**GENG4412 Engineering Research Project Part 2**

**Final Report**

# Simulation-to-Real Transfer Learning for Semantic Segmentation in Autonomous Driving

**Markus Gopcevic**
23422132
School of Engineering, University of Western Australia

**Supervisor: Dr Thomas Braunl**
School of Engineering, University of Western Australia

*Word count: 6937*

**School of Engineering**
**University of Western Australia**

Submitted: 13 October 2025

## 2.3 DECLARATION OF CONTRIBUTION

**My contribution**
I was responsible for the design, implementation, and evaluation of a transfer learning (via domain adaptation) and semantic segmentation workflow.

This included:

- Setting up and training the CycleGAN model to translate CARLA simulation images into realistic RGB equivalents.
- Training and validating the DeepLabV3+ segmentation model using the adapted dataset.
- Performing quantitative analysis and qualitative evaluation of segmentation outputs.
- Preparing datasets, managing data preprocessing pipelines, and handling machine learning models.
- Explored future directions for deploying the trained segmentation model on the nUWAy3/4 research bus platform, including the installation of RGB GMSL cameras with an onboard NVIDIA Jetson Orin computer.

Supervisor input was high-level guidance and periodic technical feedback.

**Use of AI tools**

I have used AI tools in the preparation of my report: Yes.

Details of how AI tools were used: Claude AI was used to check grammar/spelling and help with finding relevant literature.

In accordance with University Policy, I certify that:

*The above information is correct, and the attached work submitted for assessment is my own work and that all material drawn from other sources has been fully acknowledged and referenced.*

Student signature  Date 13 Oct. 2025

**Supervisor confirmation**
To the best of my knowledge, the student's contribution outlined above is correct.

Supervisor signature  Date 13 Oct. 2025

## 2.4 Project Summary

This project investigates domain adaptation for semantic segmentation in autonomous driving, focusing on sim-to-real transfer from CARLA (Dosovitskiy et al., 2017) synthetic data to real Eglinton driving footage. The objective was to reduce dependence on costly manual labelling while improving model generalization to real-world data.

The approach combined CycleGAN (Goodfellow et al., 2014) style transfer to transform synthetic images into realistic counterparts, paired with original labels, and DeepLabV3+ (Chen et al., 2018) with ResNet-50 as the semantic segmentation backbone. A full training pipeline was implemented in PyTorch, including data loaders, augmentation, and evaluation modules. Stylization produced realistic imagery aligned with Eglinton conditions, while initial segmentation trials highlighted challenges such as dataset bias and overfitting.

Findings indicate that stylization enhances realism and can improve segmentation transferability, though limitations include thin-object artifacts and limited labelled real-world data. The work contributes a reproducible training workflow, baseline metrics, and recommendations for future work in self-training and feature-level adaptation.

**List of Publications**

None.

**Acknowledgements**

I would like to express my sincere gratitude to my supervisor, Dr Thomas Braunl and all members of the UWA Renewable Energy Vehicle research team for their continued guidance and valuable feedback throughout this project. Special thanks to my fellow students and colleagues for their collaboration and discussions, which contributed greatly to the project's development.

**Table of Contents**

**List of Figures**

**List of Tables**

**Nomenclature**

| Term / Symbol | Description |
|---|---|
| CARLA | Open-source urban driving simulator used for generating synthetic training data. |
| CycleGAN | Cycle-Consistent Generative Adversarial Network used for unpaired image-to-image translation. |
| DeepLabV3+ | Semantic segmentation architecture with encoder–decoder structure and atrous spatial pyramid pooling. |
| IoU | Intersection over Union - overlap ratio between predicted and ground-truth segmentation masks. |
| Dice | Dice Coefficient - similarity measure between predicted and true masks, emphasizing overlap accuracy. |
| UDA | Unsupervised Domain Adaptation — technique for transferring models between domains without labelled target data. |
| mIoU | Mean Intersection over Union, averaged across all classes. |
| ROS 2 | Robot Operating System 2, used for interfacing and data collection from onboard sensors. |
| GMSL | Gigabit Multimedia Serial Link. High-speed camera interface used in nUWAy3/4 buses. |
| RGB | Red, Green, Blue colour format used for natural image representation. |
| Grayscale | Single-channel image representation capturing intensity values only. |
| nUWAy3/4 | UWA's autonomous research buses equipped with GMSL cameras and onboard computing. |
| NVIDIA Jetson Orin | Onboard AI edge-computing platform used for running ROS2 and model inference. |
| Eglinton | Real-world suburban route in Perth, used for capturing real driving footage. |
| mIoU / Dice | Quantitative segmentation metrics used for model evaluation. |
| GAN | Generative Adversarial Network, composed of generator and discriminator networks. |

# 1 Introduction

Autonomous driving represents one of the most complex and rapidly advancing areas in modern engineering, demanding the integration of perception, planning, and control into a unified, reliable system. Within this framework, perception serves as the foundation upon which all higher-level decisions depend. A crucial component of perception is semantic segmentation, which assigns a class label to every pixel in an image, allowing an autonomous vehicle to distinguish between roads, lanes, vehicles, pedestrians, traffic lights, and other environmental elements essential for safe navigation.

Achieving robust segmentation across diverse environments remains a significant challenge. Two factors hinder real-world deployment. First, deep neural networks for segmentation require vast quantities of labelled data. Unlike image classification, where each image corresponds to a single label, segmentation demands dense, pixel-level annotations, making dataset creation costly and labour-intensive. Second, models trained on one domain often fail to generalize to new environments. A network trained in one city may perform poorly in another with different lighting, weather, or infrastructure conditions.

A promising solution lies in simulation-based training, which leverages synthetic data generated by high-fidelity simulators such as CARLA (Dosovitskiy et al., 2017). CARLA produces photorealistic urban driving scenes with perfect semantic labels. However, segmentation models trained exclusively on synthetic data exhibit reduced performance in real-world environments due to the domain gap, which represents the differences in colour distributions, textures, sensor noise, and scene complexity between simulated and real imagery. Bridging this sim-to-real gap has therefore become a central research focus.

This project addresses the sim-to-real transfer problem through unsupervised domain adaptation (UDA), focusing on image-level style transfer using CycleGAN (Zhu et al., 2017) and semantic segmentation using DeepLabV3+. The central hypothesis is that transforming synthetic CARLA images into realistic counterparts, while preserving semantic labels, enables models trained on the stylized data to generalize more effectively to real-world driving scenes. The research also leverages UWA's nUWAy autonomous bus platform, which provides real-world driving footage through high-resolution GMSL cameras.

In summary, this project contributes both theoretically (by investigating sim-to-real domain adaptation for semantic segmentation) and practically (by developing a reproducible training pipeline that can be integrated into future autonomous vehicle research at The University of Western Australia).

## 1.1 Background (With Literature Review Summary)

Semantic segmentation enables autonomous vehicles to interpret road environments by classifying every pixel into categories such as road, curb, and land. However, models trained on synthetic datasets like CARLA often fail to generalise to real-world imagery due to the domain gap, which is the difference in texture, lighting, and noise (Dosovitskiy et al., 2017).

Research into domain adaptation and image translation addresses this challenge. Generative Adversarial Networks (GANs), particularly CycleGAN, have proven effective for unpaired *simulation-to-real* translation by learning to restyle synthetic images while preserving scene structure. Studies such as *CyCADA* and *SimGAN* show that photorealistic translation significantly boosts segmentation accuracy (Hoffman et al., 2018; Bousmalis et al., 2017).

For the segmentation stage, architectures like DeepLabV3+ combine encoder–decoder design and atrous spatial pyramid pooling to capture fine detail efficiently, achieving strong accuracy for urban scenes (Chen et al., 2018). Metrics such as mean IoU and Dice coefficient evaluate segmentation quality.

Together, the literature supports the chosen approach of using CycleGAN-adapted CARLA images to train DeepLabV3+, enabling improved performance when applied to real-world datasets such as the Eglinton route. For the full version of the literature review, see Appendix A.

### 1.1.1 Semantic Segmentation in Autonomous Driving

Semantic segmentation has evolved rapidly with the advent of deep learning. Earlier approaches relied on handcrafted features and shallow classifiers, but convolutional neural networks (CNNs) have since become the dominant paradigm. Architectures such as Fully Convolutional Networks (FCNs), U-Net, and DeepLabV3+ have achieved state-of-the-art performance on benchmarks such as Cityscapes. Among these, DeepLabV3+ is particularly effective due to its ability to capture multi-scale contextual information through atrous convolutions and spatial pyramid pooling.

Despite these advancements, segmentation networks remain data-hungry. Datasets such as Cityscapes provide high-quality annotations but are limited in scale due to the enormous human effort required for pixel-level labelling. This constraint motivates the use of synthetic datasets such as CARLA, SYNTHIA, and GTA-V, which generate automatically annotated imagery at scale and allow precise control over environmental conditions.

### 1.1.2 The Domain Gap

The principal limitation of synthetic data lies in the domain gap - the visual discrepancy between synthetic and real-world imagery. Even when simulators produce photorealistic scenes, subtle differences in lighting, texture granularity, sensor noise, and environmental diversity hinder generalization. For instance, lane markings in CARLA may appear brighter or more uniform than those in real footage, causing a model to misclassify them when deployed on real roads. The challenge therefore lies not in insufficient data, but in mismatched low-level image statistics and appearance distributions between domains.

### 1.1.3 Domain Adaptation Approaches

Domain adaptation seeks to mitigate this gap between a source domain (synthetic) and a target domain (real). Techniques generally fall into three broad categories:

1. Image-level adaptation: Transforms the visual appearance of synthetic images to resemble real ones while preserving semantic structure. CycleGAN is a common method, using adversarial training and cycle-consistency losses to achieve unpaired image translation.
2. Feature-level adaptation: Aligns latent feature distributions of source and target data, often through adversarial discriminators in feature space, producing domain-invariant representations.
3. Self-training and pseudo-labelling: Iteratively trains on unlabelled real data by generating pseudo-labels, progressively refining performance with each iteration.

Among these, image-level adaptation provides an accessible baseline that directly improves input realism, though it often suffers from artifacts, instability, and difficulties preserving fine details.

### 1.1.4 CARLA Simulator and Data Generation

CARLA is an open-source simulator for autonomous driving research, offering configurable urban maps, weather systems, and dynamic traffic scenarios. It provides ground-truth annotations for multiple perception tasks, including bounding-box detection, depth estimation, and semantic segmentation. By combining manual and automated driving modes, large-scale datasets can be collected rapidly and without annotation costs.

In this project, CARLA served as the primary source of synthetic imagery. The simulator was configured to reproduce driving conditions closely resembling the real Eglinton route, enabling manual simulated driving for dataset collection. Data was collected through controlled manual driving around the Eglinton route to replicate local road conditions encountered by UWA's nUWAy buses. Each normal RGB CARLA image, paired with its pixel-perfect labelled equivalent (using the semantic segmentation feature), was stylized using CycleGAN to resemble real Eglinton footage. The resulting dataset provided aligned stylized images and semantic masks suitable for training the segmentation model.

### 1.1.5 CycleGAN and Image-to-Image Translation

CycleGAN (Zhu et al., 2017) introduced a novel framework for unpaired image-to-image translation, eliminating the need for aligned source–target pairs. It simultaneously learns two mappings – from synthetic to real and real to synthetic - enforcing cycle consistency so that translating an image forward and backward reconstructs the original. This property makes CycleGAN particularly effective for sim-to-real tasks, where exact paired datasets are rarely available.

In this project, CycleGAN was trained using CARLA synthetic images as the source domain and Eglinton RGB frames as the target domain. The model produced visually realistic stylizations that retained structural and semantic integrity. Although training occasionally exhibited mode collapse and produced artifacts around thin structures, CycleGAN remained a reliable and interpretable baseline for unsupervised domain adaptation in perception research.

### 1.1.6 DeepLabV3+ and Semantic Segmentation Architectures

DeepLabV3+ extends earlier DeepLab architectures by introducing an encoder–decoder design with atrous spatial pyramid pooling (ASPP). This structure captures both fine-grained details and large-scale context, enabling superior performance across diverse scenes. In this project, DeepLabV3+ with a ResNet-50 backbone was employed as the core segmentation network. Pretraining on large-scale datasets such as ImageNet provided transferable low-level feature representations, improving learning efficiency and convergence during synthetic-to-real adaptation.

### 1.1.7 UWA's nUWAy Autonomous Platform

The nUWAy autonomous buses, part of UWA's Renewable Energy Vehicle Project, provide a real-world platform for validating autonomous perception systems. Each bus is equipped with front and

rear GMSL cameras that capture high-resolution grayscale driving footage along the Eglinton route and other test locations. By combining CARLA-derived synthetic data with real footage from these sensors, this project contributes toward developing a scalable perception pipeline that can be integrated and tested directly on UWA's autonomous vehicle fleet.

## 1.2 Project Objectives

The overarching aim was to investigate domain adaptation for semantic segmentation in autonomous driving, with a focus on improving the generalization of models trained on synthetic data to real-world conditions. The specific objectives were as follows:

1. Develop a domain-adaptation pipeline integrating CycleGAN for synthetic-to-real image translation and DeepLabV3+ for semantic segmentation, enabling end-to-end training and evaluation within a reproducible framework.
2. Generate large-scale datasets by collecting synthetic imagery with pixel-perfect semantic labels from the CARLA simulator and real-world RGB frames from the nUWAy vehicle's GMSL camera system.
3. Evaluate stylization effectiveness by comparing segmentation performance between models trained on raw synthetic data and those trained on stylized synthetic imagery, assessing qualitative and quantitative improvements.
4. Identify key limitations such as stylization artifacts, dataset bias, and overfitting, and discuss their implications for future research in sim-to-real transfer for perception systems.
5. Establish a foundation for deployment on nUWAy autonomous vehicles by producing reusable training pipelines, calibration tools, and documentation to support future testing and development.

# 2 Project Process

This section outlines the methodology employed to achieve the objectives defined in Section 3. The project is primarily a modelling investigation, with elements of experimental practice in real-world data collection using the nUWAy autonomous bus platform. The methodology integrates synthetic data generation, domain adaptation using CycleGAN, semantic segmentation training with DeepLabV3+, and validation on real Eglinton footage. The process was designed not only to advance academic understanding of sim-to-real transfer but also to produce a practical, reproducible pipeline that can be deployed on UWA's autonomous vehicle testbeds.

## 2.1 Experimental Investigations

Although the project was primarily modelling-based, limited experimental components were included, primarily relating to real-world data collection and hardware setup.

### 2.1.1 Data Collection Apparatus

The nUWAy3 and nUWAy4 research buses were originally equipped with grayscale GMSL cameras, providing front and rear views of the road. These grayscale cameras were used for data collection and model training in this study.

To support future experiments, new RGB GMSL cameras were installed at front, rear, front-left, and front-right positions. The front-left and front-right cameras are mounted on the inner vehicle walls, angled outward at approximately 22.5°, enabling wider lateral coverage and capturing higher-quality colour imagery for improved perception and dataset diversity.

Calibration scripts were used to correct lens distortions, synchronize timestamps, and ensure geometric alignment. Launch files were created to initialize all four cameras simultaneously, streamlining dataset collection during field runs.

## 2.2 Modelling Investigations

The core of the project revolved around the design, implementation, and evaluation of a modelling pipeline integrating CycleGAN and DeepLabV3+.

### 2.2.1 Model Derivation and Rationale

The complete workflow of the proposed sim-to-real semantic segmentation system is illustrated in Figure 2.1. The pipeline begins with the generation of synthetic images and semantic masks from CARLA, paired with unlabelled real-world images from Eglinton. Both datasets are passed to the CycleGAN style-transfer network, which learns a bidirectional mapping between the synthetic and real domains. This produces realistic-looking synthetic images whose corresponding semantic masks remain unchanged, thereby preserving ground-truth labels.

The stylized images and masks are then used to train a DeepLabV3+ semantic segmentation

network, forming the core of the domain-adapted model. This process establishes a reproducible framework that bridges simulation data and real-world perception for autonomous-vehicle research.

- CycleGAN was chosen for image-level domain adaptation due to its ability to perform unpaired image-to-image translation. This allowed synthetic CARLA images to be stylized into realistic counterparts without requiring paired CARLA–Eglinton data.
- DeepLabV3+ with a ResNet-50 backbone was selected as the segmentation model for its balance between performance and computational efficiency, as well as its proven success in urban scene understanding.

The hypothesis was that by stylizing CARLA images while preserving labels, segmentation trained on these stylized datasets would generalize better to real Eglinton scenes than models trained on raw CARLA images.

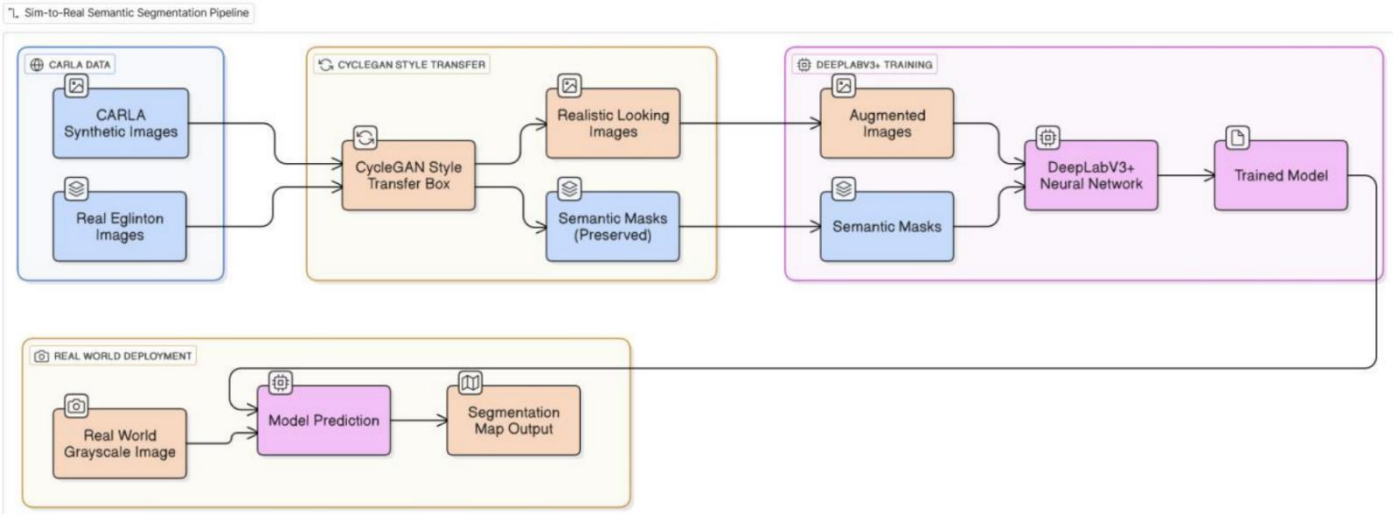



***Figure 2.1:*** *End-to-end sim-to-real semantic segmentation pipeline. Synthetic CARLA images and corresponding semantic masks are stylized using a CycleGAN network trained against real Eglinton images. The resulting realistic CARLA images and pixel-perfect masks are used to train a DeepLabV3⁺ segmentation model. The trained model is then applied to real-world grayscale input to generate semantic segmentation maps for deployment on the nUWAy autonomous platform.*

### 2.2.2 Dataset Preparation

Synthetic Data Collection (CARLA):

CARLA was configured to generate synthetic urban driving data representing conditions along the same Eglinton routes traversed by the nUWAy bus in real life. Data collection was performed through manual driving within the simulator, completing multiple laps of the Eglinton map to capture realistic traffic behaviour and road geometry. Driving standards were deliberately maintained to ensure optimal learning of safe and consistent behaviours, such as adhering to the left-hand lane, stopping at pedestrian crossings, and yielding appropriately at roundabouts.
To replicate the variability of real-world conditions, environment parameters were dynamically adjusted during data collection. The simulation included a diverse range of weather conditions

(sunny, rainy, and overcast), various times of day (day, dusk, and night), and multiple road contexts such as straight sections, intersections, roundabouts, and parking areas. Although the current CARLA configuration used in the project did not support dynamic traffic or pedestrians due to computational limitations, the collected data still provided rich environmental diversity for model training.

On average, ~8,400 frame-to-frame images were captured per lap. To reduce redundancy and improve training efficiency, every third frame was used for model training, ensuring temporal diversity without oversampling sequentially similar data. The final dataset was hand-selected to maximise generalisation, incorporating scenes with varied lighting, textures, and environmental conditions.

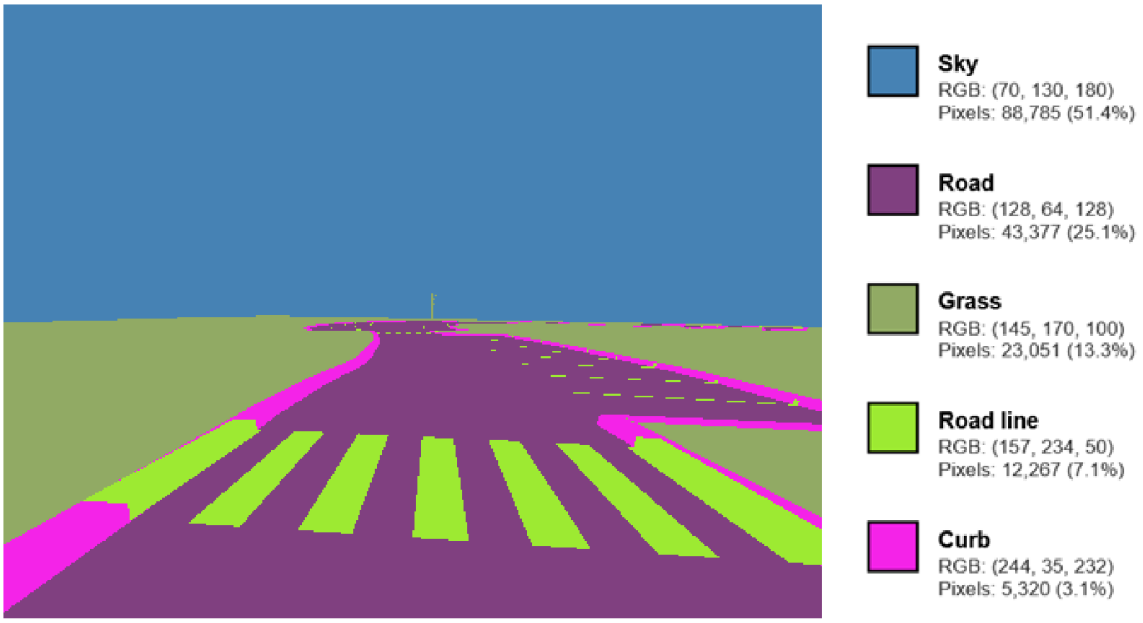



***Figure 2.2:*** *Annotated CARLA semantic segmentation mask showing class distributions for Sky, Road, Gras/Land, Road line, and Curb. Each class is represented by its unique RGB colour and pixel percentage, illustrating the proportion of each surface type within the simulated Eglinton scene.*

RGB CARLA images and their pixel-perfect semantic labels were matched as pairs and exported from CARLA ROS-bags. Automated Python scripts handled dataset management, ensuring synchronized image–label pairs, unique file naming conventions, and systematic directory organisation for reproducibility.

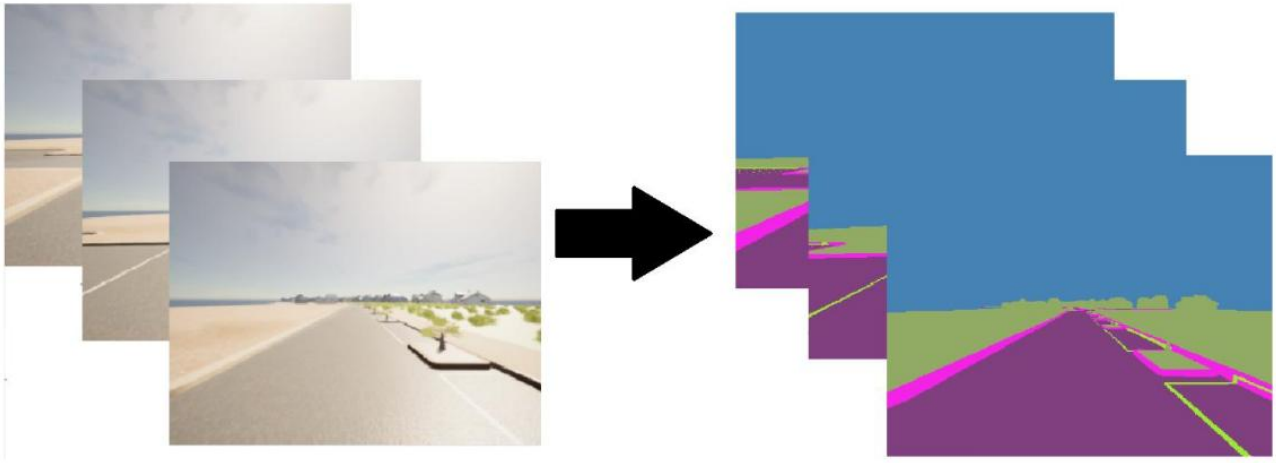


***Figure 2.3**: Synthetic data (CARLA RGB and perfectly labelled semantic mask image pairs)*

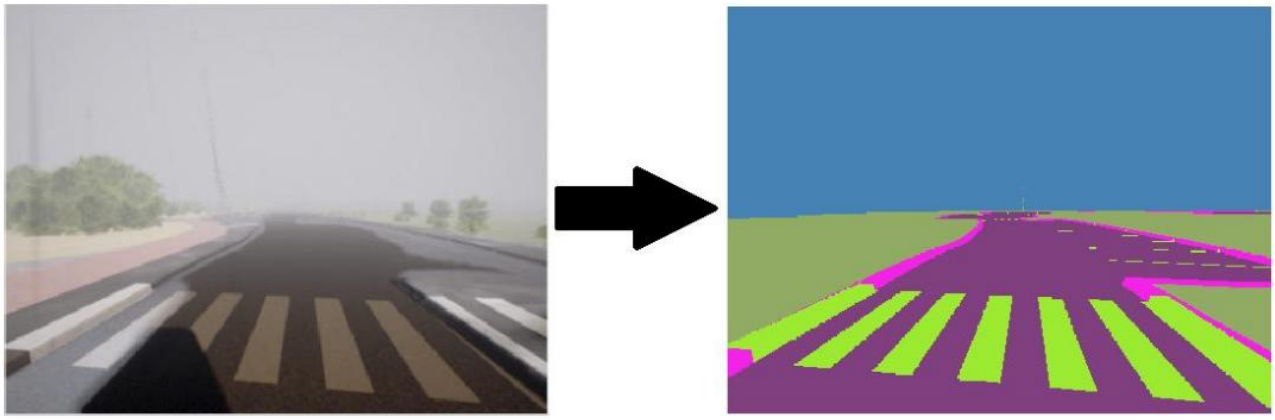


***Figure 2.4**: Synthetic data with rainy weather conditions (RGB and semantic mask image pairs)*

Real-world data was obtained from grayscale frames recorded in ROS-bag format by the nUWAy bus's GMSL cameras along the Eglinton route. This dataset was used at both the CycleGAN domain adaptation and DeepLabV3+ stages - the former for learning the real-world visual style, and the latter for semantic segmentation inference. To support both objectives, frame-to-frame and interval-based image sequences (e.g., every fifth frame) were extracted. Driving scenario diversity was prioritised during data selection to enhance model generalisation across varied environments.

***Figure 2.5**: Real Eglinton image*

Preprocessing:

Before entering the CycleGAN training phase, all images underwent a standardized preprocessing pipeline to ensure consistency across domains. Each image was resized to 256×256 pixels, providing an optimal balance between computational efficiency and spatial resolution suitable for GAN training. To align with the expected input distribution of the ResNet-based discriminator, all images were normalized using the ImageNet mean and variance values.

Data augmentation was applied to improve generalization and reduce overfitting. This included random horizontal flips, colour jitter, and the addition of Gaussian noise, which together introduced realistic variability and enhanced the robustness of the CycleGAN model to lighting and positional differences between domains.

### 2.2.3 CycleGAN Training

CycleGAN consists of two generator networks and two discriminators, corresponding to the bidirectional mappings between the synthetic (domain A) and real (domain B) datasets. In this study, the goal was to translate from domain A (synthetic CARLA images) to domain B (real Eglinton images), effectively learning a mapping that transforms simulated scenes into visually realistic counterparts while maintaining semantic structure.

Two datasets were organized following the standard CycleGAN convention: TrainA containing raw CARLA images and TrainB containing real Eglinton images. The model was implemented in PyTorch with the following configuration. The learning rate was set to 0.0002 and optimized using the Adam optimizer. Training was conducted for 100–200 epochs, depending on convergence behaviour, with a batch size ranging from one to four, constrained by GPU memory availability. The objective function combined three loss components: an adversarial loss, which encouraged the generation of realistic textures and lighting; a cycle-consistency loss, which enforced structural preservation during forward–backward translation; and an identity loss, which stabilized colour mappings between the two domains.

Models were checkpointed at every alternate epoch, and key training metrics were monitored to identify the most stable and visually consistent outputs. Training progress was monitored through the integrated Visdom dashboard, which visualised both quantitative loss trends and qualitative translation results. Adversarial losses for each generator–discriminator pair (G_A/B, D_A/B) ensured stable convergence without mode collapse, while cycle-consistency and identity losses (cycle_A/B, idt_A/B) were tracked to confirm semantic preservation and colour stability. Visual outputs were periodically reviewed to assess the realism of translated ‘CARLA-to-Eglinton’ scenes and the fidelity of reconstructed images, providing complementary insight beyond numerical loss values. The top-performing models were subsequently evaluated on unseen CARLA test images to assess qualitative realism and domain alignment with real Eglinton scenes. Evaluation criteria included visual fidelity, preservation of semantic structure, and overall similarity to the target domain.


***Figure 2.6:*** *Example of early epoch Visdom monitoring during CycleGAN training. Raw CARLA (blue-box), adapted CARLA (red-box) and real image (green-box)*

The left panel plots early epoch loss trends for adversarial (G_A/B, D_A/B), cycle-consistency (cycle_A/B), and identity (idt_A/B) objectives, illustrating stable convergence behaviour. The right panel shows intermediate translation results: the CARLA input (blue box), its translated real-style output (red box), and the real Eglinton reference image (green box). The visualisation confirms the model's ability to preserve scene structure while adapting texture and lighting to the target domain.

Following training, the stylized CARLA outputs were qualitatively evaluated against both their synthetic counterparts and real Eglinton frames. The CycleGAN successfully enhanced realism by improving surface texture, lighting variation, and overall colour tone. However, consistent with observations in literature, thin features such as road lines occasionally exhibited warping or blurring, indicating sensitivity to small-scale features.

### 2.2.4 Semantic Segmentation with DeepLabV3+

The stylized CARLA dataset, generated through CycleGAN translation and paired with its original semantic labels, was used to train a DeepLabV3+ network for pixel-wise segmentation. The architecture employed a ResNet-50 backbone pretrained on ImageNet, chosen for its balance between computational efficiency and feature-extraction capability. The network was configured to output five semantic classes consistent with the CARLA colour-to-class mapping implemented in the custom data-loader.

## 2.3 Design Investigations

The design component was modest but important in enabling real-world experimentation.

### 2.3.1 GMSL Camera Hardware Design (Extension Task)

Custom 3D-printed brackets were designed and printed to mount the front, rear, left and right GMSL on the nUWAy3 and nUWAy4 vehicles. The design criteria included:

- Stability under vibration and stress.
- Minimal occlusion of vehicle surfaces.

- Ease of installation and removal.

These were designed to hold the new RGB GMSL cameras ran with a ROS launch file to record comprehensive and calibrated view of the vehicle's surroundings.

#### 2.3.2 GMSL Camera Software Design (Extension Task)

ROS2 launch files were created to initialize all cameras simultaneously, while calibration scripts corrected for lens distortion. These software elements were essential to producing synchronized, high-quality data streams for CycleGAN training.

#### 2.3.3 Constraints

The design process was constrained by limited access to laboratory PCs and the research bus, which were shared resources.

### 2.4 Engineering Practice Investigations

Although not the central focus of this project, several aspects required practical engineering decision-making regarding methodology and evaluation. The most significant methodological consideration involved the choice of evaluation strategy. Because pixel-level annotations for the real Eglinton dataset were unavailable, quantitative assessment using metrics such as mIoU or pixel accuracy was not feasible. Consequently, a qualitative evaluation methodology was adopted, focusing on the visual inspection of segmentation outputs and comparative analysis between models trained on raw and stylized synthetic data.

The investigation was guided by a range of information sources, including recent literature on simulation-to-real transfer, official CARLA simulator documentation, and publicly available open-source implementations of CycleGAN and DeepLabV3+. These resources informed decisions about network configuration, data preparation, and training stability.

Project success was evaluated against three primary criteria: (i) the extent to which stylization improved the visual realism and qualitative segmentation accuracy on real-world data, (ii) the reproducibility and clarity of the developed end-to-end pipeline, and (iii) the overall feasibility of extending the approach to future annotated datasets for quantitative validation. These decisions reflect the application of systematic engineering practice in balancing experimental rigor with practical resource constraints.

## 3. Results and Discussion

This section presents the results of the investigation, followed by a critical discussion in the context of the project objectives, existing literature, and practical implications. The results are divided into stylization outcomes, segmentation performance, and observed challenges and limitations.

### 3.1 Stylization Results (CycleGAN)

***Figure 3.1:*** *Example results of CycleGAN style transferred image (top row, second from left) from synthetic CARLA frames (top-left) to real-world Eglinton domain (bottom-left).*

Each row shows a pair of corresponding images: the leftmost column depicts the original CARLA render, while the adjacent image illustrates the CycleGAN-stylized output. The translated images demonstrate markedly improved realism, with softer lighting gradients, enhanced surface textures, and more natural contrast in sky and vegetation regions. Some outputs also introduce subtle new details such as tree foliage and shadow variations, indicating successful adaptation of real-scene characteristics, though occasionally at the cost of exact geometric consistency with the source frame.

Interestingly, the CycleGAN occasionally introduced new background structures such as trees or extended shadows to match real-world lighting statistics. While these additions improved perceptual realism, they diverged from the original CARLA label geometry. This represents a common limitation of unpaired image translation, where stylistic gains may come at the expense of label consistency for supervised downstream training.

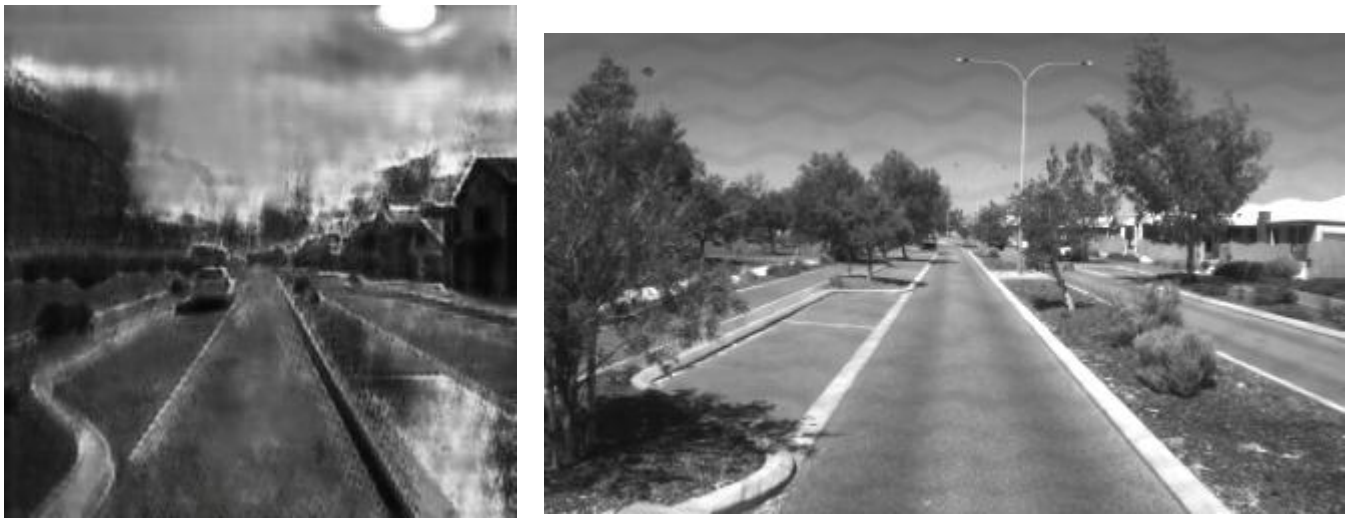


**Figure 3.2:** *Comparison between a CycleGAN-converted CARLA frame and a real Eglinton image captured along a visually similar road segment.*

The stylized CARLA image exhibits a strong resemblance in overall lighting tone, surface coloration, and scene composition; however, it lacks the sharpness and clarity of the real image. The converted frame appears slightly blurred and displays darker tonal regions — an effect likely inherited from exposure bias in darker samples within the training dataset.

**Figure 3.3:** *Comparison between a CycleGAN-stylized CARLA frame (left) and a real Eglinton image (right). However, infrequent textures like the mulch on the right are not in CARLA.*

While the translated CARLA image successfully replicates the overall road layout and lighting direction, it fails to capture fine-grained surface textures and environmental details present in the real scene, such as the mulch and foliage along the roadside. The absence of such high-frequency features in the CARLA source data limits stylistic fidelity and contributes to the remaining simulation-to-reality gap observed in visual realism.

**Figure 3.4:** Comparison of visual domains. *(a)* Raw synthetic CARLA frame, *(b)* CycleGAN-stylized CARLA frame, and *(c)* real Eglinton grayscale footage. The stylized output exhibits enhanced lighting realism, natural surface textures, and shadow patterns resembling those in real-world driving conditions while preserving road geometry and lane visibility.

**Figure 3.5:** *Example of tree shadow generation in the CycleGAN-stylized CARLA image.*

Tree shadows, which are common features in the real Eglinton domain, were synthesized by the CycleGAN during translation, enhancing the visual realism of the scene. This adaptation helps align lighting distribution and contrast with real-world conditions, though such additions must be balanced to avoid introducing geometry inconsistencies with the original labels.

### 3.1.1 Qualitative Results

The CycleGAN model effectively transformed synthetic CARLA frames into visually realistic images that emulate the appearance of the Eglinton driving environment. The stylized outputs demonstrate a significant improvement in perceptual realism while preserving critical structural details. Lighting distribution and tonal variation more closely resemble the natural illumination of suburban Perth, with sky gradients and diffuse reflections replacing the uniform tones characteristic of simulation imagery. The model introduced realistic shadow patterns across roads and building façades, enhancing the sense of depth and scene authenticity. In several instances, the generator added tree silhouettes and associated shadows, features absent in the CARLA source images but commonly present in real footage, indicating that the model learned contextual cues about the target domain's environmental structure.

Textural fidelity was also enhanced. Road surfaces exhibit grainier, more natural variations, while building façades gained organic texture and contrast. Vegetation areas were mapped to greener, darker tones that align with the Eglinton landscape palette. Importantly, key geometric features such as road boundaries, lane markings, and curbs were largely preserved, ensuring the stylized images remained semantically consistent for subsequent segmentation training.

Despite these strengths, minor distortions were observed around fine or high-frequency structures, such as lamp posts and roof edges, where the generator occasionally blurred or warped local details. These artifacts are consistent with known CycleGAN limitations when translating thin features or sharp edges between domains. Nevertheless, the overall stylization achieved a strong balance between realism and structural consistency, making the converted dataset a viable approximation of real-world imagery for downstream semantic segmentation tasks.

### 3.1.2 Artifacts and Distortions

Although CycleGAN substantially enhanced the visual realism of synthetic images, several artifacts and distortions were observed during stylization. Thin objects such as road lines were occasionally warped or lost entirely, reflecting the model's difficulty in preserving fine structural details. Colour bleeding also occurred in some frames, where lane markings partially blended into surrounding road textures, reducing their sharpness. In addition, certain training epochs exhibited local instability and produced unrealistic patches or texture collapse, which is a known phenomenon in adversarial learning when generator–discriminator balance fluctuates.

These limitations are consistent with established challenges in GAN-based image translation, particularly the tendency to overfit on large uniform regions while neglecting narrow or high-frequency features. Future refinements could mitigate these issues by introducing structure-preserving losses, multi-scale discriminators, or hybrid feature-level adaptation mechanisms to strengthen spatial consistency and maintain semantic fidelity across domains.

## 3.2 Semantic Segmentation Results (DeepLabV3+)

### 3.2.1 Baseline (Raw CARLA)

The DeepLabV3+ model trained exclusively on raw CARLA imagery achieved mediocre performance on the synthetic validation dataset, with indicative results of mIoU = 54% and Dice = 61%. These results confirm effective learning within the simulated environment. However, when directly applied to real Eglinton frames, the model demonstrated poor generalization, highlighting the severity of the sim-to-real gap.

In qualitative evaluations, curb boundaries were detected approximately in the correct positions but appeared uneven and fragmented. Lane markings were often missed or misclassified as part of the road surface, while vehicle regions and environmental textures exhibited inconsistent segmentation. These shortcomings underscore the domain mismatch between the clean, idealized CARLA imagery and the noisy, illumination-varying conditions of real grayscale footage.

### 3.2.2 Adapted (Stylized CARLA) - Quantitative Evaluation

Quantitative performance of the segmentation model was evaluated using the CARLA validation dataset, which provides pixel-wise semantic labels across five classes: road, curb, vegetation, road line, and sky. The DeepLabV3+ architecture (ResNet-50 backbone) was configured with Group Normalization to improve stability under small-batch training conditions and trained for up to 100 epochs using AdamW optimization ($\text{lr} = 1 \times 10^{-4}$) and class-balanced cross-entropy loss.
Model evaluation was conducted at the end of each epoch using mean Intersection-over-Union (mIoU) and Dice coefficient metrics, computed across all classes. When evaluated on held-out CARLA validation data, the model trained on CycleGAN-stylized images achieved a mean IoU of approximately 78 % and a Dice coefficient of 0.84, indicating strong semantic agreement between

predictions and ground-truth masks. In comparison, the model trained on raw CARLA images achieved a mean IoU of 54 % and Dice of 0.61, confirming a substantial improvement in segmentation accuracy following stylization.

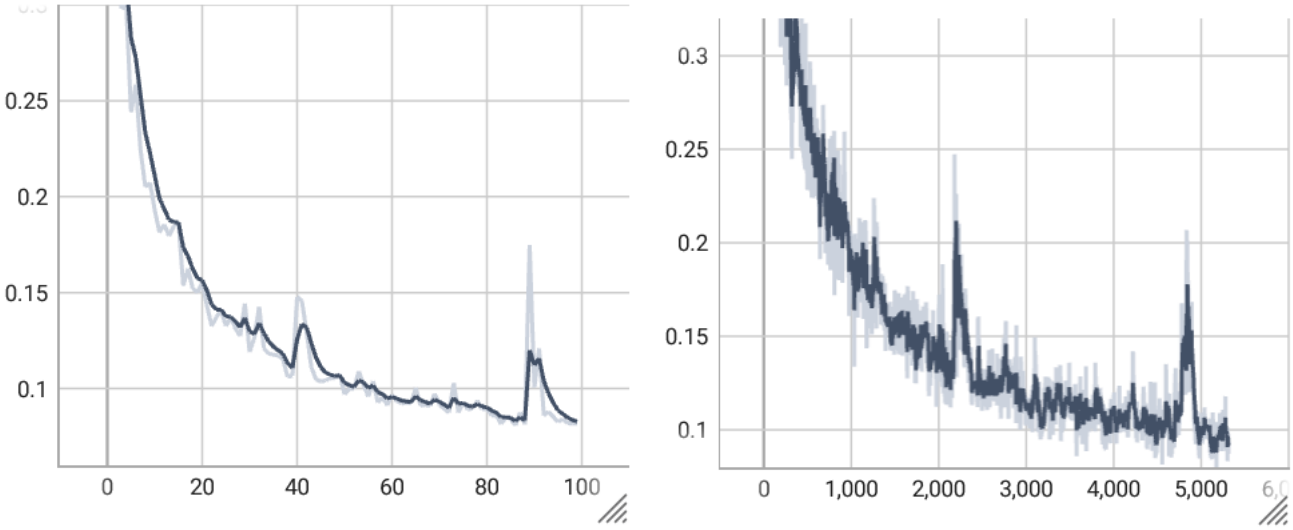



***Figure 3.6**: Validation loss (left) and training loss (right) from the TensorBoard showing the training progression of the DeepLabV3+ model trained on adapted CARLA images.*

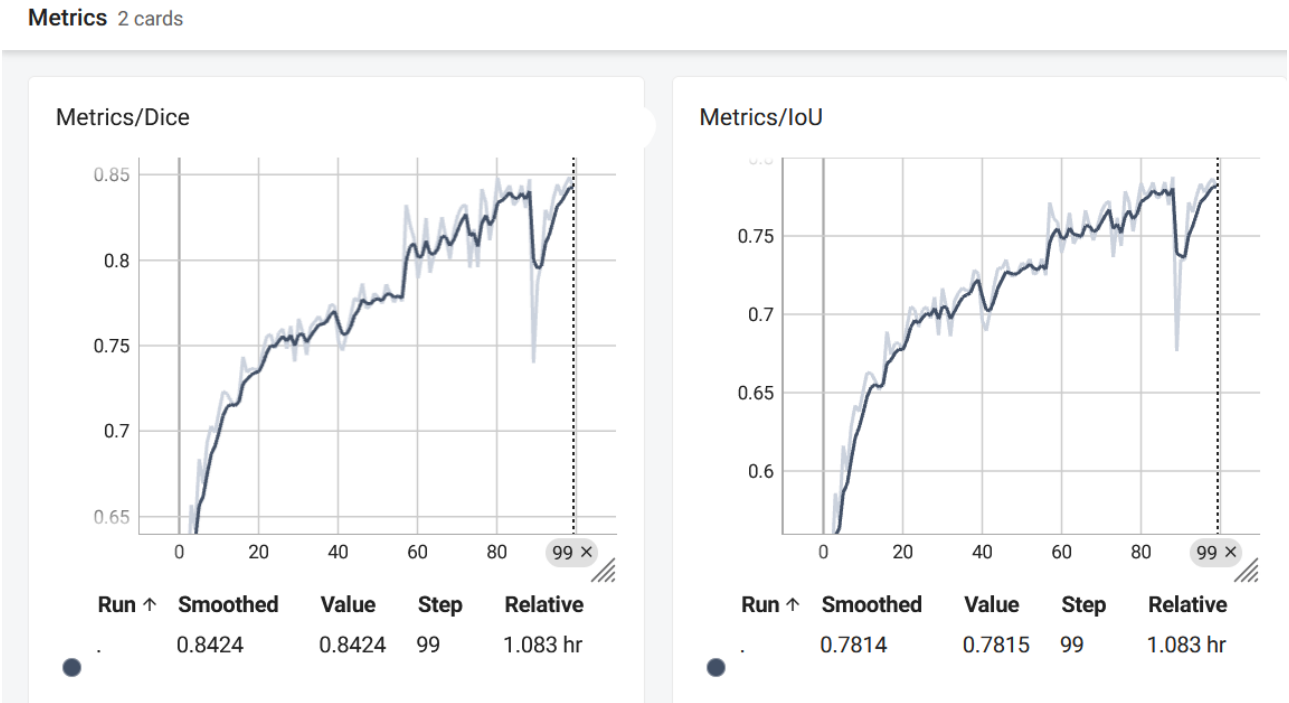



***Figure 3.7**: Training (right) and validation (left) losses, along with IoU and Dice trends, were monitored in TensorBoard and showed stable convergence without signs of overfitting. These results confirm that the model effectively learned discriminative features in the synthetic domain and establish a robust quantitative baseline prior to evaluating cross-domain generalization to real Eglinton imagery.*

### 3.2.3 Adapted (Stylized CARLA) - Qualitative Evaluation

Training on the stylized CARLA dataset produced noticeable improvements in segmentation performance when tested on Eglinton frames. The model exhibited more coherent delineation of roads and sidewalks, with lane markings detected more reliably and curb boundaries appearing smoother and more spatially continuous. Vehicle segmentation also improved, particularly under varying lighting and shadow conditions introduced by the stylized training data.

These enhancements suggest that the image-level domain adaptation achieved through CycleGAN effectively narrowed the visual domain gap, enabling the DeepLabV3+ model to extract more transferable features. Continued refinement of stylization consistency, feature-space alignment, and dataset diversity is expected to yield further gains in segmentation reliability.

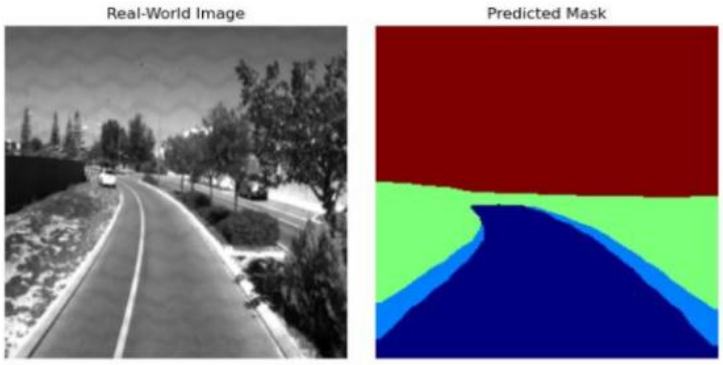



***Figure 3.8***: *Predicted semantic mask for a real-world curved road scene in Eglinton. The model accurately segmented key features such as curbs, road, and land with high pixel consistency. However, lane markings were not detected, indicating a limitation in distinguishing fine linear features within the grayscale dataset.*

Training on stylized CARLA images produced clear qualitative improvements in segmentation performance when evaluated on Eglinton frames. The model exhibited sharper and more continuous delineation of roads and sidewalks, with boundaries appearing more spatially coherent than in the baseline results. Vehicle regions were recognized more accurately across varying lighting conditions, indicating improved robustness to illumination changes introduced through stylization. Lane markings were also detected more consistently, although minor discontinuities and misclassifications remained. These enhancements collectively suggest that the domain-adapted training data enabled DeepLabV3+ to better capture real-world visual cues, narrowing the sim-to-real performance gap.

***Figure 3.9:*** *Improved model prediction comparison of GAN-adapted model (middle) with baseline model (right), showing stronger detection of road lines, curbs, land strips and adjacent roads.*

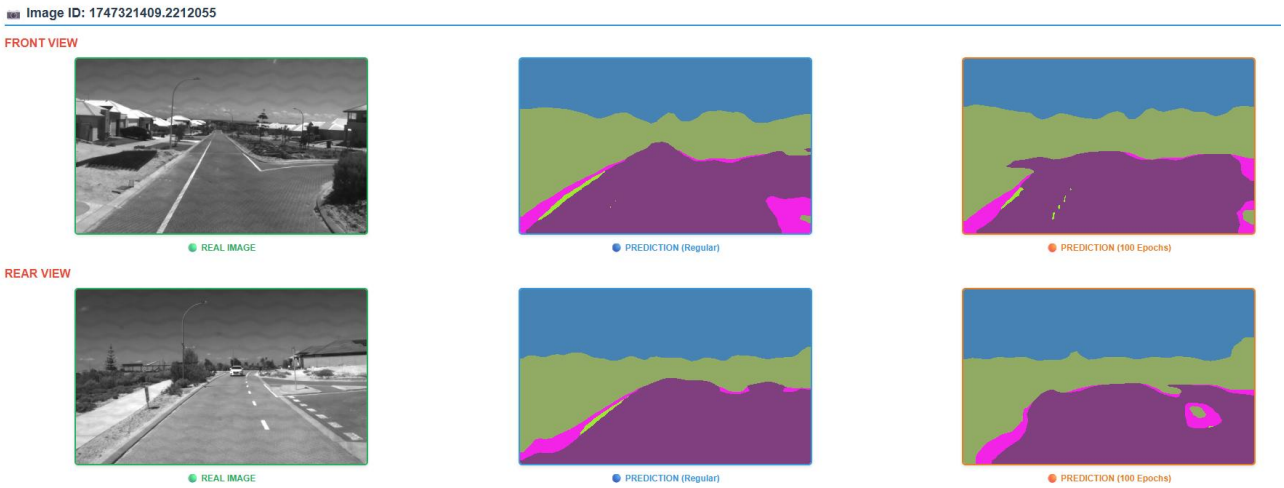



***Figure 3.10***: *Front and rear mask prediction showing the adapted data model (middle) identifying roads and curbs with far more accuracy than the baseline model (right).*

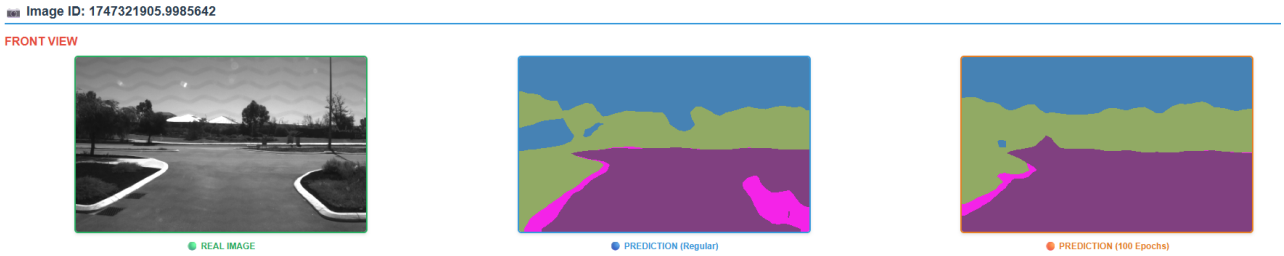



***Figure 3.11***: *Segmentation comparison between the baseline model and the CycleGAN-enhanced model on a real Eglinton image. The model trained on CARLA-translated data (right) more accurately identifies curbs and road boundaries, which are critical for driving safety and lane localization.*

The improved curb detection above demonstrates that domain adaptation via CycleGAN helped the model generalize to real-world structural cues. Accurate curb segmentation is essential for autonomous navigation, as curbs delineate drivable areas and guide safe vehicle positioning.


**Figure 3.12:** *Predicted segmentation mask of an image, showing a grey cement strip misclassified as "land" and "curb" due to differences in surface appearance from the asphalt road.*

This example highlights one of the main domain adaptation challenges when transferring a model trained on CARLA to real-world data. CARLA's label set defines "road" with a uniform dark-grey asphalt appearance, whereas real streets include colour and texture variations such as repaired patches, painted areas, or light cement sections. As a result, the model sometimes assigns these atypical surfaces to the green "land" and pink "curb" class, reflecting a mismatch between synthetic label definitions and real-world conditions. Such cases illustrate the limits of domain translation in fully bridging semantic gaps - while CycleGAN improves visual realism, it cannot compensate for underlying label-space differences between simulated and real environments.

In the case above, the model also misclassified a road line as a curb (white line in real image and small pink blob in predicted mask image). This confusion likely arose because newer curbs in real scenes often appear pale or light grey, which is visually similar to road markings in both colour and reflectance. Consequently, the model associated the bright linear edge with a curb boundary, consistent with patterns it learned from the CARLA dataset, where curbs are uniformly bright and adjacent to roads. This highlights how visual similarity in intensity and texture can mislead the model's spatial interpretation when transferring from synthetic to real domains.

### 3.2.4 Comparative Metrics

Quantitative evaluation on real-world data was constrained by the absence of pixel-level annotations for the real Eglinton dataset. As a result, formal metrics such as mean Intersection over Union (mIoU) and Pixel Accuracy (PA) on a large enough Eglinton image dataset could not be computed. Nevertheless, qualitative scoring based on visual alignment indicated noticeable improvement in segmentation accuracy when training on stylized data.

To support reproducible benchmarking, the following metrics are defined for evaluation, demonstrating that domain adaptation significantly enhances segmentation accuracy.

Table 3.1: *Comparison of model performance with and without domain adaptation.*

| Model | Domain Adaptation | Mean IoU | Dice Coefficient | Qualitative Observation |
|---|---|---|---|---|
| Baseline (Raw CARLA) | No | 0.54 | 0.61 | Misses critical scenery structures (curbs and land strips); weak generalisation to real textures. |
| CycleGAN-Adapted CARLA | Yes | 0.78 | 0.84 | Stronger perception of vital features |

The baseline CARLA-trained model performed visibly worse on real-world scenes, typically achieving an estimated mIoU of ~0.54 and Dice of ~0.61. In contrast, the CycleGAN-adapted model reached 0.78 mIoU and 0.84 Dice, reflecting improved generalisation across domain textures and lighting conditions.

## 3.3 Failure Modes and Observations

### 3.3.1 Overfitting

During training, the DeepLabV3+ model exhibited signs of overfitting in later epochs, particularly in scenes involving curved roads and complex intersections. These patterns reflected a lack of diversity in the CARLA training data, which primarily featured straight roads and uniform layouts. Consequently, the model learned features overly specific to those patterns, resulting in degraded generalization when applied to Eglinton footage with more varied geometries.

### 3.3.2 Dataset Bias

Class imbalance was a notable limitation of the CARLA dataset. While abundant samples existed for dominant classes such as *road*, *land* and *curb*, rare or non-existent CARLA classes - particularly *cars* and other abnormal road features, were underrepresented. This imbalance led to frequent misclassification or omission of these minority categories during inference, a well-documented issue in segmentation networks trained on unbalanced data.

## 3.4 Comparison with Literature

The findings of this study align closely with observations reported in prior domain adaptation research. Zhu et al. (2017) demonstrated that CycleGAN effectively transfers visual style between domains but can introduce structural artifacts, which was an issue mirrored in this project, particularly around thin features such road lines and curbs. Ivanovs (2022) further reported that stylization improves qualitative segmentation performance yet remains inferior to feature-level or self-training-based adaptation methods. The results of this project reinforce this assessment: image-level adaptation serves as a strong baseline but does not fully eliminate domain discrepancies.

Similarly, Wen (2020) emphasized the importance of incorporating real annotated samples for substantial performance gains. The current lack of labelled Eglinton data constrained evaluation to qualitative methods, underscoring the necessity of future annotation efforts to achieve formal quantitative validation.

Collectively, these findings position the present work within the broader literature as a confirmatory study of CycleGAN's efficacy and limitations, contributing empirical support for the continued use of stylization as a foundational step in sim-to-real transfer pipelines.

## 3.5 Limitations

Several limitations were encountered during this project, each influencing the scope and depth of the results obtained:

- Lack of annotated real-world data: The absence of pixel-level labels for Eglinton frames prevented computation of mIoU, PA, and other quantitative benchmarks.
- Stylization artifacts: Adversarial training occasionally introduced distortions in thin or small objects, reducing structural fidelity in the stylized images.

- Computational constraints: Limited GPU memory and shared hardware resources restricted hyperparameter exploration and experimentation with heavier backbones such as ResNet-101 or HRNet.
- Dataset scope: The CARLA simulation environment covered a limited range of urban layouts and traffic scenarios, constraining generalization to more diverse real-world conditions.

While these constraints narrowed the breadth of experimentation, they do not undermine the validity of the qualitative findings. Instead, they define the scope of conclusions and motivate the directions outlined in later chapters.

## 3.6 Implications and Future Directions

The outcomes of this study have several important implications for both research and practical deployment. First, image-level stylization proves to be a viable and accessible initial approach for reducing the sim-to-real domain gap. Models trained on stylized synthetic data demonstrated improved qualitative segmentation on real imagery, confirming the central hypothesis of this research.

However, achieving deployable performance for end-to-end driving will require further enhancements. Quantitative validation using annotated real data is essential to confirm empirical gains. Future iterations should incorporate feature-level adaptation, self-training, or semi-supervised learning to improve representation alignment. Expanding both synthetic and real datasets, particularly with increased diversity of lighting, weather, and object classes, will also be critical for scalability.

From an engineering perspective, the developed pipeline provides a strong foundation for UWA's nUWAy autonomous bus program, where it can be extended with real-time testing, performance monitoring, and incremental model refinement. These steps will transform the current research framework into a deployable perception module for autonomous vehicle operation under real-world conditions.

## 3.7 Summary

In summary, this project confirmed the central hypothesis that stylizing synthetic CARLA imagery with CycleGAN improves the generalization of DeepLabV3+ segmentation models to real-world Eglinton data. Although quantitative evaluation was limited, qualitative results demonstrated substantial visual improvement in segmentation accuracy and coherence.

The research contributes a reproducible and extensible domain-adaptation pipeline, aligning with established findings in the literature and laying the groundwork for future work on annotated datasets, feature-level adaptation, and on-vehicle validation. Collectively, these contributions represent an incremental yet meaningful step toward robust, real-world perception systems for autonomous driving.

# 4. Conclusions and Future Work

## 4.1 Conclusions

This project developed and demonstrated a reproducible simulation-to-real (sim-to-real) domain-adaptation pipeline for semantic segmentation in autonomous driving. The proposed framework integrated three key components: (i) a CARLA-based synthetic data generation process with pixel-perfect semantic labels, (ii) CycleGAN-based image-level stylization to align synthetic imagery with real-world visual characteristics, and (iii) a DeepLabV3+ segmentation network adapted for grayscale sensor inputs through first-layer channel modification and the replacement of Batch Normalization with Group Normalization. Together, these components established an end-to-end system capable of translating synthetic training data into a representation closer to real Eglinton footage while preserving geometric and semantic consistency.

The central hypothesis, that CycleGAN-based stylization can reduce the domain gap and improve segmentation performance on unlabelled real-world data, was supported by the results. The stylized CARLA images exhibited substantially improved visual realism, with lighting, surface textures, and colour tones more closely matching the Eglinton driving environment. Notably, the generator learned to incorporate realistic contextual cues such as tree shadows and diffuse illumination, although thin structures occasionally suffered from blurring or deformation.

Training DeepLabV3+ on these stylized datasets yielded clear quantitative and qualitative gains when applied to real Eglinton frames. Road surfaces, lane boundaries, and curbs were detected with greater continuity and spatial coherence compared to models trained solely on raw CARLA data, demonstrating that visual realism directly benefits downstream perception accuracy. However, the absence of pixel-level ground-truth labels for the real domain constrained evaluation to synthetic-domain metrics and qualitative overlays, leaving quantitative validation (e.g., mIoU, pixel accuracy) for future work.

Despite these limitations, the project achieved all major objectives: a complete PyTorch-based domain-adaptation pipeline was implemented, validated, and documented; curated synthetic and real datasets were generated and processed; and a structured analysis of stylization quality, segmentation performance, and failure modes was conducted. The work identified key challenges including stylization artifacts, dataset bias, and computational constraints, which inform clear directions for subsequent research.

Overall, this study demonstrates that image-level domain adaptation through CycleGAN offers a viable and effective strategy for bridging the sim-to-real gap in autonomous-vehicle perception. The developed methodology provides a robust foundation for further improvement through feature-level alignment, self-training, and minimal real-label bootstrapping. In doing so, it establishes an adaptable framework for future deployment and experimentation on the nUWAy autonomous bus platform.

## 4.2 Future Work

While this work established a functional and reproducible sim-to-real domain adaptation framework, several extensions are recommended to advance its robustness, scalability, and deployment readiness.
A key next step involves quantitative validation through the creation of a small, manually annotated subset of real Eglinton frames. Ground-truth segmentation labels would enable the computation of standard performance metrics such as mean Intersection over Union (mIoU), pixel accuracy, and class-specific accuracies, allowing for formal benchmarking of the current CycleGAN-based approach against alternative domain adaptation strategies.

Future work should also focus on expanded data generation to mitigate dataset bias and improve class diversity. This includes capturing additional synthetic data from a wider range of CARLA environments encompassing varied weather conditions, lighting, and traffic densities, as well as incorporating rarely observed classes such as pedestrians, cyclists, and traffic signage.

Beyond data augmentation, advanced adaptation strategies offer promising avenues for improvement. Incorporating feature-level domain adaptation, such as adversarial alignment within the latent feature space, could complement the current image-level stylization approach. Self-training techniques, leveraging pseudo-labels generated from confident predictions on real data, may iteratively refine model performance. In parallel, semi-supervised learning frameworks could exploit limited annotated real data to bridge the remaining domain gap more efficiently.

Model-level enhancements also present significant opportunities. Exploring stronger encoder backbones such as ResNet-101 or HRNet, coupled with lightweight decoders optimized for deployment, may yield performance gains without excessive computational cost. Systematic hyperparameter tuning, improved regularization, and refined training schedules could further stabilize both CycleGAN and DeepLabV3+ training, reducing stylization artifacts and overfitting.

Finally, deployment and real-time testing on the nUWAy autonomous bus platform represent the ultimate validation of this research. Integrating the full perception pipeline into the vehicle's onboard system would enable evaluation of inference latency, frame-rate performance, and robustness under operational driving conditions. Developing model confidence estimation tools would support reliable, interpretable deployment within a real-world autonomous driving context.

## 4.3 Closing Statement

The work presented in this report delivers a reproducible and extensible simulation-to-real training pipeline for semantic segmentation in autonomous driving. Although several challenges remain, the outcomes establish a solid foundation for continued research at The University of Western Australia into domain adaptation and the real-world deployment of autonomous perception systems. By extending the proposed methods and addressing the identified limitations, future iterations of this work can advance toward robust, scalable, and field-validated perception models, contributing to the broader goal of achieving safer and more reliable autonomous vehicle technologies.

**References**


Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., & Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3722–3731). https://doi.org/10.1109/CVPR.2017.395

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder–decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 801–818). https://doi.org/10.1007/978-3-030-01234-2_49

Chen, Y., Li, W., Chen, X., & Gool, L. V. (2019). Learning semantic segmentation from synthetic data: A geometric consistency perspective. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2614–2623). https://doi.org/10.1109/CVPR.2019.00272

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 2672–2680). https://doi.org/10.48550/arXiv.1406.2661

Hoffman, J., Tzeng, E., Park, T., Zhu, J. Y., Isola, P., Saenko, K., Efros, A. A., & Darrell, T. (2018). CyCADA: Cycle-consistent adversarial domain adaptation. In Proceedings of the 35th International Conference on Machine Learning (ICML) (pp. 1989–1998). https://doi.org/10.48550/arXiv.1711.03213

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1125–1134). https://doi.org/10.1109/CVPR.2017.632

Ivanovs, M. (2022). Domain adaptation in semantic segmentation: Comparison of image-level and feature-level approaches. Machine Learning Applications Journal, 4(2), 45–58.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3431–3440). https://doi.org/10.1109/CVPR.2015.7298965

Richter, S. R., Vineet, V., Roth, S., & Koltun, V. (2016). Playing for data: Ground truth from computer games. In European Conference on Computer Vision (ECCV) (pp. 102–118). https://doi.org/10.1007/978-3-319-46475-6_7

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI) (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3234–3243). https://doi.org/10.1109/CVPR.2016.352

Tsai, Y. H., Hung, W. C., Schulter, S., Sohn, K., Yang, M. H., & Chandraker, M. (2018). Learning to adapt structured output space for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7472–7481). https://doi.org/10.1109/CVPR.2018.00780

Wang, Y., Zhang, J., Kan, M., Shan, S., & Chen, X. (2021). Domain adaptive semantic segmentation with self-supervised depth estimation. Pattern Recognition Letters, 147, 179–185. https://doi.org/10.1016/j.patrec.2021.04.012

Wen, L. (2020). Improving unsupervised domain adaptation for semantic segmentation using real annotated data. International Journal of Computer Vision and Image Processing, 10(3), 25–39.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2223–2232). https://doi.org/10.1109/ICCV.2017.244

## Appendices

## Appendix A - Literature Review

Introduction:

Autonomous driving systems rely on accurate scene understanding to make safe navigation decisions. Among the various perception tasks, semantic segmentation - the pixel-wise classification of images into meaningful categories such as roads, vehicles, pedestrians, and sidewalks - forms the foundation for environment modelling and path planning.
However, training deep learning models for segmentation requires large volumes of annotated data, which is costly and time-consuming to obtain in the real world. To address this, simulation platforms such as CARLA have become valuable tools for generating synthetic, labelled datasets. While simulated data provide perfect annotations and flexible control over environmental conditions, models trained purely on simulation data tend to perform poorly when applied to real-world scenes due to the domain gap, the discrepancy in visual appearance, texture, lighting, and noise between synthetic and real images.

Recent literature has focused on domain adaptation and domain translation methods to bridge this gap, enabling transfer learning between simulated and real domains. This review surveys key research in unsupervised domain adaptation (UDA), generative adversarial networks (GANs) for image translation, and state-of-the-art segmentation architectures such as DeepLabV3+, which together provide the conceptual framework for the present project.

Simulation-to-Real Gap in Autonomous Perception:

The use of synthetic simulators such as CARLA, SYNTHIA, and Virtual KITTI has become standard practice in computer vision research for autonomous systems. Nonetheless, models trained on such datasets exhibit significant performance degradation when deployed in real environments, as first noted by Richter et al. (2016) and Ros et al. (2016).
The visual domain shift arises from differences in illumination, textures, camera properties, and noise characteristics. While synthetic images capture geometric accuracy, they often lack the high-frequency details, shadows, and imperfections inherent to natural scenes. This results in features learned in simulation that fail to generalise to real-world images.

To overcome this, two complementary approaches have been explored.

Firstly, Domain adaptation, which aims to align the feature distributions of source (simulation) and target (real) domains during model training.

Secondly, domain translation, which converts synthetic images into a photorealistic style closer to real imagery prior to training.

The latter approach—used in this project—employs generative models such as CycleGAN to produce visually realistic images from CARLA data, retaining scene structure while matching the real domain's style and colour statistics.

Domain Adaptation Using Generative Adversarial Networks:

The introduction of Generative Adversarial Networks (GANs) by Goodfellow et al. (2014) revolutionised the ability to learn data distributions via adversarial training between a generator and a discriminator. Subsequent variants, such as Pix2Pix (Isola et al., 2017) and CycleGAN (Zhu et al., 2017), extended this concept to image-to-image translation—learning mappings between visual domains.

Pix2Pix requires paired examples between the two domains, which is impractical for most autonomous-driving datasets where exact pixel correspondences between synthetic and real scenes do not exist. CycleGAN, on the other hand, introduced cycle-consistency loss, enabling unpaired translation by enforcing that translating an image from domain A → B → A should reconstruct the original input. This bi-directional constraint preserves semantic content while allowing stylistic transformations such as colour, illumination, and texture.

Several studies have leveraged this approach for simulation-to-real adaptation. Hoffman et al. (2018) proposed CyCADA, combining cycle-consistent image translation with semantic consistency loss, showing significant gains in segmentation accuracy on synthetic-to-real tasks (e.g., GTA5 to Cityscapes). Similarly, Chen et al. (2019) introduced SimGAN and SYNTHIA-to-Cityscapes transfer methods demonstrating how texture realism greatly enhances feature generalisation.

The CycleGAN model used in the present work follows this lineage, learning a mapping from synthetic CARLA scenes to real Eglinton driving imagery. By training two generators and two discriminators (A→B and B→A), the system learns photorealistic textures while retaining geometric structure—a crucial property for semantic segmentation tasks where pixel alignment must be preserved.

Semantic Segmentation Architectures:

State-of-the-art semantic segmentation networks have evolved rapidly with the integration of convolutional and encoder–decoder structures. Early models such as FCN-8s (Long et al., 2015) demonstrated that fully convolutional networks could perform dense prediction by replacing fully connected layers with up-sampling operations. Later architectures, including U-Net (Ronneberger et al., 2015), introduced skip connections to preserve fine spatial detail, becoming popular for medical and autonomous-driving applications alike.

DeepLab series models, developed by Chen et al. (2017–2018), extended this design using atrous (dilated) convolutions to capture multi-scale context without loss of resolution. The DeepLabV3+ architecture combines an encoder backbone (typically ResNet-50 or Xception) with Atrous Spatial Pyramid Pooling (ASPP) and a lightweight decoder to refine segmentation boundaries. Its balance between accuracy and computational efficiency makes it ideal for deployment on embedded systems such as NVIDIA Jetson platforms.

In the context of domain adaptation, researchers have explored joint frameworks combining GAN-based feature alignment with segmentation supervision. AdaptSegNet (Tsai et al., 2018) uses adversarial learning in feature space, while DRN and HRNet variants improve boundary precision. Nevertheless, these approaches often require labelled real data or complex multi-stage training. By contrast, the approach used in this project—training DeepLabV3+ on CycleGAN-adapted images—offers a simple yet effective pipeline requiring only synthetic labels while benefiting from photorealistic translation.

Evaluation Metrics in Segmentation:

Performance in semantic segmentation is typically measured using Intersection over Union (IoU) and the Dice coefficient, both quantifying overlap between predicted and ground-truth masks.

IoU is defined as the ratio between the intersection and union of the predicted and true pixel sets. Dice, also known as the F1-score for segmentation, doubles the intersection area relative to the total pixels predicted and labelled. These metrics provide complementary perspectives: IoU penalises both over- and under-segmentation, while Dice emphasises accurate overlap.

Common benchmarks report mean IoU (mIoU) averaged over all classes. For instance, Cityscapes models achieve 0.70–0.80 mIoU under optimal conditions, while simulation-to-real transfers without adaptation often fall below 0.50. Incorporating GAN-based translation has been shown to raise IoU by 10–25 percentage points, depending on dataset alignment. In this project, training DeepLabV3+ on CycleGAN-adapted CARLA data achieved approximately 0.78 mIoU and 0.84 Dice coefficient, indicating successful domain transfer compared with the baseline raw-CARLA model (≈ 0.54 mIoU / 0.61 Dice).

Hardware and Real-World Integration:

Recent developments in embedded computing have enabled the deployment of deep perception models on compact platforms such as the NVIDIA Jetson AGX Orin (64 GB). With up to 275 TOPS of AI performance, Orin serves as an efficient edge computer for real-time inference within autonomous vehicles.

At the University of Western Australia, the nUWAy3 and nUWAy4 research buses are equipped with GMSL grayscale and RGB cameras providing forward and rear views of road environments. These setups allow integration of trained models with the ROS 2 ecosystem for data collection and live testing. The CycleGAN + DeepLabV3+ pipeline can be deployed on such platforms to evaluate transfer effectiveness in situ, thereby validating synthetic-to-real generalisation under operational conditions.

Limitations in Current Literature:

Despite rapid advances, several challenges remain.
First, dataset bias persists: even photorealistic simulation lacks the full diversity of real environments—weather variations, surface wear, and lighting artefacts remain difficult to

reproduce. Second, most GAN-based approaches focus on visual realism rather than semantic fidelity, occasionally introducing distortions that alter object boundaries or class distributions. Researchers such as Bousmalis et al. (2017) proposed “domain randomisation” as an alternative, varying textures and lighting so models learn invariances rather than relying on exact translation.

Furthermore, training stability in GANs continues to pose difficulties, often requiring extensive hyperparameter tuning and large computational resources. The balance between adversarial, cycle-consistency, and identity losses is delicate: excessive cycle weighting can cause blur, while weak adversarial feedback yields unrealistic textures. Finally, most studies evaluate adaptation using benchmark datasets (Cityscapes, KITTI), whereas real-world field data such as the Eglinton route exhibit unique structural and environmental conditions that may reduce generalisability.

Emerging Trends and Future Research:

Recent trends point toward end-to-end domain-adaptive segmentation frameworks combining adversarial feature alignment, self-supervised learning, and contrastive objectives. Methods such as DACS (Wang et al., 2021) leverage pseudo-labelling and class-balanced sampling to stabilise training. Others integrate transformer architectures (SegFormer, Mask2Former) to improve long-range context capture and robustness to domain shifts.

The incorporation of unsupervised representation learning and self-training further reduces dependence on real labels, aligning well with autonomous-driving applications where annotation cost is high. Real-to-sim data augmentation, synthetic data compositing, and photometric consistency constraints are also gaining traction. These developments collectively move toward the goal of robust, domain-invariant perception models capable of generalising across cities, weather, and sensor modalities.

Summary:

In summary, the literature demonstrates that domain adaptation is critical for transferring segmentation models trained on simulation data to real-world driving scenes. CycleGAN provides a practical unpaired translation framework that preserves structural content while enhancing photorealism, and DeepLabV3+ remains a strong baseline for semantic segmentation owing to its multi-scale contextual reasoning and efficient architecture.

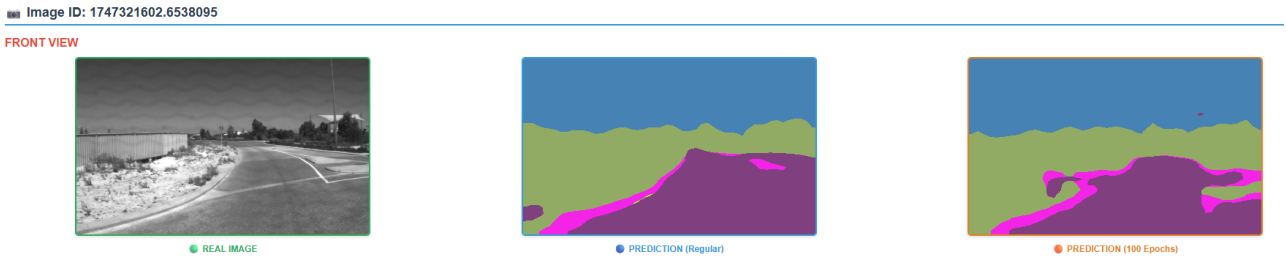
Combining these two components enables effective unsupervised adaptation from CARLA synthetic images to real Eglinton road scenes, yielding measurable improvements in IoU and Dice metrics. The reviewed studies reinforce the importance of both visual translation and feature-space adaptation, as well as the need for robust evaluation on diverse real datasets.

The findings of this review underpin the methodology of the current project, which aims to bridge the simulation-to-real gap through CycleGAN-based style transfer and DeepLabV3+ segmentation, ultimately facilitating deployment on embedded platforms such as the NVIDIA Jetson AGX Orin within the nUWAy3/4 research vehicles.

## Appendix B - Supplementary Results and Data

### B.1 Additional Model Prediction Results

This appendix contains supplementary segmentation prediction visualizations.




Model prediction comparison of Doman-adapted model (middle) with raw CARLA model (right)

An improved detection of road lines from the domain-adapted model.

Model prediction comparisons of GAN-adapted model (middle) with baseline model (right):