically assess the impact of different input labels and post-processing techniques, facilitating rapid iteration and evaluation of model outputs.







VI. Discussion

A. Analysis:

Progression from Milestone 1 to 3 saw significant advancements. Milestone 1 established the project foundation, while Milestone 2 introduced key enhancements like gamma correction and sharpness adjustment. Milestone 3 showcased improved convergence and fidelity through model retraining with a larger dataset and integration of a user-friendly Streamlit UI app.

B. Comparison:

DAGAN's utilization demonstrates superior performance in generating realistic scenes compared to traditional methods. The attention mechanisms and post-processing techniques contribute to higher-quality image synthesis, crucial for autonomous vehicle perception systems.

C. Implications:

The project's findings are pivotal for advancing autonomous driving technology. The synthesized images aid in training robust perception systems, while the Streamlit UI facilitates efficient model evaluation and iteration.

VII. Conclusion

A. Summary of the Findings:

Throughout the project milestones, significant advancements were achieved in the development and enhancement of the Dual Attention Generative Adversarial Network (DAGAN) model for semantic scene synthesis in autonomous vehicles. Milestone 1 laid the groundwork by selecting the Cityscapes Dataset and implementing a pre-trained DAGAN model, setting the stage for subsequent improvements. Milestone 2 showcased enhanced image realism through the implementation of gamma correction, sharpness adjustment, and other post-processing techniques. Milestone 3 further refined the model through retraining with an expanded dataset and the integration of a user-friendly Streamlit UI application.

B. Contributions of the Project to the Field of Computer Vision:

The project's contributions extend to the field of computer vision by advancing the state-of-the-art in semantic scene synthesis for autonomous vehicles. By leveraging advanced attention mechanisms and post-processing techniques, the DAGAN model demonstrates superior performance in generating realistic synthetic scenes, crucial for training perception systems in autonomous driving applications. Additionally, the development of the Streamlit UI application enhances model evaluation and iteration, fostering a more dynamic and user-friendly development environment.

C. Limitations and Future Work:

Despite the achievements, the project encountered several limitations and challenges. Computational constraints, including CPU limitations and restricted training data subsets, hindered model performance and output quality. Image quality complications such as blurred outcomes and color deviations also posed challenges. However, prospective solutions such as data augmentation and resource upgrades, including transitioning to GPU or cloud-based resources, offer avenues for future improvement. Furthermore, future work will focus on refining the DAGAN architecture, exploring hyperparameter tuning, and enhancing the user interface to further improve model efficiency and output quality.

D. Project Summary:

In summary, the project journey encompassed the evolution of the DAGAN model from its initial implementation to its refinement and enhancement through successive milestones. The project's findings contribute to advancing autonomous driving technology by providing high-quality synthetic training data and user-friendly tools for model evaluation and development. While challenges were encountered, the project's achievements lay the groundwork for future advancements in semantic scene synthesis and autonomous vehicle perception systems.

VIII. References/Literature

1) **DAGAN:** Dual Attention Generative Adversarial Network.
   Paper: https://arxiv.org/abs/2008.13024
2) **CityScapes Dataset:** A Dataset for Semantic Urban Scene Understanding.
   Paper: https://arxiv.org/abs/1604.01685
3) **StyleGAN:** A Style-Based Generator Architecture for Generative Adversarial Networks.
   Paper: https://arxiv.org/abs/1812.04948
4) **CycleGAN:** Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.
   Paper: https://arxiv.org/abs/1703.10593

IX. Appendices

A. Project GitHub Repository

https://github.com/Pranav-Polavarapu/Semantic-Scene-Synthesis

X. Individual Contributions

A. Pranav Polavarapu

- Set the Technical ground, focusing on Generative AI and GANs for semantic scene synthesis.
- Implemented the Dual-Attention Generative Adversarial Network (DAGAN) model and associated methodologies.
- Conducted dataset analysis and preparation, including preprocessing and sampling.
- Collaborated in applying right foundational CVIP techniques learned through the course.
- Led model training pipelines, optimization strategies, and parameter tuning.
- Developed the user interface application using Streamlit and integrated it with the model for real-time visualization.

B. Paige Shoemaker

- Set the domain direction towards autonomous driving, emphasizing the importance of realistic training scenarios.
- Conceptualized the project's objectives and milestones, ensuring alignment with autonomous driving requirements.
- Researched and evaluated datasets, contributing to the selection of the Cityscapes dataset.
- Collaborated in applying right foundational CVIP techniques learned through the course.
- Analyzed model optimization strategies and monitored training progress.
- Collaborated on the design and features of the user interface application, ensuring usability and effectiveness.