UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Semantic Segmentation for Autonomous Driving on Adverse Visual Conditions

Victor Hugo Sillerico Justo

Monograph - MBA in Artificial Intelligence and Big Data

ICMC USP SÃO CARLOS

Victor Hugo Sillerico Justo

Semantic Segmentation for Autonomous Driving on Adverse Visual Conditions

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence and Big Data

Advisor: Prof. Dr. Valdir Grassi Jr.

Original version

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

J96s

Justo, Victor Hugo Sillerico
Semantic Segmentation for Autonomous Driving on Adverse Visual Conditions / Victor Hugo Sillerico Justo; orientador Valdir Grassi Junior. -- São Carlos, 2024.
58 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Artificial Intelligence. 2. Semantic Segmentation. 3. Autonomous Driving. I. Junior, Valdir Grassi, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Victor Hugo Sillerico Justo

Segmentação Semântica para Direção Autônoma em Condições Visuais Adversas

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial e Big Data

Versão original

São Carlos

2024

This work is dedicated to my family, whose unwavering love, support, and encouragement have been my greatest source of strength.
Your belief in me has inspired and sustained me throughout this journey, and for that, I am forever grateful.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God for providing me with strength, wisdom, and perseverance throughout this journey. I extend my deepest gratitude to my family, whose constant love and support have been my foundation. Their encouragement and belief in my abilities have motivated me to achieve this milestone.

I would also like to express my sincere appreciation to my advisor, Prof. Dr. Valdir Grassi Junior, for his invaluable guidance and mentorship. His expertise, patience, and insightful feedback were instrumental in shaping the direction of this work. I am grateful for the time and dedication he invested in helping me navigate the challenges of this project.

Finally, I would like to acknowledge the program MBA in Artificial Intelligence and Big Data for providing me with the knowledge, resources, and opportunities to pursue this project. I am particularly grateful for the scholarship offered by the program, which made this academic journey possible. The financial support, along with the invaluable learning environment and encouragement from both the professors and colleagues, has been instrumental in my personal and professional growth.

"The only source of knowledge is experience."
Albert Einstein

ABSTRACT

JUSTO, V. **Semantic Segmentation for Autonomous Driving on Adverse Visual Conditions**. 2024. 58 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

This monograph addresses the challenge of semantic segmentation for nighttime autonomous driving, a critical task in ensuring safety and reliability in autonomous vehicle systems under low-light conditions. The study evaluates the performance of four deep learning models—UNet, FPN, PSPNet, and DeepLabV3+—trained on a hybrid dataset comprising both real and synthetic nighttime images. The real images were sourced from the ACDC dataset, while the synthetic images were generated using the CARLA driving simulator. The models were trained to segment six key classes: road, vegetation, building, sky, car, and background. Quantitative evaluation using Intersection over Union (IoU) and F1-score metrics demonstrated promising results across the models. According to those metrics, the best model was the Feature Pyramid Network, achieving a mean F1-score of 88.32% and a mean IoU of 81.12%. However, qualitative analysis revealed that while synthetic data helps increase the volume of nighttime scenarios, it alone is insufficient to achieve high-quality segmentation performance, particularly in complex nighttime environments. This highlights the limitations of relying solely on synthetic datasets to improve real-world application segmentation outcomes. To make this work easier to reproduce, the CARLA-Night dataset has been made available on Kaggle (https://www.kaggle.com/datasets/victorsillericojusto/carla-night), and all the code developed for the project can be accessed through a GitHub repository (https://github.com/victorsillerico/segmentation-nighttime.git).

Keywords: Semantic Segmentation. Autonomous Driving.

RESUMO

JUSTO, V. **Segmentação Semântica para Direção Autônoma em Condições Visuais Adversas**. 2024. 58 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Esta monografia aborda o desafio da segmentação semântica para direção autônoma noturna, uma tarefa crítica para garantir a segurança e a confiabilidade em sistemas de veículos autônomos sob condições de pouca luz. O estudo avalia o desempenho de quatro modelos de aprendizado profundo — UNet, FPN, PSPNet e DeepLabV3+ — treinados em um conjunto de dados híbrido que compreende imagens noturnas reais e sintéticas. As imagens reais foram obtidas do conjunto de dados ACDC, enquanto as imagens sintéticas foram geradas usando o simulador de direção CARLA. Os modelos foram treinados para segmentar seis classes principais: estrada, vegetação, edifício, céu, carro e fundo. A avaliação quantitativa usando Intersection over Union (IoU) e métricas de pontuação F1 demonstrou resultados promissores em todos os modelos. De acordo com essas métricas, o melhor modelo foi o Feature Pyramid Network, alcançando uma pontuação F1 média de 88,32% e uma IoU média de 81,12%. No entanto, a análise qualitativa revelou que, embora os dados sintéticos ajudem a aumentar o volume de cenários noturnos, eles sozinhos são insuficientes para atingir um desempenho de segmentação de alta qualidade, particularmente em ambientes noturnos complexos. Isso destaca as limitações de confiar apenas em conjuntos de dados sintéticos para melhorar os resultados de segmentação em aplicações do mundo real. Para facilitar a reprodução deste trabalho, o conjunto de dados CARLA-Night foi disponibilizado no Kaggle (https://www.kaggle.com/datasets/victorsillericojusto/carla-night), e todo o código desenvolvido para o projeto pode ser acessado por meio de um repositório GitHub (https://github.com/victorsillerico/segmentation-nighttime.git).

Palavras-chave: Segmentação Semântica. Direção Autônoma.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

UNET	U-shaped Network
PSPNET	Pyramid Scene Parsing Network
DEEPLABV3	Deep Convolutional Neural Networks for Semantic Segmentation
FPN	Feature Pyramid Network
ASPP	Atrous Spatial Pyramid Pooling
PICOC	Population, Intervention, Comparison, Outcome, and Context
DCNN	Deep Convolutional Neural Networks
FCN	Fully Convolutional Networks
CMoDE	Convoluted Mixture of Deep Experts
GAN	Generative Adversarial Networks
IAPM	Image-Adaptive Processing Module
LGF	Learnable Guided Filter
SOD	Semantic-Oriented Disentanglement
IAPARSER	Illumination-Aware Parser
SGD	Stochastic Gradient Descent
ACDC	Adverse Conditions Dataset with Correspondences
GPS	Global Positioning System
COLAB	Google Colaboratory
CARLA	Car Learning to Act
RESNET	Residual Neural Network
IOU	Intersection Over Union

CONTENTS

1 INTRODUCTION

1.1 Context

Driver-less technology has been the focus of extensive research efforts by industry and academia. Different platforms will take on-road action in the next years including single passenger self-driving cars, delivery robots, and heavy-duty autonomous trucks. However, further research is necessary to improve the navigation stack including perception, prediction, planning, and control to get fully autonomous driving. A critical task to improve the navigation capabilities of robots is semantic segmentation, which is useful for discerning and delineating objects within an image with remarkable precision (Sellat; Bisoy; Priyadarshini, 2022). As an essential component of scene understanding, semantic segmentation involves the classification of individual pixels or regions in an image into semantically meaningful categories, thus providing a fine-grained understanding of the visual content (Siam *et al.*, 2018).

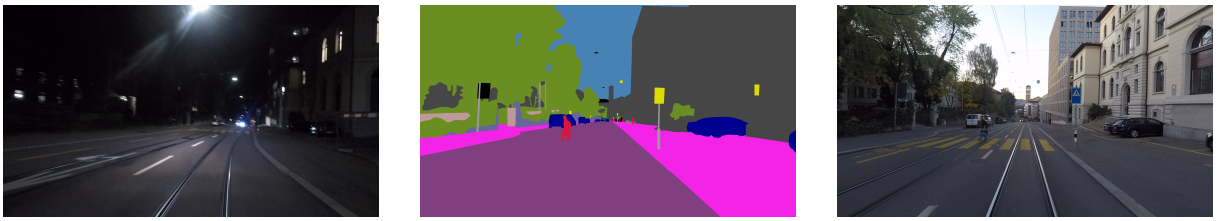
Significant progress has been made in semantic segmentation in recent years, and learning-based techniques have achieved promising results in terms of performance and generalization capabilities (Schwonberg *et al.*, 2023). However, it is difficult to design learning-based models for segmentation tasks due to application-dependent aspects, and datasets dominated by images captured under normal conditions that reduce the capability of the model to deal with any environmental context. In particular, adverse visual conditions, such as low-light, fog, rain, or snow, can impact the quality and clarity of images (Sakaridis; Dai; Gool, 2021). By including such challenging conditions in the training dataset, the model learns to handle real-world scenarios more effectively, which improves the robustness of the model, allowing it to perform well in varying environmental conditions.

This project will focus on semantic scene understanding for autonomous driving applications in low-light conditions. It involves comparing various learning methods documented in the literature. Therefore, this work aims to identify a neural network model that excels in semantic segmentation of scenes within challenging visual domains, particularly those involving nighttime and low-light conditions. Furthermore, we intend to assess the model's ability to generalize by testing it with images from diverse contexts.

1.2 Justification and Motivation

Scene understanding of outdoor environments requires a robust visual perception system that can parse input images under changing environmental conditions throughout the day and across seasons (Sakaridis; Dai; Gool, 2021). Robots should be equipped with perception models to deal with challenging environmental factors to be operable and

ensure safety in the real world (Valada *et al.*, 2017). For instance, adverse weather and illumination conditions (e.g. fog, rain, snow, low light, nighttime, glare, and shadows) create visibility problems for the sensors that power automated systems (Fig. 1). Further research in this topic is necessary because the performance of current vision algorithms is still mainly evaluated under clear weather conditions (e.g. good weather, favorable lighting), and even the top-performing algorithms undergo severe performance degradation under adverse conditions (e.g. night, reduce illumination).



(a) Nighttime image. (b) Segmented image. (c) Daytime image.

Figure 1 – Semantic segmentation on adverse environmental conditions.

Semantic segmentation models have to consider the trade-off between robustness and efficiency, as well as the intrinsic limitations related to computational/memory bounds and data-scarcity (Schwonberg *et al.*, 2023). Therefore, based on the great progress that learning-based solutions have made in recent years, this work aims to study, train, and evaluate Deep Neural Network models that can achieve robust results in semantic segmentation tasks under challenging environmental conditions.

1.3 Research Questions and Objectives

In this project, we consider the hypothesis that it is possible to train a neural network model using images collected in non optimal illumination conditions, so that the model is capable to extract useful information for semantic scene understanding in autonomous driving applications. Given the challenges and problems currently faced in semantic segmentation for autonomous vehicles, the aforementioned hypothesis motivates the following research questions:

Q1 *"How do different semantic segmentation techniques compare in their accuracy and robustness for segmenting key objects (pedestrians, vehicles, lane markings) in nighttime driving scenarios?"*

Q2 *"What specific modifications to existing segmentation frameworks are most effective in improving performance under adverse lighting conditions like night driving?"*

Given these research questions, the following objectives are defined for the development of this research project:

- Carry out a literature review on semantic segmentation methods for autonomous driving applications in challenging domains related to illumination conditions.
- Collect image-based datasets to train Deep Learning models that are suitable for semantic segmentation in adverse visual conditions.
- Test learning-based models available in the literature with images taken on adverse lighting conditions, and analyze their generalization capabilities.
- Create a novel dataset of artificial urban images using the CARLA simulator in realistic nighttime conditions, aimed at enhancing the performance of semantic segmentation models for autonomous driving in low-light environments.
- Evaluate the performance of neural network models trained for semantic segmentation, and compare the results to identify the best solution.

2 TECHNICAL BACKGROUND

2.1 Nighttime Semantic Segmentation

Nighttime semantic segmentation is an area of computer vision that focuses on accurately labeling the different objects and regions in an image captured at night. This task is challenging because nighttime images have lower light and different visual properties compared to daytime images, which can fool standard segmentation models (Valada *et al.*, 2017).

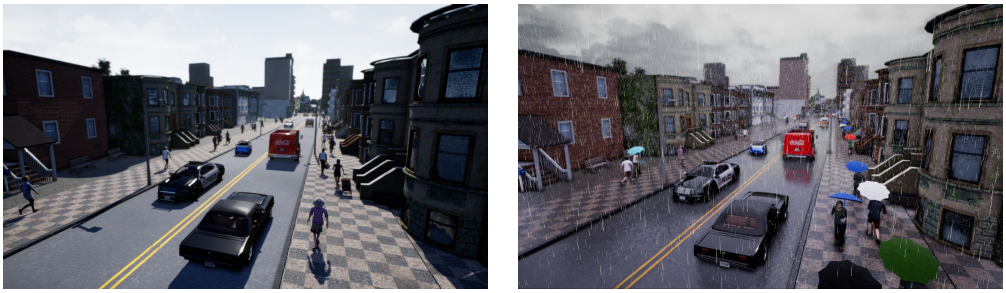
In the critical field of autonomous vehicles, nighttime semantic segmentation becomes even more crucial. Here, precise understanding of the surrounding environment at night is essential for safe navigation. Since autonomous vehicles rely on camera data, the ability to segment objects like pedestrians, vehicles, and lanes even in low-light conditions is paramount (Wu *et al.*, 2023). Nighttime semantic segmentation research aims to develop robust models that can overcome these challenges and ensure the continued safe operation of self-driving cars.

2.2 CARLA Simulator

CARLA (Car Learning to Act) is an open-source simulator to carry out research about autonomous vehicles. CARLA has been created to support prototyping, training, and validation of modern autonomous urban driving systems, including both perception and control. CARLA provides open-source code, protocols, and open digital assets (urban layouts, buildings, vehicles) that can be used freely. Moreover, CARLA is flexible, it supports different environmental conditions including weather and time of day (see Fig. 2), it also support setup of a wide range of sensors, and provides useful signals such as GPS coordinates, speed, acceleration, and data on collisions and other infractions. (Dosovitskiy *et al.*, 2017).

CARLA supports development, training, and detailed performance analysis of autonomous driving systems. Dosovitskiy *et al.* (2017) used CARLA to evaluate modern approaches to autonomous driving. They tested a classic modular pipeline compound by dedicated subsystems for visual perception, planning, and control. They also tested models that have received much attention by the research community in recent years, such as those approaches based on deep networks trained end-to-end using either imitation learning or reinforcement learning.

Figure 2 – A street in Town 2 of CARLA Simulator showing different weather conditions.



(a) Clear day. (b) Daytime rain.

Source: Extracted from (Dosovitskiy *et al.*, 2017)

2.3 Baseline Models

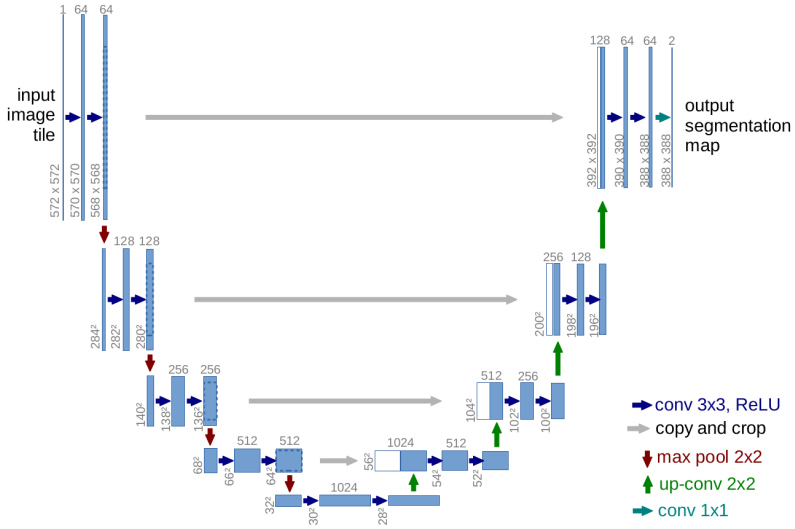
Building effective nighttime semantic segmentation models often starts with establishing strong baseline models. These baselines are typically established by daytime segmentation models that are then adapted or enhanced to perform better under low-light conditions. This adaptation can involve various techniques like incorporating strategies to handle the specific challenges of nighttime imagery, such as reduced color information and increased noise. By using well-performing daytime models as a foundation, researchers can develop nighttime segmentation models that are more robust and effective for tasks like autonomous driving.

U-Net, PSPNet, DeepLabV2, and FPN were used as baseline models in this project, all of them have strengths and weaknesses for nighttime use. U-Net and DeepLabV2 might need adjustments to handle the limitations of capturing long-range dependencies or needing large datasets for training. PSPNet excels in capturing global context crucial for low-light conditions. FPN is useful for nighttime segmentation, where generating multi-scale feature maps with strong semantic information at various resolutions is essential for capturing details across different scales.

2.3.1 U-Net

Ronneberger, Fischer and Brox (2015) designed U-Net for semantic segmentation tasks. It excels at pixel-wise classification, meaning it assigns a specific class label (e.g., car, person, road) to each pixel in an image. U-Net's strength lies in its unique structure: a contracting encoder path that captures contextual information and a corresponding expanding decoder path that recovers spatial resolution (see Fig. 3). Originally developed for biomedical image analysis, U-Net's versatility has led to its adoption in various fields like autonomous driving and satellite imagery analysis.

Figure 3 – U-Net architecture.

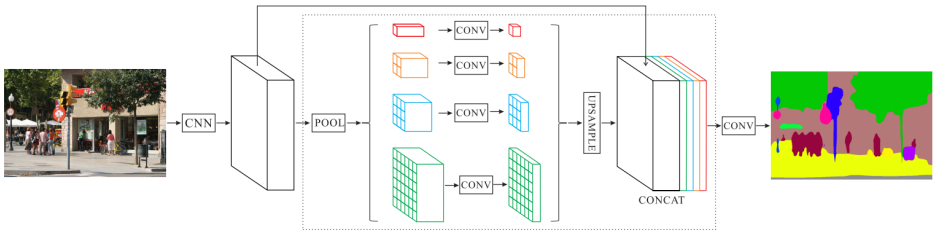


Source: Extracted from (Ronneberger; Fischer; Brox, 2015)

2.3.2 PSPNet

PSPNet (Zhao *et al.*, 2017) addresses semantic segmentation by incorporating global scene context through a unique pyramid pooling module (see Fig. 4). This module goes beyond traditional single-scale feature extraction by capturing information at various resolutions, making it possible to understand the big picture of a scene while preserving details of individual objects. PSPNet utilizes the pyramid pooling module to generate multi-scale feature representations that are upsampled and combined, creating a comprehensive scene understanding. Finally, a decoder refines this information and assigns class labels to each pixel, resulting in improved segmentation accuracy, particularly for complex scenes.

Figure 4 – PSPNet architecture.



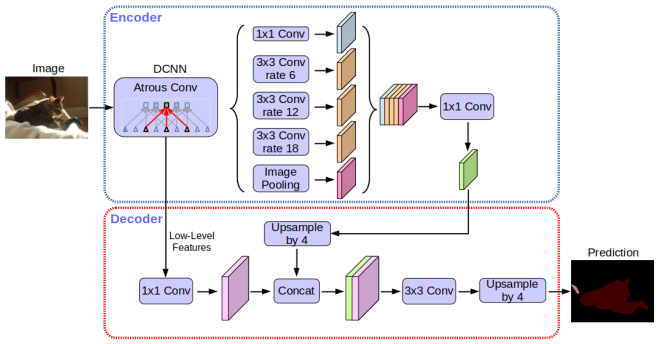
Source: Extracted from (Zhao *et al.*, 2017)

2.3.3 DeepLab

Chen *et al.* (2018) developed DeepLab using dilated convolutions to capture long-range dependencies in the image without increasing the number of parameters or losing resolution. DeepLab models typically consist of an encoder backbone (see Fig. 5), often a pre-trained classification network like ResNet, modified with Atrous convolutions.

Additional modules like Atrous Spatial Pyramid Pooling (ASPP) can be incorporated to capture objects at various scales. DeepLab has evolved through several versions, each offering improvements in accuracy and efficiency, making it a popular choice for various semantic segmentation applications.

Figure 5 – DeepLab architecture.

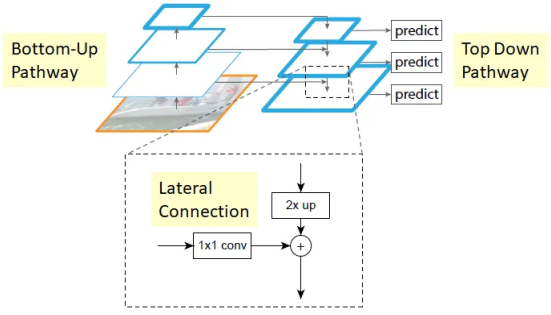


Source: Extracted from (Chen *et al.*, 2018)

2.3.4 Feature Pyramid Network

A Feature Pyramid Network (FPN) (Lin *et al.*, 2017) is a feature extractor that generates multi-scale feature maps from a single-scale image, making it suitable for tasks like object detection. It works independently of the backbone architecture, constructing a feature pyramid through two pathways: a bottom-up pathway and a top-down pathway (see Fig 6). The bottom-up pathway uses the feedforward process of the backbone network to compute feature maps at multiple scales, typically halving the resolution at each stage. The top-down pathway enhances these maps by upsampling coarser, semantically rich features and combining them with higher-resolution maps from the bottom-up process via lateral connections. This combination results in feature maps that are semantically strong yet accurately localized.

Figure 6 – FPN architecture.



Source: Extracted from (Lin *et al.*, 2017)

2.4 Related Works

The Systematic Literature Review was done considering the the PICOC (Population, Intervention, Comparison, Outcome, and Context) criteria to break down the objectives into searchable keywords and help formulate research questions (see Table 1). This strategy enjoys widespread use in medical and social science research (Petticrew; Roberts, 2008), and was adapted by Carrera-Rivera *et al.* (2022) to define the scope and objectives of Literature Reviews for computer science research.

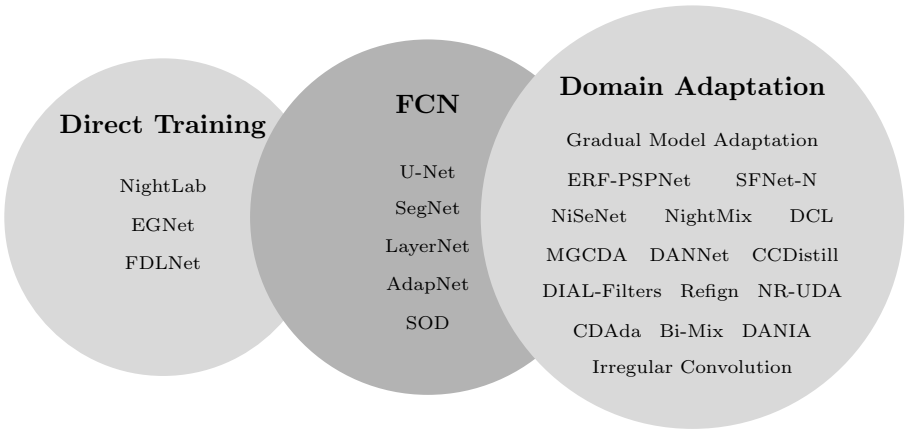
Table 1 – PICOC criteria for Systematic Literature Review.

Population	Real-world images captured in nighttime driving conditions
Intervention	A deep learning architecture for robust segmentation in low-light
Comparison	U-Net, PSPNet, DeepLab, and RefineNet as baseline models
Outcome	Mean IoU, and F1-score for pedestrians, vehicles, and lane markings
Context	Autonomous vehicles requiring accurate and reliable segmentation

Source: Made by the author.

Considering the research questions defined in Chapter 1, and search sources like IEEE Xplore, ACM, Scopus, Web of Science, it was possible to elaborate a taxonomy of nighttime semantic segmentation (see Fig. 7). Therefore, the literature work in semantic segmentation is categorized into three main subcategories: (1) Fully Convolutional Networks. (2) Direct Training Models. (3) Domain Adaptation Models.

Figure 7 – Taxonomy of nighttime semantic segmentation models.



Source: Made by the author.

2.4.1 Fully Convolutional Networks Methods

Semantic segmentation has seen a dramatic improvement thanks to Deep Convolutional Neural Networks (DCNN), especially with the rise of Fully Convolutional Networks

(FCNs). They achieve this through a two-part architecture: an encoder that shrinks the image while capturing key features, and a decoder that expands the information back to the original size for segmentation output. For instance, LayerNet (Li; Liu; Yang, 2023) uses a multi-head decoder and a well-designed hierarchical module to model, extract and fuse multistage features of different depths in nighttime images.

Several refinements on FCNs have been introduced to improve segmentation accuracy (Siam *et al.*, 2018). For example, Valada *et al.* (2017) include context and multi-modal information to refine the inference results by reducing the sensitivity to appearance variations, which is typical on perception systems using unimodal images as inputs. Its Convoluted Mixture of Deep Experts (CMoDE) fusion scheme selects class-specific features from expert networks based on the current scene representation, and learns deeper representations from the mixture of kernels.

2.4.2 Model Adaptation Methods

The lack of large-scale labeled datasets in the nighttime scenes motivates the development of domain adaptation approaches that transfer the knowledge from the day-time scenes to night-time (Dai; Gool, 2018a; Romera *et al.*, 2019; Sakaridis; Dai; Gool, 2019; Wu *et al.*, 2021; Lee *et al.*, 2023).

Supervised and unsupervised semantic segmentation research constantly produce new approaches and advancements in domain adaptation. For instance, Dai and Gool (2018b) proposed an unsupervised method that progressively adapts the semantic models trained on daytime scenes to nighttime scenes by dividing the twilight time between day and night into three subgroups considering the elevation of the sun, and makes incremental adjustments to the network using pseudo labels. Alternatively, Cheng *et al.* (2022) fused supervised daytime scenes and unsupervised night-time scenes. The supervision information in the daytime scene and the texture information specific to the night-time scene are fully utilized, and the model is adapted to both the daytime scene and the night-time scene.

Other methods explore Curriculum Learning [1] (self-learning) for domain adaptation to bridge the inter-domain and intra-domain gap together without additional data or network. For example, CDAda (Xu *et al.*, 2021) uses entropy minimization and a pseudo-label self-training method to adjust the model according to the level of difficulty attached to the domain, which enables smoother semantic knowledge transfer. Sakaridis, Dai and Gool (2022) developed a curriculum framework where the model progressively learns from easier (brighter) to harder (darker) nighttime images. This approach leverages correspondences between daytime reference images and their darker counterparts to guide the label inference of the model for nighttime scenes.

[1] The process of introducing learning from easy to complex, is typically used to promote the optimization of non-convex problems

Given the limitations of traditional methods in low-light scenarios, researchers are exploring the potential of Generative Adversarial Networks[2](GANs) to improve semantic segmentation performance in nighttime environments. Creswell *et al.* (2018) embedded GANs in domain adaptation frameworks generating promising results. Moreover, Song *et al.* (2022) improved the performance of night segmentation with a system compounded by an appearance transferring module that transfers unlabeled images, acquired during both daytime and nighttime, into a shared latent feature space, which encodes the image content of both scenes at the semantic level.

Recent works (Romera *et al.*, 2019; Sun *et al.*, 2019) also use GANs to effectively reduce the domain gaps by learning the mapping of input images to output images, and improving the segmentation performance from two perspectives, including direct inference of night images and real-time online conversion inference of night images through style conversion. Yang, Han and Liu (2023) combine a fuzzy information complementing strategy using generative models to fill in missing information, with a network that fuses different processing stages to capture richer spatial context. The scheme further incorporates irregular convolutional attention modules to focus on specific regions and extract detailed boundaries of moving targets.

Others methods (Anoosheh *et al.*, 2019; Schutera *et al.*, 2021) have embedded pre-trained image enhancement modules on the pipeline to translate night-time images into their day-time counterparts. For instance, Wang *et al.* (2022) developed a nighttime segmentation framework composed of two parts: an image enhancement module which introduces semantic information, and a segmentation network with strong feature extraction capability. Furthermore, Yang *et al.* (2021) introduced a Bidirectional Mixing (Bi-Mix) framework that leverages the information between coarsely aligned day-night image pairs to improve translation-adaptation and the segmentation-adaptation processes.

In order to enhance the results before and after the segmentation network, Liu *et al.* (2023) exploited the intrinsic features of driving-scene images under different illuminations using DIAL-Filters, which consist of an Image-Adaptive Processing Module (IAPM) and a Learnable Guided Filter (LGF). In addition, Brüggemann *et al.* (2023) proposed the REFING framework that leverages existing correspondences between images in normal and adverse conditions. It achieves this in two stages: first, an uncertainty-aware dense matching network aligns the normal image to its adverse counterpart. Second, an adaptive label correction mechanism refines the adverse condition prediction using the aligned normal image prediction.

Multi-stage approaches, e.g. NiSeNet (Nag; Adak; Das, 2019), that use twilight as an intermediate domain between day and night to perform a multi-stage adaptation, have

[2] A type of deep learning system where two neural networks compete to create new data that is indistinguishable from real data

been proved to introduce more computational burden. Twilight scene images are more difficult to capture due to its strict definition according to the solar elevation angle. In order to overcome this challenge, DANNet (Wu *et al.*, 2021) performs domain adaptation in one-stage by using an adversarial training with labeled daytime data and unlabeled roughly aligned day-night image pairs. Moreover, DANIA (Wu *et al.*, 2023), a one-stage adaptation framework for nighttime semantic segmentation, leverages a labeled daytime dataset (the source domain) and an unlabeled dataset that contains coarsely aligned day-night image pairs (the target daytime and nighttime domains). It does not need to train additional day-night image transfer models as a separate pre-processing stage.

Other one-stage domain adaptation network is CCDistill (Gao *et al.*, 2022). It extracts the content and style knowledge contained in features, and measures the level of illumination difference between two images to deal with the lack of labels for nighttime images. The adaptation is achieved using the invariance of the same kind of difference.

2.4.3 Direct Training Methods

Even though acquiring large amounts of high-quality nighttime data can be expensive and time-consuming, some methods use different strategies train the model directly on labeled night-time images. For instance, EGNet (Tan *et al.*, 2021) considers the HSV color space and uses the channel V to generate discriminating features in under- and over-exposed regions. The NightLab (Deng *et al.*, 2022) architecture is other example of direct training that uses a Hardness Detection Module to divide objects into simple and difficult categories. Finally, Xie *et al.* (2023) explored the image frequency distributions for night-time scene parsing. However, these methods do not take into account the negative impact and degradation effect of lighting on semantic segmentation tasks but instead implicitly force the network to learn the entangled representations of various content and lighting.

Disentangling the accidental scene events, such as reduced illumination, improves the performance of computer vision tasks in different ways. For instance, Deep Learning models can capture the isolated factors of variation affecting the represented entities, improving their robustness to diverse conditions. Previous works have explored different techniques for disentangling the image representations (Baslamisli *et al.*, 2018), such as learning domain invariant representations across different domains. Wei *et al.* (2023) proposed Disentangle Then Parse (DTP) approach that consists of two key components. First, DTP has a Semantic-Oriented Disentanglement (SOD) framework to extract the reflecting component to consistently identify the semantics under cover of varying and complicated lighting conditions. Second, DTP has an Illumination-Aware Parser (IAParser) to explicitly learn the correlation between semantics and lighting, and include the illumination features to make precise predictions.

Table 2 – Intersection over Union (IoU) results obtained by methods for semantic segmentation of nighttime images.

Reference	Model	Backbone	IoU Score(%)
Wei *et al.* (2023)	SOD	DeepLabV3+	63.7
Ding, Li and Tian (2023)	DCL	PSPNet	50.4
Li, Liu and Yang (2023)	LayerNet	ResNet	65.3
Wang *et al.* (2022)	SFNet-N	ResNet50	56.9
Sakaridis, Dai and Gool (2022)	MGCDA	RefineNet	42.5
Xu *et al.* (2021)	CDAda	RefineNet	45.0
Wu *et al.* (2021)	DANNet	ResNet101	45.2
Tan *et al.* (2021)	EGNet	ResNet101	45.3
Nag, Adak and Das (2019)	NiSeNet	DeepLabV3+	45.56
Valada *et al.* (2017)	Adapnet	ResNet50	71.72
Cheng *et al.* (2022)	NightMix	Zero-DCE	46.96
Sun *et al.* (2019)	ERF-PSPNet	CycleGAN	45.09
Yang, Han and Liu (2023)	Irregular-Conv	DeblurGAN	94.2
Deng *et al.* (2022)	NightLab	ReLAM	62.82
Xie *et al.* (2023)	FDLNet	UperNet	52.68
Dai and Gool (2018b)	GMA	RefineNet	41.6
Gao *et al.* (2022)	CCDistill	RefineNet	47.5
Yang *et al.* (2021)	Bi-Mix	RefineNet	46.5
Wu *et al.* (2023)	DANIA	PSPNet	52.6
Brüggemann *et al.* (2023)	REFIGN	DeepLabV2	65.5
Liu *et al.* (2023)	DIAL-Filters	ResNet-101	51.21
Song *et al.* (2022)	NR-UDA	ResNet	16.93

Source: Made by the author.

2.5 Final Considerations

The reviewed related works showcase various approaches for dealing with the challenge of nighttime semantic segmentation. From established supervised learning methods using daytime data to cutting-edge techniques like unsupervised learning and GANs, researchers are actively exploring avenues for robust segmentation in low-light conditions. Table 2 summarizes the Intersection over Union[3] (IoU) results presented in the reviewed research works. It is difficult to directly compare the performance of the models because

[3] It is a metric used to evaluate the accuracy of a model's predictions compared to the ground truth (actual labels) for each pixel in an image.

they do not use the same datasets and backbone networks. Most of the models use RefineNet, PSPNet, DeepLab and its variations as backbone and make adjustment in the architecture to support the segmentation task in low-light conditions.

Therefore, nighttime semantic segmentation remains an active area of research with significant potential for various applications. By addressing data acquisition challenges, developing robust and generalizable models, exploring sensor fusion, and optimizing for efficiency, it is possible to push the boundaries of this critical field.

In this project, the focus was on methods within the category of Fully Convolutional Networks (FCNs), which are well-suited for pixel-wise prediction tasks. The models employed, such as U-Net, PSPNet, DeepLabV3+, and FPN, are all part of this FCN family. These architectures leverage end-to-end training with convolutional layers to directly generate segmentation maps, making them highly effective for this application. This choice of approach reflects our aim to prioritize models that excel at learning complex spatial hierarchies, which are crucial for accurately identifying objects like vehicles, buildings, vegetation, and lane markings in low-light conditions.

3 METHODOLOGY

3.1 Project Study

In this project, baseline models for semantic segmentation were implemented to compare their performance with nighttime autonomous driving scenes. The main goal was to determine which architecture presents the best results in terms of accuracy and robustness to segment images of light-constrained scenarios, considering a training set composed of real images and synthetic images. The code of the implementation can be accessed through a GitHub repository (https://github.com/victorsillerico/segmentation-nighttime.git).

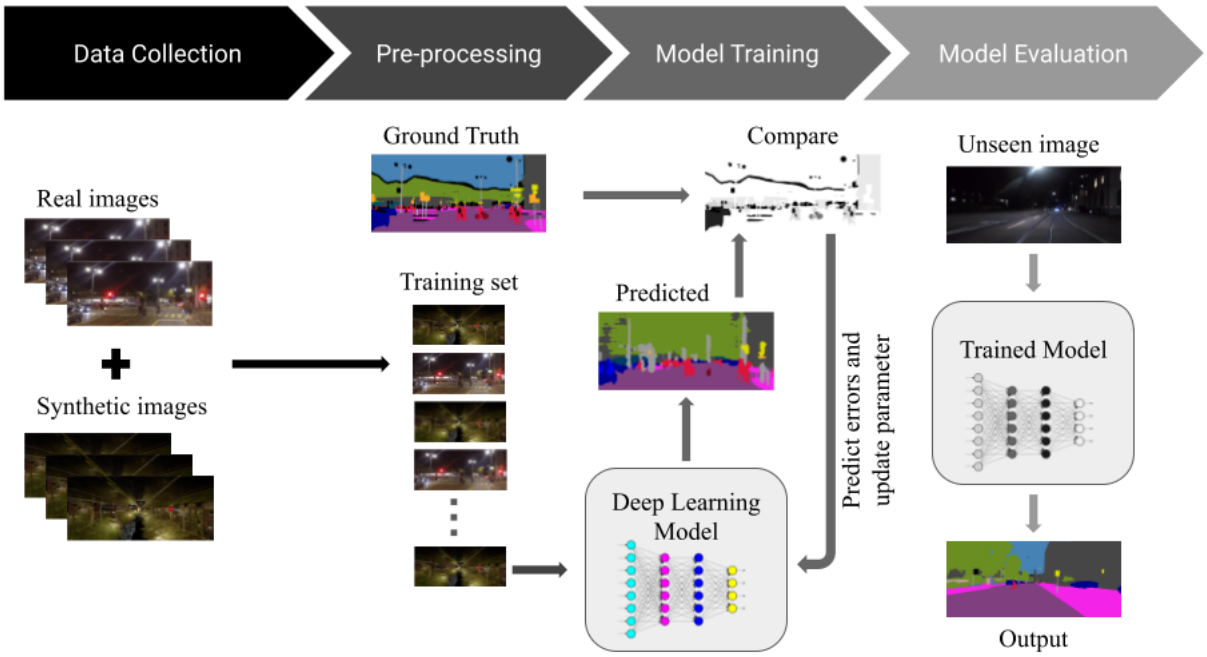
The proposed pipeline is presented in Fig. 8. Initially, nighttime urban images were collected from specialized datasets for autonomous driving, and combined with a set of synthetic images collected and annotated to create segmentation maps, labeling each pixel according to its class (e.g., road, pedestrian, vehicle). The images for the synthetic dataset were created using the autonomous vehicle simulator CARLA[1], in which a vehicle in autopilot mode navigates around urban scenarios taking images under different weather and daytime conditions. Then, data augmentation techniques, such as rotation, cropping, and brightness adjustment, are applied to enhance data variety and volume. The dataset is divided into training, validation, and testing sets, typically in a 70-20-10 ratio.

Training a model for semantic segmentation in nighttime autonomous driving involves several key steps. Initially, nighttime urban images are collected and annotated to create segmentation maps, labeling each pixel according to its class (e.g., road, pedestrian, vehicle). Various CNN architectures like U-Net, SegNet, DeepLab, and PSPNet are configured with specific hyperparameters (learning rate, epochs, batch size). During training, images are passed through the model to generate segmentation predictions, and a loss function (e.g., cross-entropy or Jaccard index) measures the error against true labels. Gradients are calculated and weights updated via backpropagation using optimizers like Adam or SGD. The model's performance is validated after each epoch to prevent overfitting.

Post-training, the model is evaluated on the test set using metrics such as accuracy, IoU, and F1-score. Fine-tuning, including hyperparameter adjustments and transfer learning from pre-trained models, is conducted based on validation and test results to enhance performance, ensuring robust and accurate segmentation in low-light conditions critical for safe nighttime autonomous driving.

[1] https://carla.org//

Figure 8 – Proposed pipeline.



Source: Made by the author.

3.2 Datasets

Many large-scale image-based datasets have been proposed for urban scene understanding, targeting autonomous driving (AD) scenarios, e.g. Cityscapes (Cordts *et al.*, 2016), BDD100K (Yu *et al.*, 2020), CamVid (Brostow; Fauqueur; Cipolla, 2009), KITTI (Geiger; Lenz; Urtasun, 2012). They include images captured under normal visual conditions during daytime and in clear weather. However, the perception capabilities of autonomous vehicles impose strict requirements on algorithms to maintain satisfactory performance in adverse domains.

New datasets have been proposed in response to this need for large-scale driving datasets specialized for challenging perceptual conditions, in terms of size, domain adversity, and featured tasks. For instance, the Adverse Conditions Dataset with Correspondences (ACDC) (Sakaridis; Dai; Gool, 2021) includes 4006 images evenly distributed across four common adverse conditions that include fog, nighttime, rain, and snow. Each image taken under adverse conditions is accompanied by a high-quality, fine pixel-level semantic annotation, a corresponding image of the same scene under normal conditions, and a binary mask. The recordings were made using a 1080p GoPro Hero 5 camera at 30 frames per second, with the camera positioned differently depending on the condition (in front for nighttime and normal conditions).

Dark Zurich (Sakaridis; Dai; Gool, 2022) dataset is also suited to train models

for semantic segmentation of images collected in light-constrained scenarios. It contains 8779 images captured at nighttime, twilight, and daytime, along with the respective GPS coordinates of the camera for each image. These GPS annotations are used to construct cross-time-of-day correspondences, to match each nighttime or twilight image to its daytime counterpart. This dataset has 201 nighttime images (151 test + 50 validation) with fine pixel-level semantic annotations for the 19 evaluation classes of Cityscapes.

Nightcity (Tan *et al.*, 2021) dataset is other alternative to develop models for scene parsing in reduced illumination conditions. It is based on a collection of real nighttime driving videos (which were captured using a Driving Recorder during car driving) over the Internet from various cities (e.g. Los Angeles, New York, Chicago, Hong Kong, London,Tokyo and Toronto). These videos cover urban street, highway and tunnel scenarios. Then, 297 diverse images were selected with no obvious motion blur from these videos for manual annotation, following the approach used to construct the Cityscapes dataset.

Table 3 provides a summary of such autonomous driving oriented semantic segmentation datasets with their most important characteristics: the number of classes, the number of annotated samples, whether images are real or rendered, whether the dataset contains video sequences (and not only temporally uncorrelated images), the geographical location (for what concerns simulated datasets, we report the simulated area indicated, if available), and whether the dataset allows setting arbitrary conditions (seasonal, weather, daylight, etc.). In addition, Table 4 presents a summary of the classes available in these different datasets, to ease the comprehension of the compatibility between different models.

Table 3 – Datasets for research on urban scene Semantic Segmentation.

Dataset name	Clases	Annotated samples	Real or sim.	Video seq.	Environment/ geography	Visual conditions
Cityscapes	30	5000	Real	Yes	Germany	-
BDD100K	19	10000	Real	No	United States	-
CamVid	32	701	Real	Yes	United Kingdom	-
KITTI	28	400	Real	Yes	Germany	-
ACDC	19	4006	Real	Yes	Switzerland	Daytime
Dark Zurich	19	8779	Real	Yes	Switzerland	Daytime
Nightcity	20	4297	Real	Yes	Various cities	Daytime

Source: Made by the author.

3.2.1 CARLA-Night Dataset

The CARLA simulator was used to create a dataset of artificial images for semantic segmentation in nighttime autonomous driving (see Fig. 9), a high-fidelity, open-source

Table 4 – The various categories for which annotations are provided in different semantic segmentation datasets.

Classes	Cityscapes	BDD100K	CamVid	KITTI-v2	ACDC	Dark Zurich	Nightcity
Bicycle	x	x	x		x	x	x
Bridge	x		x				
Building	x	x	x	x	x	x	x
Bus	x	x	x		x	x	x
Car	x	x	x	x	x	x	x
Caravan	x			x			
Fence	x	x	x		x	x	x
Guard rail	x			x			
Lane marking			x				
Motorcycle	x	x	x		x	x	x
Parking	x		x				
Person	x	x	x		x	x	x
Rail track	x						
Rider	x	x			x	x	x
Road	x	x	x	x	x	x	x
Sky	x	x	x	x	x	x	x
Sidewalk	x	x	x	x	x	x	x
Terrain	x	x		x	x	x	x
Train	x	x	x	x	x	x	x
Traffic light	x	x	x	x	x	x	x
Traffic sign	x	x	x	x	x	x	x
Truck	x	x	x	x	x	x	x
Tunnel	x		x	x			
Vegetation	x	x	x	x	x	x	x
Wall	x	x	x	x	x	x	x

Source: Made by the author.

driving simulation environment. The simulation was configured to replicate nighttime driving conditions by adjusting the lighting settings, such as reducing ambient light and enabling streetlights, while avoiding additional weather effects like rain or fog. The dataset has 1000 images of nighttime scenarios in urban environments, with their corresponding

full-color semantic annotations, and the label IDs according to the CityScapes dataset. The CARLA-Night Dataset (https://www.kaggle.com/datasets/victorsillericojusto/carla-night) was released under the MIT license on Kaggle to facilitate the reproducibility of the experiments in this work and to contribute to the research community interested in training models with artificial images.

A variety of urban and rural environments were simulated to capture diverse road scenes, including streets with parked cars, buildings, vegetation, and roads. Through CARLA's built-in API, vehicles were controlled and spawned to ensure a range of dynamic objects across scenes.

For each scene, semantic segmentation labels were automatically generated, providing pixel-level annotations for essential objects like roads, cars, pedestrians, buildings, road signs, etc. This process allowed for the collection of a large, labeled dataset under consistent nighttime conditions, offering a reliable resource for training deep learning models.

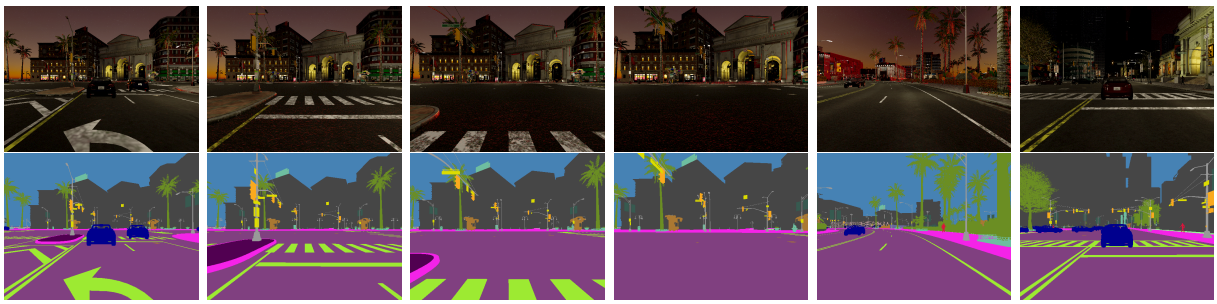


Figure 9 – Images collected using the autonomous driving simulator CARLA. The first row presents the original RGB images and their corresponding annotations in the second row.

3.3 Evaluation Metrics

After a model is trained for segmenting images according to a set of classes, it is necessary to evaluate how well it performs. Additionally, it is also important to verify that the model generalizes beyond the training dataset to new data it was not trained on. Finally, it is expected that the model makes high confidence, correct predictions, and for higher confidence thresholds to not result in many more false negatives.

Evaluation metrics are required to measure the performance of semantic segmentation models. There are many ways to quantify the similarity between the predicted (prediction) and annotated segmentation (ground truth), the most common are Precision and Recall (Sensitivity), Dice Coefficient (F1-Score), and Jaccard Index (IoU). For semantic segmentation, the evaluation unit is an individual pixel, which can be one of four categories:

- True Positive (TP): the pixel was classified correctly as a class of interest.
- True Negative (TN): the pixel was classified correctly as the background class.
- False Positive (FP): the pixel was incorrectly assigned a class of interest
- False Negative (FN): the pixel was incorrectly assigned the background class or a different class

3.3.1 F1-Score

It is also known as Dice Coefficient, and represents the harmonic mean of precision and recall. It scores the overlap between predicted segmentation and ground truth, and penalize false positives.

$$F1score = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.1}$$

3.3.2 Intersection Over Union (IoU)

It is also known as Jaccard Index, and represents the area of the intersection over union of the predicted segmentation and the ground truth.

$$IoU = \frac{TP}{TP + FP + FN} \tag{3.2}$$

3.4 Experiments

The Pytorch framework and the Segmentation Models library were the main tools to prototype the different architectures described in the literature. The Google Colaboratory (Colab) environment was the main tool for testing the initial version of the code. However, the GPU resources available for free in Colab were not enough to train and validate the models. The resources provided by the Intelligent Systems Laboratory (LASI) at the University of São Paulo (USP) helped to train the models with full-size images (1920x1080). The experiments were conducted using a system with an NVIDIA RTX 3090 GPU (24 GB VRAM, CUDA 12.2), AMD Ryzen 9 5950X CPU (3.5 GHz, 16 cores, 32 threads), 125 GB DDR4 RAM, and a 1.8 TB NVMe SSD. The operating system used was Ubuntu 20.04 with Python 3.8 and PyTorch 2.0.

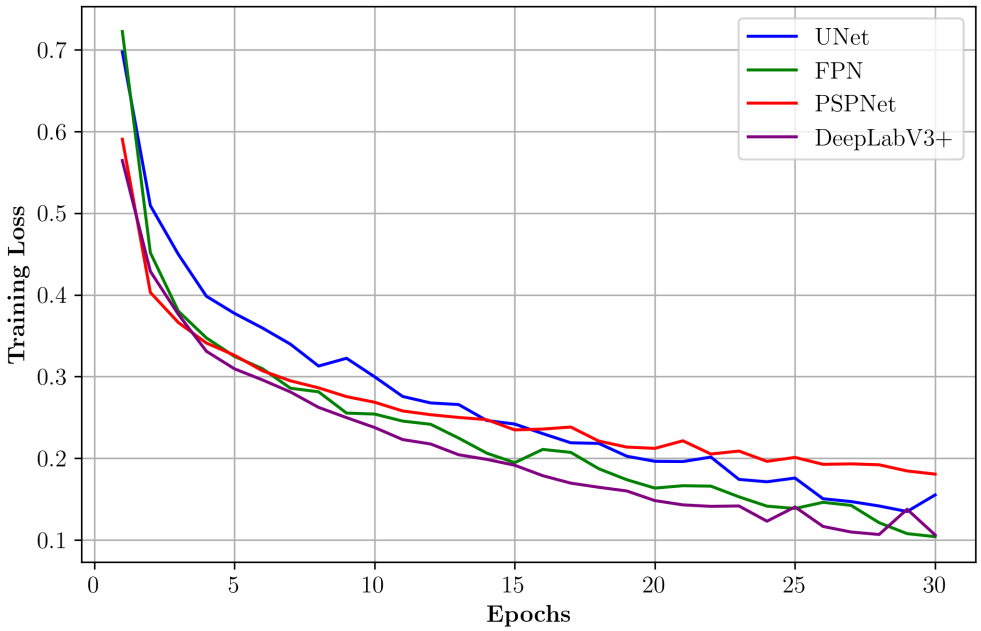
UNet, FPN, PSPNet, and DeepLabV3+ models were trained and the predicted segmentation masks were compared to define the best model in terms of generalization capabilities. Baseline models establish a minimum level of performance expectation. To ensure convergence of the models, values for epochs and batch size were adjusted experimentally taking into account common values described by the research community. The

additional testing sets were used to evaluate the generalization capability of the proposed architecture.

Cross-entropy loss was used as the primary loss function. It is widely used for pixel-wise classification tasks, making it well-suited for segmentation problems where each pixel is assigned to a specific class. The choice of cross-entropy was driven by its effectiveness in handling multiclass segmentation problems, as it computes the divergence between the predicted class probabilities and the true labels.

The training procedure for the semantic segmentation models—UNet, FPN, PSPNet, and DeepLabV3—was conducted using the Adam optimizer with a learning rate of 0.001. A batch size of 4 was used during training, while a smaller batch size of 2 was used for validation to accommodate memory constraints. Each model was trained for 30 epochs, providing sufficient iterations for the networks to learn the complex features required for nighttime autonomous driving. Although techniques to prevent overfitting, such as dropout or weight decay, were not explicitly used during the training process, the model's performance across epochs was monitored to ensure that the training remained stable and effective without excessive overfitting (see Fig. 10 and Fig. 11).

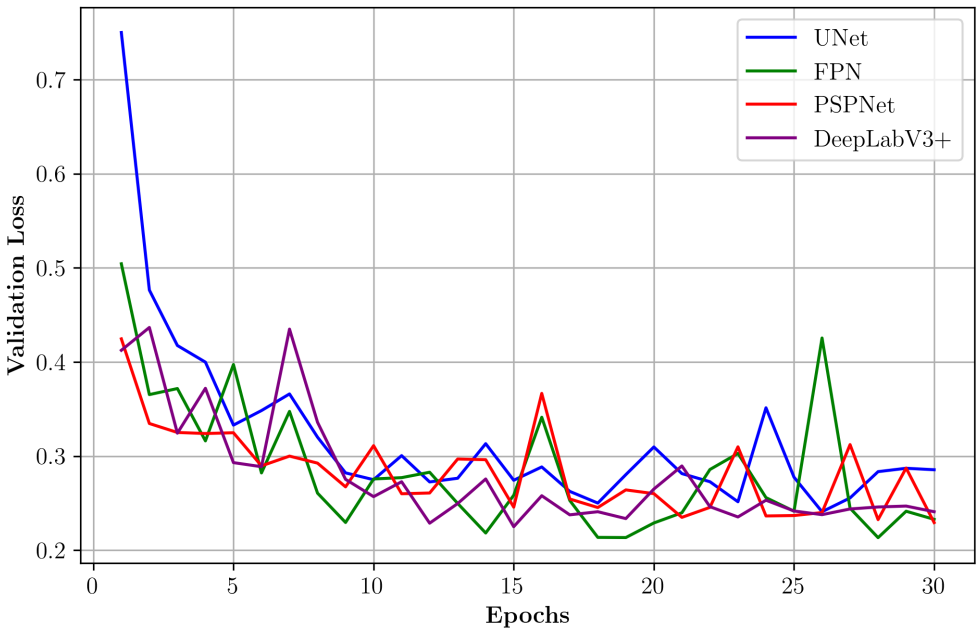
Figure 10 – Training loss over epochs for different models.



Source: Made by the author.

The backbone architecture was initialized with ResNet34 for UNet, FPN, and DeepLabV3+, and ResNet50 for PSPNet, both with weights pretrained on the ImageNet dataset. By leveraging these pretrained weights, the models could benefit from the knowledge gained from large-scale, diverse images, improving the network's ability to capture meaningful features from the nighttime data. This transfer learning approach allowed for

Figure 11 – Validation loss over epochs for different models.



Source: Made by the author.

more efficient training, as the models did not need to learn low-level features from scratch. The pretrained ResNet34 and ResNet50 backbones were fine-tuned on the nighttime hybrid (real and synthetic) dataset, resulting in improved segmentation accuracy and faster convergence, particularly for challenging low-light scenarios. The use of ImageNet-pretrained weights helped boost the model's generalization capacity, especially when detecting objects that may be less visible at night. These pre-trained models are accessible through the Segmentation Models PyTorch library[2], which provides a wide range of pre-trained networks.

For evaluating the performance of the semantic segmentation model for nighttime autonomous driving, Intersection over Union (IoU) and F1-score were employed as the primary evaluation metrics. IoU is particularly useful for segmentation tasks as it measures the overlap between predicted and ground truth areas, providing insight into the model's ability to accurately segment both frequent and less frequent classes. F1-score, which combines precision and recall, was used to further assess how well the model detects objects across different categories. Given the challenging nighttime conditions, special attention was paid to ensuring that both IoU and F1-score were calculated for small and rare objects, such as pedestrians and cyclists, which are more difficult to detect at night due to poor visibility. These metrics allowed us to monitor the model's performance not only on common elements like roads and cars but also on harder-to-detect objects, which are critical for safe autonomous driving in low-light conditions.

[2] https://segmentation-modelspytorch.readthedocs.io/en/latest/

It was expected that the use of synthetic images of nighttime environments in training would improve performance in nighttime segmentation tasks compared to baseline methods trained under normal conditions. The inclusion of these artificial images had the goal to enhance the model's ability to generalize and accurately segment scenes in low-light scenarios. Additionally, modifications made to the segmentation architectures, such as incorporating specialized layers or adaptive algorithms tailored for nighttime features, were expected to further boost the accuracy and robustness of the models. These enhancements would lead to significant improvements in handling the unique challenges posed by nighttime imagery, ultimately resulting in more reliable and precise segmentation outcomes.

4 RESULTS

A dataset with real and synthetic images was used to train four models (UNet, PSPNet, DeepLabV3+, and FPN) for semantic segmentation in nighttime autonomous driving. The CARLA simulator was used to collect 1,000 images of nighttime scenarios in urban environments, which were then merged with 506 images from the night set of the ACDC dataset, totaling 1,506 images that were split 80% for training and 20% for validation. An additional set of 500 images was used to evaluate the model performance in unseen data. This chapter presents the quantitative and qualitative results of the experiments and the overall development process.

4.1 Quantitative Analysis

4.1.1 Model Efficiency

Several factors influence how long it takes to train a model. In this project, key factors include the model's complexity (such as the number of parameters and the type of architecture), the dataset size, the batch size, the available hardware, and the choice of optimizer. Table 5 shows each model's training and inference times.

The four models used a batch size of 4 for the training set and 2 for the validation set. Moreover, UNet, FPN, and DeepLabV3+ used ResNet34 as the backbone. FPN had the shortest training time, taking 144.02 minutes, followed by DeepLabV3+ with 155.60 minutes, and UNet requiring 168.99 minutes. PSPNet, however, used ResNet50 as the backbone, which has more layers than ResNet34 to allow for more complex feature extraction, explaining why it had the longest training time of 188.68 minutes.

The test set with 500 nighttime images was used to calculate the inference time which refers to the total time taken by the model to process an input image and generate a segmentation map. This includes the time it takes to load the model, prepare the input, run the input through the network, and output the result. PSPNet was the fastest model, although it used a backbone with more layers than the other models. On the other side, UNet had the models' biggest inference time.

4.1.2 Performance Metrics

F1-Score and IoU were used to evaluate the performance of the semantic segmentation model because they provide a clear understanding of how well a model's predictions align with the ground truth, particularly for pixel-wise classification. The same validation set was used to obtain the metric values presented in Tables 6 and 7.

Table 5 – Time for training and prediction with standard deviation for inference time.

Model	Training Time (min)	Inference Time (min)	Inference Std Dev (min)
UNet	168.99	0.47	0.02
FPN	144.02	0.44	0.01
PSPNet	188.68	0.42	0.01
DeepLabV3+	155.60	0.45	0.02

Source: Made by the author.

The F1-Score results (see Table 6) indicate that the four models perform well for segmenting the road class, achieving score of 98%. This result was expected since the datasets used in the project are tailored for autonomous vehicle applications, thus all the images contain road information. Similarly, the sky class presents a score higher than 89% for each of the four models because it is a common element in all the images and has distinctive and consistent features that facilitate its identification. The building class presents different values between 87% and 91% for the studied models. The vegetation and car classes present values lower than 82% because these classes appear less than the others in the images, making it difficult to properly segment those objects. Overall, FPN ended up being the best model for nighttime semantic segmentation with a mean F1-Score value of 88.32%, and UNet was the less efficient with 84.47%.

Table 6 – F1-Score values calculated for each model.

Model	**F1-Score (%)**						mF1-Score
	road	building	vegetation	car	sky	background	
UNet	98.35	87.69	74.87	77.35	89.33	79.22	84.47
FPN	98.81	91.05	81.01	82.43	93.54	83.10	88.32
PSPNet	98.15	88.14	78.38	73.31	92.67	83.14	85.63
DeepLabV3+	98.90	87.70	80.81	81.81	92.24	83.74	87.53

Source: Made by the author.

IoU results (see Table 7) describes how well the segmentation models identify specific classes in an image, focusing on the overlap between predicted and ground truth pixels. The four models got poor performance in the segmentation of cars because they do not appear regularly in the images and most of their appearances are in a small size, which limits the segmentation task. Similarly, the performance was not good for buildings and vegetation, mainly because of their irregular form in the images. On the other hand, the models got an IoU value over 97% for the road class, and over 82% for the sky class, this good performance is expected since both classes represent a significant part of the images and present distinguishable characteristics. Considering the mean IoU, FPN had

the best performance with 81.12%, and the worst performance corresponds to the UNet model with 75.92%.

Table 7 – Intersection over Union (IoU) values calculated for each model.

Model	IoU (%)						mIoU
	road	building	vegetation	car	sky	background	
UNet	96.81	79.98	61.59	68.07	82.20	66.89	75.92
FPN	97.68	85.13	69.53	73.67	88.44	72.27	81.12
PSPNet	96.47	80.99	66.80	63.31	86.98	72.16	77.79
DeepLabV3+	97.86	81.31	69.13	72.95	86.89	73.22	80.23

Source: Made by the author.

The overall results demonstrate that training the model with a combined dataset of real images from the ACDC dataset and artificial images generated by the CARLA simulator leads to improved performance compared to training with only real images from ACDC (see Table 8). The inclusion of artificial images provided a more diverse and robust training set, enabling the model to generalize better to various nighttime driving scenarios. This hybrid approach resulted in higher accuracy across key metrics, such as mean F1-score and mean IoU, highlighting the value of supplementing real-world data with synthetic images for enhancing segmentation performance in challenging nighttime conditions.

Table 8 – Metric values calculated for different datasets.

Model	ACDC		ACDC + CARLA	
	mF1-Score(%)	mIoU(%)	mF1-Score(%)	mIoU(%)
UNet	64.39	53.22	84.47	75.92
FPN	74.95	63.16	88.32	81.12
PSPNet	73.13	60.92	85.63	77.79
DeepLabV3+	76.35	64.94	87.53	80.23

Source: Made by the author.

4.2 Qualitative Analysis

Fig. 12 and Fig. 13 present examples of predictions for each trained semantic segmentation model. It helped to make a visual inspection and have an idea of how good the results are when compared with the ground truth of the sample images.

The results for a test set (see Fig. 12) show that the performance of the model is limited when used with real images. Most of the pixels in the images represent the road

and sky classes, which makes them an easy target to be learnt by the models. However, the classes that does not appear regularly in the images, and are smaller compared to the others, represent a great challenge that impact the generalization capabilities of the models. PSPNet has the worst performance even tought it had a Resnet50 as backbone, it was not able to efectively segment the clases. The other models do not have a consistent performance, in some cases UNet and FPN were able to segment cars, but failed badly to segment vegetation.

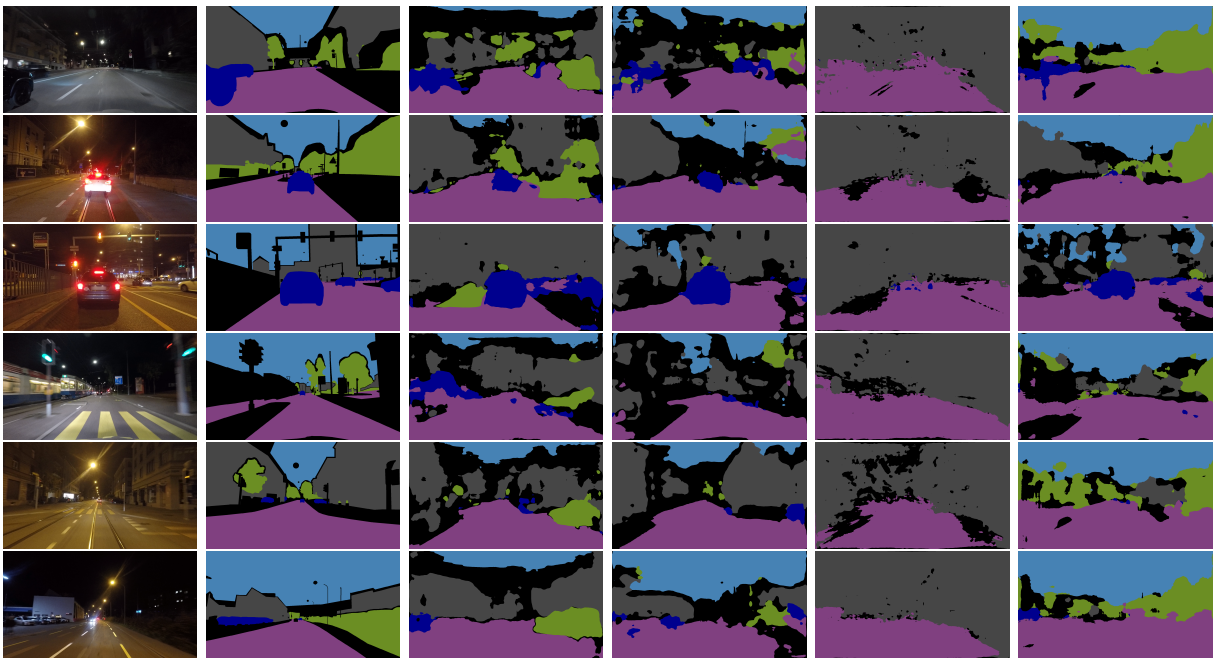


Figure 12 – Qualitative results for semantic segmentation of real images. Each row corresponds to a test sample, and from the left to the right the images correspond to the original image, its ground truth, and the four predicted masks.

When tested with a set of synthetic images collected with CARLA simulator, the models present a different performance (see Fig. 13). The four models were able to segment the road class effectively, but they had a poor performance in defining the boundaries of the building and sky classes, which appeared to be merged in most of the cases. Once again, PSPNet had the worst performance, even though it was expected to get better results with synthetic images, it had problems segmenting all the classes. UNet and FPN had performed the best compared to ground truth, but they also failed to segment the building classes.

By analyzing the images of Fig. 12 and Fig. 13, it is evident that there are some classes the models tend to misclassify. For example, the model may frequently confuse vegetation with background, or buildings with sky, particularly in low-light conditions, where object boundaries are less distinct.

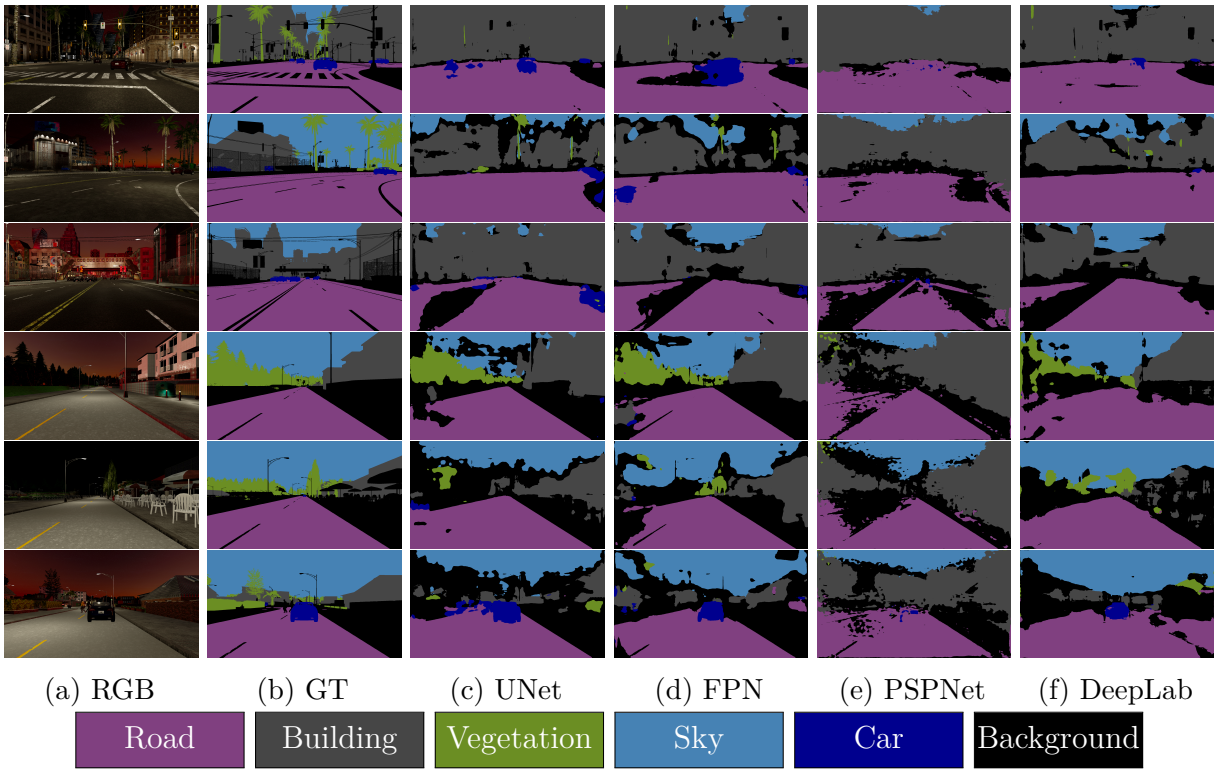


Figure 13 – Qualitative results for semantic segmentation of synthetic images. Each row corresponds to a test sample, and from the left to the right the images correspond to the original image, its ground truth, and the four predicted masks.

5 CONCLUSIONS

This project, conducted a literature review to identify state-of-the-art methods for semantic segmentation of nighttime images. UNet, FPN, PSPNet, and DeepLabV3+ models were trained with a dataset that combines real images of the ACDC dataset with synthetic images collected with the CARLA simulator. The hybrid (real and synthetic) dataset was formatted according to the 19 classes specified in the Cityscapes dataset. Still, only 6 classes were selected for prediction (road, building, vegetation, car, sky, background) which are the most representative in the ACDC dataset.

The analysis of the quantitative and qualitative results demonstrates that using images from a simulator positively impacts the training process and helps to improve the model's overall performance. Still, the improvements are not enough to achieve the level of robustness required for autonomous driving applications. F1-Score and IoU, the most common metrics for evaluating semantic segmentation models, were used to compare the selected architectures. Based on the metrics above, the best model was the FPN with 88.32% of mean F1-Score and 81.12% of mean IoU.

An additional set of real images was used to test the generalization capabilities of the model. The predicted masks show that the models can better segment some classes, like road and sky, than other classes. However, all the models failed to effectively segment irregular objects such as buildings and vegetation. The UNet and FPN models, when compared to the other trained models, were the ones that performed best in the task of semantic segmentation of some classes, as well as having greater generalization capacity.

In response to the first research question (Q1) that lead the direction of this project, the comparison of different semantic segmentation techniques revealed that certain models outperform others in terms of accuracy and robustness when segmenting key objects, such as buildings, vehicles, and vegetation in nighttime driving scenarios. Techniques like PSPNet and DeepLabV3+ showed superior performance compared to simpler architectures like U-Net, particularly in handling complex features and low lighting conditions. Models incorporating pyramid pooling and dilated convolutions proved more effective in capturing contextual information, which is crucial for segmenting objects in low-visibility environments. However, despite these improvements, all models exhibited some degree of performance degradation in the nighttime setting, indicating that segmenting objects under adverse lighting conditions remains a significant challenge.

Regarding to the second research question (Q2), the most effective modifications to existing segmentation frameworks for nighttime conditions involved the use of hybrid datasets and data augmentation. Specifically, combining real images from the ACDC

dataset with artificial images generated in the CARLA simulator significantly boosted model performance. This approach provided a more diverse range of lighting scenarios, allowing the models to generalize better in low-light conditions. This strategy, along with leveraging deeper architectures like FPN, helped mitigate the challenges posed by adverse lighting in nighttime driving, leading to more robust and accurate segmentation results.

In conclusion, none of the trained models achieve the expected performance to be part of critical systems, such as the perception module of intelligent vehicles. Furthermore, using artificial images, created by a simulator, to build upon a large dataset for training and validation does not guarantee an optimal performance of semantic segmentation models in nighttime conditions.

5.1 Future Works

There are several avenues to enhance further the performance of semantic segmentation models for nighttime autonomous driving. One potential direction is to consider more diverse and challenging datasets, capturing a broader range of nighttime scenarios and adverse weather conditions. Realistic autonomous driving simulators, such as CARLA, can help to collect images of adverse visual conditions scenarios including foggy, cloudy, rainy, and nighttime scenes. Additionally, integrating advanced data augmentation techniques, such as domain adaptation, could improve model generalization.

Another area of interest lies in exploring more sophisticated neural architectures, such as attention mechanisms or transformer-based models, which have shown promise in improving segmentation accuracy. Furthermore, real-time implementation and optimization of these models for edge devices could significantly contribute to practical deployment in autonomous vehicles. Finally, leveraging multi-modal sensor data, such as LiDAR and thermal imaging, could offer complementary information to RGB inputs, enhancing robustness in low-visibility conditions.

Finally, exploring multimodal Large Language Models (LLMs) that can generate textual descriptions of images could be a valuable area of research. Although these models typically have slower inference times, they offer the potential to combine visual and language understanding, allowing the system to interpret and describe complex nighttime driving scenes. This could contribute to more robust perception modules in intelligent vehicles.

REFERENCES

ANOOSHEH, A. *et al.* Night-to-day image translation for retrieval-based localization. *In*: **2019 International Conference on Robotics and Automation (ICRA)**. [*S.l.: s.n.*], 2019. p. 5958–5964.

BASLAMISLI, A. S. *et al.* Joint learning of intrinsic images and semantic segmentation. *In*: **Proceedings of the European Conference on Computer Vision (ECCV)**. [*S.l.: s.n.*], 2018.

BROSTOW, G. J.; FAUQUEUR, J.; CIPOLLA, R. Semantic object classes in video: A high-definition ground truth database. **Pattern recognition letters**, Elsevier, v. 30, n. 2, p. 88–97, 2009.

BRÜGGEMANN, D. *et al.* Refign: Align and refine for adaptation of semantic segmentation to adverse conditions. *In*: **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**. [*S.l.: s.n.*], 2023. p. 3174–3184.

CARRERA-RIVERA, A. *et al.* How-to conduct a systematic literature review: A quick guide for computer science research. **MethodsX**, v. 9, p. 101895, 2022. ISSN 2215-0161. Available at: https://www.sciencedirect.com/science/article/pii/S2215016122002746.

CHEN, L.-C. *et al.* Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 40, n. 4, p. 834–848, 2018.

CHENG, J. *et al.* Night-time semantic segmentation with unsupervised learning and cross attention. *In*: . [*S.l.: s.n.*], 2022. v. 189. Cited by: 0. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85162198510&partnerID=40&md5=0a63becd7cc1867067a524d0cad9bd2c.

CORDTS, M. *et al.* The cityscapes dataset for semantic urban scene understanding. *In*: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [*S.l.: s.n.*], 2016. p. 3213–3223.

CRESWELL, A. *et al.* Generative adversarial networks: An overview. **IEEE Signal Processing Magazine**, v. 35, n. 1, p. 53–65, 2018.

DAI, D.; GOOL, L. V. Dark model adaptation: Semantic image segmentation from daytime to nighttime. *In*: **2018 21st International Conference on Intelligent Transportation Systems (ITSC)**. [*S.l.: s.n.*], 2018. p. 3819–3824.

DAI, D.; GOOL, L. V. Dark model adaptation: Semantic image segmentation from daytime to nighttime. *In*: **2018 21st International Conference on Intelligent Transportation Systems (ITSC)**. [*S.l.: s.n.*], 2018. p. 3819–3824.

DENG, X. *et al.* Nightlab: A dual-level architecture with hardness detection for segmentation at night. *In*: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S.l.: s.n.*], 2022. p. 16938–16948.

DING, F.; LI, J.; TIAN, W. Dual-level consistency learning for unsupervised domain adaptive night-time semantic segmentation. *In*: **2023 IEEE International Conference on Multimedia and Expo (ICME)**. [*S.l.: s.n.*], 2023. p. 420–425. ISSN 1945-788X.

DOSOVITSKIY, A. *et al.* CARLA: An open urban driving simulator. *In*: **Proceedings of the 1st Annual Conference on Robot Learning**. [*S.l.: s.n.*], 2017. p. 1–16.

GAO, H. *et al.* Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. *In*: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S.l.: s.n.*], 2022. p. 9913–9923.

GEIGER, A.; LENZ, P.; URTASUN, R. Are we ready for autonomous driving? the kitti vision benchmark suite. *In*: IEEE. **2012 IEEE conference on computer vision and pattern recognition**. [*S.l.: s.n.*], 2012. p. 3354–3361.

LEE, H. *et al.* Gps-glass: Learning nighttime semantic segmentation using daytime video and gps data. *In*: **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops**. [*S.l.: s.n.*], 2023. p. 4001–4010.

LI, H.; LIU, C.; YANG, Y. Layernet: A one-step layered network for semantic segmentation at night. **IEEE Computer Graphics and Applications**, v. 43, n. 6, p. 9–21, Nov 2023. ISSN 1558-1756.

LIN, T.-Y. *et al.* Feature pyramid networks for object detection. *In*: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [*S.l.: s.n.*], 2017. p. 2117–2125.

LIU, W. *et al.* Improving nighttime driving-scene segmentation via dual image-adaptive learnable filters. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 33, n. 10, p. 5855–5867, 2023.

NAG, S.; ADAK, S.; DAS, S. What's there in the dark. *In*: **2019 IEEE International Conference on Image Processing (ICIP)**. [*S.l.: s.n.*], 2019. p. 2996–3000. ISSN 2381-8549.

PETTICREW, M.; ROBERTS, H. **Systematic reviews in the social sciences: A practical guide**. [*S.l.: s.n.*]: John Wiley & Sons, 2008.

ROMERA, E. *et al.* Bridging the day and night domain gap for semantic segmentation. *In*: **2019 IEEE Intelligent Vehicles Symposium (IV)**. [*S.l.: s.n.*], 2019. p. 1312–1318.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. *In*: SPRINGER. **Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18**. [*S.l.: s.n.*], 2015. p. 234–241.

SAKARIDIS, C.; DAI, D.; GOOL, L. V. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *In*: **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**. [*S.l.: s.n.*], 2019.

SAKARIDIS, C.; DAI, D.; GOOL, L. V. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. *In*: **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**. [*S.l.: s.n.*], 2021. p. 10765–10775.

SAKARIDIS, C.; DAI, D.; GOOL, L. V. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 44, n. 6, p. 3139–3153, 2022.

SCHUTERA, M. *et al.* Night-to-day: Online image-to-image translation for object detection within autonomous driving by night. **IEEE Transactions on Intelligent Vehicles**, v. 6, n. 3, p. 480–489, 2021.

SCHWONBERG, M. *et al.* Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving. **IEEE Access**, v. 11, p. 54296–54336, 2023.

SELLAT, Q.; BISOY, S. K.; PRIYADARSHINI, R. Semantic segmentation for self-driving cars using deep learning: a survey. *In*: MISHRA, S. *et al.* (ed.). **Cognitive Big Data Intelligence with a Metaheuristic Approach**. Academic Press, 2022, (Cognitive Data Science in Sustainable Computing). p. 211–238. ISBN 978-0-323-85117-6. Available at: https://www.sciencedirect.com/science/article/pii/B9780323851176000029.

SIAM, M. *et al.* A comparative study of real-time semantic segmentation for autonomous driving. *In*: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops**. [*S.l.: s.n.*], 2018.

SONG, C. *et al.* Nighttime road scene parsing by unsupervised domain adaptation. **IEEE Transactions on Intelligent Transportation Systems**, v. 23, n. 4, p. 3244–3255, 2022.

SUN, L. *et al.* See clearer at night: Towards robust nighttime semantic segmentation through day-night image conversion. *In*: . [*S.l.: s.n.*], 2019. v. 11169. Cited by: 40; All Open Access, Green Open Access. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077332282&doi=10.1117%2f12.2532477&partnerID=40&md5=8d7a5f5a35dac1b8aa1e106642888f98.

TAN, X. *et al.* Night-time scene parsing with a large real dataset. **IEEE Transactions on Image Processing**, v. 30, p. 9085–9098, 2021.

VALADA, A. *et al.* Adapnet: Adaptive semantic segmentation in adverse environmental conditions. *In*: **2017 IEEE International Conference on Robotics and Automation (ICRA)**. [*S.l.: s.n.*], 2017. p. 4644–4651.

WANG, H. *et al.* Sfnet-n: An improved sfnet algorithm for semantic segmentation of low-light autonomous driving road scenes. **IEEE Transactions on Intelligent Transportation Systems**, v. 23, n. 11, p. 21405–21417, Nov 2022. ISSN 1558-0016.

WEI, Z. *et al.* Disentangle then parse: Night-time semantic segmentation with illumination disentanglement. *In*: **2023 IEEE/CVF International Conference on Computer Vision (ICCV)**. [*S.l.: s.n.*], 2023. p. 21536–21546. ISSN 2380-7504.

WU, X. *et al.* Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. *In*: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S.l.: s.n.*], 2021. p. 15769–15778.

WU, X. *et al.* A one-stage domain adaptation network with image alignment for unsupervised nighttime semantic segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 45, n. 1, p. 58–72, Jan 2023. ISSN 1939-3539.

XIE, Z. *et al.* Boosting night-time scene parsing with learnable frequency. **IEEE Transactions on Image Processing**, v. 32, p. 2386–2398, 2023.

XU, Q. *et al.* Cdada: A curriculum domain adaptation for nighttime semantic segmentation. *In*: **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops**. [*S.l.: s.n.*], 2021. p. 2962–2971.

YANG, G. *et al.* Bi-mix: Bidirectional mixing for domain adaptive nighttime semantic segmentation. **arXiv preprint arXiv:2111.10339**, 2021.

YANG, X.; HAN, J.; LIU, C. A semantic segmentation scheme for night driving improved by irregular convolution. **FRONTIERS IN NEUROROBOTICS**, v. 17, JUN 12 2023. ISSN 1662-5218.

YU, F. *et al.* Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *In*: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [*S.l.: s.n.*], 2020. p. 2636–2645.

ZHAO, H. *et al.* Pyramid scene parsing network. *In*: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S.l.: s.n.*], 2017.