Advancements in Road Lane Mapping: Comparative Analysis of Deep Learning-based Semantic Segmentation Methods Using Aerial Imagery

by

Xuanchen (Willow) Liu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Geography

Waterloo, Ontario, Canada, 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The rapid advancement of autonomous vehicles (AVs) underscores the necessity for high-definition (HD) maps, with road lane information being crucial for their navigation. The widespread use of Earth observation data, including aerial imagery, provides invaluable resources for constructing these maps. However, to fully exploit the potential of aerial imagery for HD road map creation, it is essential to leverage the capabilities of artificial intelligence (AI) and deep learning technologies. Conversely, the domain of remote sensing has not yet fully explored the development of specialized models for road lane extraction, an area where the field of computer vision has made significant progress with the introduction of advanced semantic segmentation models.

This research undertakes a comprehensive comparative analysis of twelve deep learning-based semantic segmentation models, specifically to measure their skill in road lane marking extraction, with a special emphasis on a novel dataset characterized by partially labeled instances. This investigation aims to examine the models' performance when applied to scenarios with minimal labeled data, examining their efficiency, accuracy, and ability to adapt under conditions of limited annotation and transfer learning.

The outcome of this study highlights the distinct advantage of Transformer-based models over their Convolutional Neural Network (CNN) counterparts in the context of extracting road lanes from aerial imagery. Remarkably, within the state-of-the-art models, such as Segmenting Transformers (SegFormer), Shifted Window (Swin) Transformer, and Twins Scaled Vision Transformer (Twins-SVT) exhibit superior performance. The empirical results on the Waterloo Urban Scene dataset mark substantial progress, with mean Intersection over Union (IoU) scores ranging from 33.56% to 76.11%, precision from 64.33% to 77.44%, recall from 66.0% to 98.96%, and F1 scores from 44.34% to 85.35%. These findings underscore the benefits of model pretraining and the distinctive attributes of the dataset in strengthening the effectiveness of models for HD road map development, announcing new possibilities in the advancement of autonomous vehicle navigation systems.

Acknowledgments

I extend my deepest gratitude to my parents, Da Liu and Jingbo Hu. From my earliest memories, you have treated my opinions and choices with the utmost respect, as if I were already an adult. You granted me the freedom to be myself, supporting my dreams with sacrifices that I am only beginning to comprehend. You've never demanded anything from me, only asking me to be a kind, healthy, and happy person. I've always known your unwavering support and acceptance, feeling proud of whatever path I chose to follow. I want to express my profound love for both of you and my heartfelt thanks for everything you've done for me. Simply being your child is a blessing I cherish deeply.

I express my sincere gratitude to Professor Jonathan Li, my supervisor, for his consistent guidance, thorough teaching, and steadfast support throughout my study. His mentorship extended beyond academic excellence, imparting lessons on leadership and genuinely caring about my well-being. His respect for each student and the nurturing environment he creates have confirmed the wisdom of my decision to select such an outstanding supervisor two years ago.

I would like to extend my gratitude to my committee members, Professor Michael A. Chapman and Professor Linlin Xu, for their participation in my oral defense and the valuable insights they provided. Their expertise and advice have been greatly appreciated and played a crucial role in the refinement of my work.

Special thanks to Dr. Frederick E. Clark, my former instructor at the University of Alberta. Getting to know him personally has been a privilege. Over the past five years, Dr. Clark has transcended the role of an instructor to become not just a mentor but also one of my closest friends. I have been able to discuss anything with him, always receiving responses filled with wisdom. The welcoming nature of both Dr. Clark and Sib has made Edmonton feel like a second home to me.

I'm grateful to a special person in my life, Kyle Gao. On the day I joined our GIM group meeting, I told my mom that everyone was great except for one annoying guy, who

unexpectedly became my boyfriend in the later days. It's not always easy for an INTP to grasp why Willow (ENFJ) is upset again, but you've made significant changes to yourself for the sake of my happiness, changes I know were not easy for you. Your care for me and our cats, along with your amazing culinary skills, have brought joy (and extra pounds) to our little family. In a book by Maugham that I've read, "Of Human Bondage," there's a line I particularly love: "It's no use crying over spilt milk, because all of the forces of the universe were bent on spilling it" Standing at the crossroads of fate and choice, indeed, we encounter various incidents and possibilities of spilling milk. However, because it's you, I'm willing to pour new milk into the cup countless times alongside you.

I owe a deep debt of gratitude to my friends in Alberta, Shuxin Qiao and Zepeng Xiao. Since the first year of my undergraduate studies, they have stood by me as my best friends. Initially, my understanding of many science courses was limited, but they patiently assumed the role of my tutors without any conditions. When I began my master's degree and faced challenges with programming, they became my tutors again, guiding me through the basics, including teaching me how to execute my first print “Hello World”. I am deeply grateful for their steadfast support and encouragement of my dreams. It is my sincere hope that this friendship will endure for a lifetime.

I'm also grateful to my friend Dr. Hongjie He for his patience and kind guidance since I joined the group. He has been like a big brother within our circle of friends, always possessing a generous heart and a readiness to help others. Additionally, my heartfelt thanks go to my other great friends in Waterloo—Kristie Hu, Jing Du, Dening Lu, and Jirui Hu. When I first arrived in Waterloo, knowing few people, and having even fewer friends, you all entered my life and changed it for the better. As we move forward, potentially to different corners of the world after graduation, I cherish the hope that our friendship endure, transcending both time and distance.

Lastly, I'm thankful for the financial support from the Caivan Future Cities Graduate Scholarship. To everyone who has offered warmth, help, and friendship, though I cannot name all, your kindness is deeply remembered. I haven’t taken a single step alone. Thanks for walking with me.

Dedication

Dedicated to my mother, Jingbo Hu, and my father, Da Liu, for your unwavering support, encouragement, and love.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADAS	Advanced Driver Assistance Systems
AFNB	Asymmetric Fusion Non-local Block
AI	Artificial Intelligence
ANN	Asymmetric Non-local Neural Networks
APNB	Asymmetric Pyramid Non-local Block
ASPP	Atrous Spatial Pyramid Pooling
aAcc	All Pixel Accuracy
AVs	Autonomous Vehicles
BDD100	Berkeley Deep Drive 100K
CalTech	California Institute of Technology
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CV	Computer Vision
DSLR	Digital Single-Lens Reflex
DWTs	Discrete Wavelet Transforms
FC-DenseNet	Fully Convolutional Dense Network
FCN	Fully Convolutional Networks
FCNN	Fully Convolutional Neural Networks

FN	False Negatives
FP	False Positives
GANs	Generative Adversarial Networks
GCLNet	Geometric Constrained Network
GIM Lab	Geospatial Intelligence and Mapping Laboratory
GIS	Geographic Information Systems
GPU	Graphics Processing Units
HD	High Definition
IoU	Intersection over Union
ISPRS	International Society for Photogrammetry and Remote Sensing
JPU	Joint Pyramid Upsampling
KITTI	Karlsruhe Institute of Technology & Toyota Technological Institute
LaneIoU	Lane Intersection over Union
LiDAR	Light Detection and Ranging
LMD	Lane Marking Detector
LR-ASPP	Lite Reduced Atrous Spatial Pyramid Pooling
mAcc	Mean Accuracy
MFPN	Multi-scale Feature Pyramid Network
mIoU	Mean Intersection over Union

MLP	Multi-Layer Perceptron
NAS	Network Architecture Search
PSPNet	Pyramid Scene Parsing Network
RCNet	Road Classification Network
RGB	Red Green Blue
SegFormer	Segmenting Transformers
SGD	Stochastic Gradient Descent
SVMs	Support Vector Machines
Swin	Shifted Window Transformer
TN	True Negatives
TP	True Positives
Twins-PCPVT	Twins Pooled Convolutional Pyramid Vision Transformer
Twins-SVT	Twins Scaled Vision Transformer
UAVs	Unmanned Aerial Vehicles
ViT	Vision Transformer

Chapter 1 Introduction

1.1 Background and Motivation

The rapid development of autonomous driving technologies heralds a new era in transportation, underpinned by the critical need for high-definition (HD) maps (Azimi et al., 2019a). These maps, rich in detail and precision, are pivotal for the safe and efficient operation of autonomous vehicles (AVs), offering an unparalleled level of navigational accuracy far beyond traditional navigation tools. HD maps encompass a comprehensive array of environmental data, including the traffic signals, road features, and especially precise lane markings, which are particularly important (Zhou et al, 2021). This detailed representation is essential for enabling AVs to understand their surroundings, make informed decisions, and navigate safely.

The surge in the availability of various earth observation data types has provided a powerful resource for the creation of these essential HD road maps (Azimi et al., 2019a). Data sources such as the Light Detection and Ranging (LiDAR) point clouds and satellite images offer valuable insights, yet they come with their limitations (Zhou et al, 2021). On the one hand, LiDAR, for example, may suffer from uneven point distribution and various densities, leading to a lack of uniform detail in various regions and potential information gaps (Azimi et al., 2019a). On the other hand, satellite imagery often faces challenges like lower spatial resolution, which hinders the capture of subtle details, and its infrequent revisit may fail to reflect the most current changes on the roads (Azimi et al., 2019b).

In contrast, aerial imagery, such as those from drones or unmanned aerial vehicles (UAVs), distinguishes itself with higher resolution and clarity, along with more adaptable data collection methods, significantly outperforming satellite imagery and LiDAR (Azimi et al., 2019a). Among its many advantages is the high-resolution capability that allows for capturing intricate details such as lane markings. Additionally, there is flexibility in data collection, which permits targeted gathering of information at the best possible times and from the most effective angles to ensure optimal lighting and weather conditions. These examples illustrate how aerial imagery overcomes the limitations of LiDAR and satellite imagery, making it exceptionally suitable for creating HD maps.

However, the full potential of aerial imagery in creating HD road maps cannot be realized without the aid of Artificial Intelligence (AI) or deep learning algorithms (Rehman et al., 2023). These algorithms play a crucial role in processing and interpreting the complex data contained within aerial images, enabling the automated extraction of essential elements such as lane markings. The complexity and variability of road environments demand sophisticated deep learning models that can accurately identify and classify road features across diverse conditions, underscoring the necessity of integrating deep learning algorithms into the HD map creation process.

Despite the advantages offered by deep learning, existing methods for the automated extraction of lane markings in aerial images are fraught with challenges (Azimi et al., 2019a). These include difficulties in dealing with the high variability of road conditions, the complexity of interpreting dense urban environments, the need for high levels of accuracy in feature extraction, and the added complexity posed by the narrow width of road lanes. The limitations of current deep learning approaches highlight the need for ongoing research and development to refine and enhance these technologies for better performance in the specific context of HD map creation.

In the field of remote sensing, the methodologies available for automated lane markings extraction are limited and often not optimized for the unique challenges posed by aerial imagery. Conversely, the field of computer vision boasts a wealth of advanced semantic segmentation models that have shown great promise in various applications. However, these remain underexplored for their potential in remote sensing tasks such as lane markings extraction. This discrepancy between available technologies and their application in enhancing HD maps points to a significant gap in current research efforts.

Addressing this gap requires a focused comparative study of semantic segmentation models derived from the field of computer vision, specifically tailored to the task of extracting road lane markings from aerial imagery. Such a study is imperative to identify the most effective models that can overcome the existing limitations and significantly improve the accuracy and efficiency of HD map creation. This research endeavor seeks to fill this void by systematically evaluating these models, thereby contributing to the

advancement of autonomous driving technologies through the development of more detailed and reliable HD road maps. Through this comparative analysis, we aim to establish a benchmark for future innovations in the field, ensuring that AVs navigate with unparalleled precision and safety.

1.2 Objectives of the Study

This study is driven by the principal aim of performing an exhaustive comparative experimental analysis of twelve deep learning-based semantic segmentation models to thoroughly evaluate their effectiveness in the task of road lane extraction.

- The first objective is to study the models' performance and relevance, offering insights into their performance, efficiency, and parameters.

- The second objective is to test the capabilities of state-of-the-art deep learning - based segmentation models on a newly created dataset with partial labels under few-shot and transfer learning conditions, comparing efficiency, accuracy, and ability to learn and adapt to datasets with only a small portion of labeled data (e.g., 1%).

1.3 Structure of the Thesis

The thesis is structured into the following five chapters:

Chapter 1 sets the study's groundwork by outlining motivations, identifying research gaps, and defining objectives, thus framing the inquiry.

Chapter 2 offers a concise review of semantic segmentation in road lane detection, structured into four main sections: the advancement of aerial view road detection techniques, dataset evaluation for model training, the development of segmentation models with an emphasis on deep learning, and the shift from conventional to deep learning-based lane detection methods.

Chapter 3 delves into the methodology with a structured exposition, starting with an overview of the general workflow. It then elaborates on the dataset used for analysis, dataset preprocessing, a thorough examination of each model's architecture and features, exploration of transfer learning strategies, and the selected evaluation metrics.

Chapter 4 begins with an overview of the training environment, followed by a detailed discussion on model training, encompassing training procedures, parameter configurations, optimization process, data partitioning, and fine-tuning processes. The chapter showcases both the quantitative assessments of model performance and qualitative visual representations of findings across two separate datasets, concluding with a discussion section.

Chapter 5 concludes the thesis and summarizes the significant insights gained from the study, highlights the limitations and challenges encountered, and proposes the directions for future research.

Chapter 2 Related Works

This chapter provides a focused review of semantic segmentation for road lane detection across four sections. Section 2.1 examines road lane marking extraction methods with its definition and characteristics, comparing ground-level detail to aerial coverage, and evaluating manual, machine learning, and deep learning approaches for detection and analysis. Section 2.2 evaluates available aerial view datasets for model training. Section 2.3 explores the evolution of semantic segmentation models, emphasizing the impact of deep learning advancements. Lastly, Section 2.4 examines the progression of aerial view road lane detection methods, transitioning from traditional to deep learning-based approaches.

2.1 Road Lane Marking Extraction Methods and Perspectives

Road lanes are crucial elements within road infrastructure, delineating paths for vehicle movement and promoting smooth traffic circulation while also communicating key information about traffic regulations. These lanes are marked by a variety of symbols on the road surface, which can vary greatly in shape, size, length, and color (Gao et al., 2006), reflecting the diversity of traffic rules and cultural norms across countries (Shinar et al., 2003). From dashed lines to solid lines, and from arrows to pedestrian crossings, each symbol serves a specific role. Geometrically, lane markings are crafted with clear boundaries to ensure they are easily distinguishable (Liu et al., 2012). The variety in lane markings, with their unique shapes and the broad spectrum of colors (ranging from the standard white and yellow to more distinctive hues in certain areas), highlights the critical need for precise and accurate lane extraction for a wide range of applications, especially autonomous driving technologies (Azimi et al., 2019a).

Transitioning from theoretical significance to practical applications of road lane information, ground-level observation emerges as a primary method of inquiry. It's common for people to see and recognize lane markings at ground level as guides for navigation and safety on the road. At this ground perspective, traditional in-site surveying and mapping techniques become important, offering a lens to view and capture the

intricate details of lane markings with high resolution. Yet, this approach has its downsides: its reach is limited by a narrow field of view, and lane markings can be hidden by obstacles or vehicles, making the process time-consuming and expensive (Wu et al., 2014). Additionally, it can also be influenced by bad weather conditions, further complicating the accuracy and efficiency of the system. While in-site surveying offers unmatched detail and precision in data collection, its applicability and efficiency are greatly reduced by these challenges.

Alternatively, aerial imagery offers a means to collect road lane information from above. Though the resolution might be slightly lower than that achieved through in-site surveying, the advantage of aerial imagery lies in its extensive coverage and the ability to capture spectral features and information across the red, green, and blue (RGB) bands, which are sufficient for extracting road information (Azimi et al., 2019b). Aerial views provide a comprehensive perspective of road layouts, revealing patterns and alignments that ground-level surveying cannot match. Within aerial imagery, three primary methods emerge for extracting road lane markings.

The first method is manual interpretation and editing, using tools like the Geographic Information Systems (GIS) to carefully label road lane markings. This technique is known for its accuracy, drawing on the detailed analysis of skilled professionals (Long et al., 2021). However, its drawbacks are noteworthy: it is labor-intensive, costly, and impractical for large-scale projects due to the requirement for expert knowledge and significant time and financial investment. Following manual techniques, traditional machine learning methods offer an alternative, employing algorithms to recognize patterns and classify road features. Despite their utility, these methods also necessitate expert knowledge for feature selection and algorithm tuning, and they struggle with generalization across varied environments, limiting their scalability and adaptability.

In contrast, deep learning emerges as a powerful method for the automated extraction of road lane markings from aerial photos (Long et al., 2021). This approach reduces the dependency on manual intervention and expertly crafted features, showcasing strong generalization across various environments. These models require comparatively lower

computational power, offering speed and efficiency that enable the processing of large datasets. Deep learning's ability to learn from vast amounts of data and its adaptability to new, unseen environments make it an invaluable tool for contemporary and future applications in road lane extraction, setting a new standard for accuracy and efficiency in this critical task.

2.2 Datasets for Lane Segmentation

In the context of enhancing lane segmentation model performance, the selection of training datasets plays a pivotal role. For models to accurately capture lane markings, datasets must offer high-resolution images that reveal detailed features of lanes, provide annotations for a variety of lane types, be readily accessible for research purposes, and be large enough to support effective model training and validation (Long et al., 2021).

Before selecting a dataset for analysis, it is crucial to confirm that the aerial imagery offers a resolution that allows for clear visibility of road lane markings. To determine the suitable resolution, the Nyquist-Shannon sampling theorem needs to be applied (Por et al., 2019). This theorem mandates that the sampling rate must be at least twice the size of the smallest detail that needs to be accurately reconstructed in the data. Therefore, identifying the smallest feature of interest in the imagery is essential. For example, if the smallest discernible feature in a dataset is a zebra crossing stripe measuring 30cm in width, the resolution must be at least 15cm to capture this feature effectively. Such a resolution is necessary to accurately reconstruct the image or signal from the sampled data. Choosing the right resolution is vital to ensure the dataset includes enough detail to accurately represent real-world features, thus maintaining data integrity and ensuring that important information is not lost during data processing.

Currently, the collection of HD map data predominantly utilizes mobile mapping systems equipped with sensor-laden vehicles, including LiDAR and digital cameras. Datasets obtained at ground level are abundantly accessible, serving the needs of analyses based on the perspective of vehicle-mounted cameras. Notable examples of such datasets include the TuSimple Lane Detection Challenge Dataset (Yoo et al., 2020), the Road and

Lane Dataset from the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) (Geiger et al., 2013), the California Institute of Technology (Caltech) Lanes Dataset (Chao et al., 2019), and the Berkeley Deep Drive 100K (BDD100K) dataset (Yu et al., 2020). Nonetheless, these datasets are mainly designed to support models that adopt a driver's viewpoint, lacking the comprehensive coverage essential for thorough traffic management and analysis. This approach faces challenges in data analysis due to the restricted field of view of the sensors and physical obstructions. Additionally, the task of mapping extensive urban areas using this technique proves to be both time-consuming and demanding in terms of resources.

In contrast, aerial image datasets play a crucial role in providing a comprehensive perspective on traffic patterns, making them invaluable for the implementation of large-scale traffic management systems. As outlined in Section 2.1, despite the availability of numerous aerial imagery sources that could potentially meet the outlined requirements, there exists a notable scarcity of aerial datasets that offer detailed annotations of lane markings. The process of annotating a substantial volume of datasets to the extent required for effective deep learning training entails significant expenditure of time and financial resources. Furthermore, while numerous aerial imagery benchmarks featuring annotations may appear appropriate for tasks like lane extraction within the realm of semantic segmentation, as investigated by the Long et al. in their survey about benchmark dataset for aerial image, the majority do not meet the precise requirements of this undertaking. For instance, datasets such as the International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam and Vaihingen (Rottensteiner et al., 2014) provide high-resolution images yet fall short in delivering lane-specific information. Similarly, the Massachusetts Roads (Azimi et al., 2019b) and SpaceNet datasets (Van Etten et al., 2021) present urban annotations but lack the requisite granularity for precise lane segmentation.

The SkyScapes dataset, an aerial image resource, stands out with its 13 cm resolution images coupled with detailed annotations that encompass a wide variety of lane markings across 12 classes, encompassing both urban and suburban environments (Azimi et al., 2019b). Furthermore, it is freely accessible, making it an invaluable asset for the accurate detection and classification of road lanes from an aerial perspective.

This review underscores a noticeable gap in detailed aerial-level datasets compared to ground-level offerings, emphasizing the SkyScapes dataset's role. Its detailed annotations and high-resolution imagery position it as a key resource for developing sophisticated multi-class road lane detection models. Accordingly, this thesis will focus on leveraging the SkyScapes dataset to devise and refine road lane detection methodologies, setting a robust foundation for subsequent model development discussions.

2.3 Existing Semantic Segmentation Models

In the field of computer vision, semantic segmentation performs pixel-wise labeling and is well suited for complex real-world tasks. This technique divides an image into distinct segments corresponding to various objects or regions. Deep learning has significantly transformed this area, moving away from traditional methods based on clustering and contours (Weinland et al., 2011; Sonka et al., 2013). By enabling accurate segmentation of unknown images at a pixel level, deep learning algorithms not only enhance precision but also provide a richer understanding of visual scenes (Guo et al., 2018).

There are two primary categories of existing semantic segmentation models: Convolutional Neural Network (CNN)-based models and Transformer-based models. These models represent distinct approaches in the field, each leveraging unique architectures to analyze and interpret image data effectively.

2.3.1 CNN-based Models

Following the advancements brought by CNN in semantic segmentation, the field witnessed a significant shift with the advent of Encoder-Decoder architecture. The Encoder-Decoder model, initially conceptualized for neural machine translation (Sutskever et al., 2014) to map sequences from one domain to another, offered a novel approach to semantic segmentation (Badrinarayanan et al., 2017; Ronneberger et al., 2015). It consists of two parts: an encoder that compresses the input into a feature-rich representation, and a decoder that reconstructs the target output from this representation

(Sutskever et al., 2014).

The selection of CNN-based models in semantic segmentation is based on their contributions to the field, highlighted by innovative structures and methods that have significantly shaped further research. Models are selected for historical importance, unique architecture, and influence on later studies, and performance, offering a benchmark with a blend of models with exceptional performance and models with historically significant impact on the development of semantic segmentation. These models are briefly described in this section:

FCN (Fully Convolutional Network): This model modifies CNNs to process full images directly. By substituting fully connected layers with convolutional layers, it outputs spatial maps suitable for inputs of any size, a significant advancement over traditional CNNs that demand fixed-size inputs (Long et al., 2015). This "fully convolutional" design enables adaptable and size-invariant segmentation.

FastFCN (Fast Fully Convolutional Network): This model introduces critical innovations over conventional CNN approaches (Wu et al., 2019). Its main advancement is the Joint Pyramid Upsampling (JPU) module, which efficiently merges multi-scale features, bypassing the extensive pooling and upsampling layers typical of CNNs and FCNs. This allows for high-resolution semantic segmentation with reduced computational demand, facilitating quicker processing speeds than traditional models.

U-Net: This model leverages a convolutional neural network (CNN) in an encoder-decoder framework (Ronneberger et al., 2015). Its notable feature, "skip connections," concatenates feature maps from the encoding path with the decoder's upsampled output, enhancing detail localization. This structure effectively maintains spatial context, addressing the common challenge of detail loss in deeper layers found in traditional CNN segmentation approaches.

MobileNetV3: This model innovates within CNN architectures, optimizing for mobile device constraints without compromising performance (Howard et al., 2019). By

integrating Hardware-Aware Network Architecture Search (NAS) and the NetAdapt algorithm, it fine-tunes its structure for optimal functionality on mobile CPUs. MobileNetV3 introduces architectural enhancements, such as the Lite Reduced Atrous Spatial Pyramid Pooling (LR-ASPP), to improve semantic segmentation efficiency. Additionally, a streamlined segmentation decoder is incorporated to boost performance in dense pixel prediction tasks, ensuring computational efficiency is maintained.

ANN (Asymmetric Non-local Neural Networks): This model is a CNN-based framework that introduces two innovations: the Asymmetric Pyramid Non-local Block (APNB) and the Asymmetric Fusion Non-local Block (AFNB) (Zhu et al., 2019). APNB reduces computation and memory usage by applying pyramid sampling to non-local blocks, maintaining performance while addressing the high resource demands of traditional non-local operations. AFNB improves segmentation by fusing multi-level features and addressing long-range dependencies, overcoming typical CNN limitations in capturing these dependencies efficiently.

DeepLabV3: This model marks a notable development in semantic segmentation, integrating atrous convolutions and Atrous Spatial Pyramid Pooling (ASPP) within a CNN architecture (Chen et al., 2017a). Atrous convolutions are employed to broaden the receptive field, preserving the resolution of feature maps, and enhancing the model's ability to assimilate expansive contextual details without downsampling. The ASPP module leverages atrous convolutions at varied dilation rates to efficiently capture information across multiple scales, ensuring precise segmentation of objects of different sizes.

DeepLabV3+: This model enhances DeepLabV3 by adding an encoder-decoder structure for better detail and edge precision in semantic segmentation (Chu et al., 2021). It improves on outlining object boundaries and pixel labeling by refining the ASPP module with a decoder to efficiently capture object edges. Depth-wise separable convolution in the ASPP and decoder minimizes computational complexity, ensuring efficient, high-performance segmentation.

Pyramid Scene Parsing Network (PSPNet): This model employs a CNN-based structure, elevating scene parsing through its Pyramid Pooling Module, which aggregates multi-scale contextual information for superior global comprehension (Zhao et al., 2017). Utilizing features from four distinct pyramid scales, it captures a wide array of global details, essential for the precise parsing and interpretation of complex scenes. This strategic approach overcomes the challenge of fusing global contextual insights, significantly boosting scene parsing accuracy by thoroughly analyzing the interconnected relationships present within images.

SegNeXt: This model is CNN-based and introduces a novel convolutional attention mechanism to enhance computational efficiency (Guo et al., 2022). It leverages convolutional operations for spatial hierarchy management and local feature extraction, key for segmentation, avoiding the computational load of transformers' self-attention. This mechanism efficiently encodes spatial context with specialized convolutions, aiming to balance computational and parameter efficiency with high segmentation accuracy across various datasets.

2.3.2 Transformer-based Models

The landscape of semantic segmentation underwent another significant transformation with the introduction of Transformer-Based Models. Originally designed for natural language processing (Vaswani et al., 2017), the Transformer architecture, with its self-attention mechanism, offered a novel approach to handling image data in semantic segmentation tasks (Dosovitskiy et al., 2020; Ranftl et al., 2021; Zheng et al., 2021). This transition marked an exciting development in the field, leveraging the ability of Transformers to model long-range dependencies and global context effectively (Strudel et al., 2021).

A key milestone in this evolution was the adaptation of the Vision Transformer (ViT) for semantic segmentation. Pioneered by Dosovitskiy et al., ViT departed from conventional convolutional approaches, treating images as sequences of patches, and applying the self-attention mechanism to capture complex spatial relationships across the entire image

(Dosovitskiy et al., 2020). This approach allowed for a more complete understanding of the image context, leading to improvements in segmenting intricate scenes where global comprehension is crucial (Dosovitskiy et al., 2020).

Subsequent models built on the Transformer architecture further refined its application in semantic segmentation (Ranftl et al., 2021; Zheng et al., 2021). These models demonstrated an enhanced ability to segment objects and scenes with a high degree of accuracy, particularly in challenging scenarios involving occlusions, varying object scales, and complex backgrounds. By effectively capturing both local and global features through self-attention, Transformer-Based Models have set new benchmarks in the field (Strudel et al., 2021). The integration of Transformer technology into semantic segmentation represents a significant leap forward, underscoring the field's dynamic nature and its continuous pursuit of more advanced, efficient, and accurate segmentation methods. Transformer-based models are briefly described here:

Twins-SVT: This model, particularly through its variants Twins Pooled Convolutional Pyramid Vision Transformer (Twins-PCPVT) and Twins Scaled Vision Transformer (Twins-SVT), revolutionizes spatial attention in vision transformers with a novel and streamlined design (Chu et al., 2021). This innovation is marked by a simplified yet potent spatial attention mechanism that stands in contrast to the complex and resource-intensive approaches of traditional models. By employing a direct and efficient spatial attention strategy, both Twins-PCPVT and Twins-SVT architectures achieve high computational efficiency through optimized matrix multiplications, ensuring robust model performance without the burden of excessive computational demands.

Segmenting Transformers (SegFormer): This model is transformer-based and capitalizes on self-attention mechanisms to grasp global dependencies for improved scene understanding (Xie et al., 2021). It innovatively omits positional encoding, avoiding issues related to varying input image resolutions during testing. Additionally, SegFormer integrates a lightweight Multi-Layer Perceptron (MLP) decoder to blend multiscale features from the encoder, efficiently marrying local and global context for accurate segmentation outcomes.

Shifted Window (Swin) Transformer: This model is transformer-based and features a novel shifted window design that diverges from the fixed-size patches (Liu et al., 2021). This design enables adaptive feature extraction across scales, essential for semantic segmentation. It achieves linear computational complexity with image size, improving upon the quadratic complexity of standard transformers. Its hierarchical structure facilitates efficient processing of images at different resolutions, effectively extracting local and global features with enhanced accuracy and scalability.

Despite their strengths, these models' application in road lane detection remains underexplored, highlighting the necessity for a comparative study. Accordingly, this thesis undertakes a comprehensive evaluation of both CNN and Transformer-based models to identify their efficacy in road lane detection, addressing a critical research gap.

2.4 Evolution of Lane Marking Detection Methods

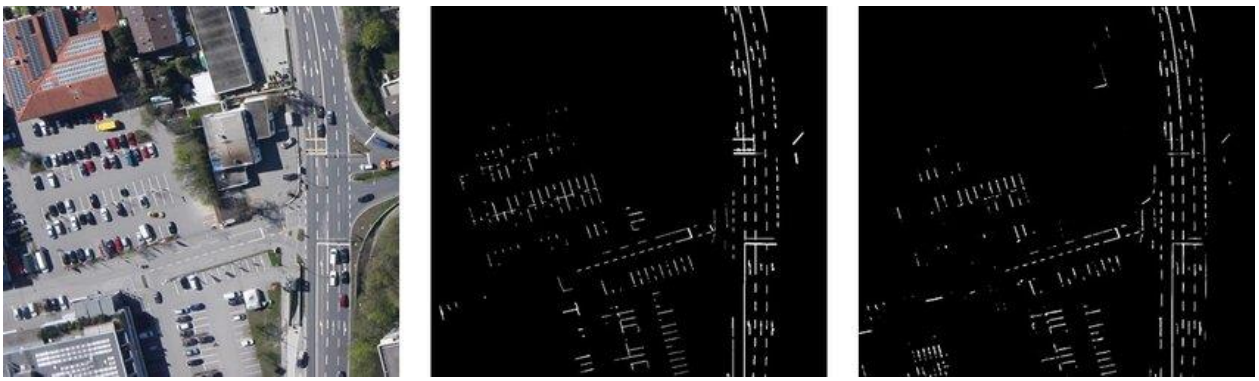
This section explores the progression of lane marking detection techniques, initially discussing the challenges associated with traditional road lane extraction methods from aerial imagery. It then transitions to deep learning, examining the application in general road extraction before focusing on its use for road lane extraction from aerial imagery.

Before the advent of deep learning, significant efforts were made in the realm of aerial imagery analysis to extract road lanes using traditional computational methods. However, compared to deep learning, traditional methods have notable limitations. The manual feature engineering process, requiring extensive domain expertise, contrasts with the automatic feature learning of deep learning, offering a more efficient and adaptive approach to lane detection (Li et al., 2021). Traditional techniques' sensitivity to environmental changes often led to inconsistent effectiveness, highlighting a significant gap that deep learning methods bridge with their robustness and precision. This advancement is crucial for autonomous vehicle development, where accurate and reliable lane detection under diverse conditions is paramount. Deep learning's superiority in handling complex scenarios significantly pushes the boundaries beyond the capabilities

of traditional approaches.

The exploration of deep learning methods to extract features from aerial imagery has surged, yielding significant advancements across various applications. Among these, road extraction serves as a pivotal area closely aligned with the challenges of road lane identification. For instance, Gao et al. (2018) leveraged a multi-feature pyramid network (MFPN) to adeptly handle roads of various widths in aerial imagery. A method refined by the same team with semantic segmentation and tensor voting to connect disjointed road segments (Gao et al., 2019). Concurrently, Wei et al. used of the Generative Adversarial Networks (GANs) for extracting both road pavement and centerlines (2020). The innovative approaches by Zhang et al (2019) and another Zhang et al in GAN-based road feature and obscured road reconstruction (2019), showcases the evolving capabilities of deep learning in comprehensive road network mapping. These advancements underscore the dynamic progression of deep learning applications. Yet, they also highlight an unexplored opportunity in applying such technologies for road lane extraction from aerial imagery, a critical area with substantial implications for automated navigation and urban planning.

However, research on road lane detection using aerial imagery, as of January 2024, remains limited, with a few seminal contributions, notably by Azimi et al (2019a; 2019b). Their Aerial LaneNet, unveiled in 2018, utilizes the Symmetric Fully Convolutional Neural Networks (FCNN) augmented with Wavelet Transforms to address the unique challenges of aerial imagery (Azimi et al., 2019a). The strategic use of the Discrete Wavelet Transforms (DWTs) with FCNNs is key in maintaining high-frequency details vital for lane marking identification, addressing the complexity of segmenting small features like lane markings from high-altitude images. This methodology enhances the model's multi-resolution analysis capability, crucial for accurate lane marking recognition across backgrounds.



(a) (b) (c)

Figure 2.1: Examples of results using Aerial LaneNet approach with the best performance. (a) Input aerial images, (b) Ground truth lane markings, and (c) Predictions of the extracted lane markings, adapted from Azimi et al. (2019a).

Despite its pioneering approach, Aerial LaneNet's performance, achieving a mean Intersection over Union (mIoU) of 77.98%, sets a benchmark in the domain (Azimi et al., 2019a). However, it is primarily configured for binary classification, distinguishing between lane markings and the background. This limitation highlights the need for future research directed towards multi-class segmentation to effectively tackle more complex scenarios in aerial view road lane detection.

Building on the Aerial LaneNet foundation, Azimi et al. (2019b) introduced SkyScapesNet to delve into urban infrastructure complexities through aerial imagery. This model addresses the critical need for detailed and precise segmentation of urban landscapes, crucial for autonomous driving and urban planning. Specifically designed to segment multiple classes of road lanes and accurately identify small-scale urban features, SkyScapesNet utilizes a multi-task approach. By combining dense semantic segmentation with semantic edge detection, it leveraged the SkyScapes dataset, which comprises 31 semantic categories, to enhance the segmentation of densely populated urban areas.

SkyScapesNet, based on the Fully Convolutional DenseNet (FC-DenseNet) architecture, includes features designed for aerial images and uses a multi-task approach to improve object boundaries and feature identification (Azimi et al., 2019b). With an IoU of 40.13% and precision of 65.93%, it effectively classifies urban features, marking progress in aerial

urban segmentation.

The limited research on road lane detection using aerial imagery, highlighted by notable contributions from Azimi et al. (2019a, 2019b), reveals a significant gap in the field, especially in multi-class segmentation challenges. With only two main studies addressing this area, there is a clear need for comparative research to evaluate and advance deep learning techniques for aerial imagery. Such studies are crucial for improving autonomous driving providing more accurate and detailed segmentation of road lanes from aerial perspectives.

This section has outlined the evolution of lane marking detection from traditional computational techniques to advanced deep learning approaches, highlighting the shift towards more sophisticated methods in aerial imagery analysis. While traditional methods provided a foundation, deep learning has introduced unparalleled precision and adaptability, significantly advancing road lane extraction. Despite notable contributions, particularly from Azimi et al., there remains a crucial need for further research, especially in multi-class segmentation for comprehensive aerial view analysis.

Chapter 3 Methodology for Automated Lane Marking Extraction

This chapter outlines the methodology employed for automated lane marking extraction. Initially, the focus is on introducing the two primary training datasets utilized in this study. Subsequently, a graphical representation detailing the general workflow is presented to provide a clear overview. Following this, the chapter delves into the training environment setup, before progressing to the imaging pre-processing techniques employed. An in-depth description of the models' training process is then provided, leading to an explanation of the transfer learning approach adopted. Finally, the evaluation metrics used to assess the effectiveness of the extraction methodology are detailed.

3.1 Description of Training Datasets

3.1.1 SkyScapes Dataset

The SkyScapes Dataset is a comprehensive set of aerial images taken over Munich, Germany (Azimi et al., 2019b). Munich is the capital of Bavaria and the third-largest city in Germany, stands as a pivotal economic and transportation hub in the region. With an extensive area of over 310 km² and a population exceeding 1.5 million, the city showcases an advanced public transportation system. The presence of well-defined road lane markings further enhances traffic flow and safety, highlighting Munich's crucial role in linking major European transportation corridors.

Within this context, the SkyScapes Dataset provides a detailed aerial perspective of Munich. Captured using a helicopter-mounted Digital Single-Lens Reflex (DSLR) camera system (Azimi et al., 2019b), it consists of 16 RGB images with a resolution of 5,616×3,744 pixels and offers a ground sampling distance of approximately 13 cm per pixel. Spanning an area of 5.7 km², the dataset covers both urban and rural environments as shown in the Figure 3.1, meticulously showcasing traffic conditions and the interplay of Munich's comprehensive transportation infrastructure with its geographical landscape.

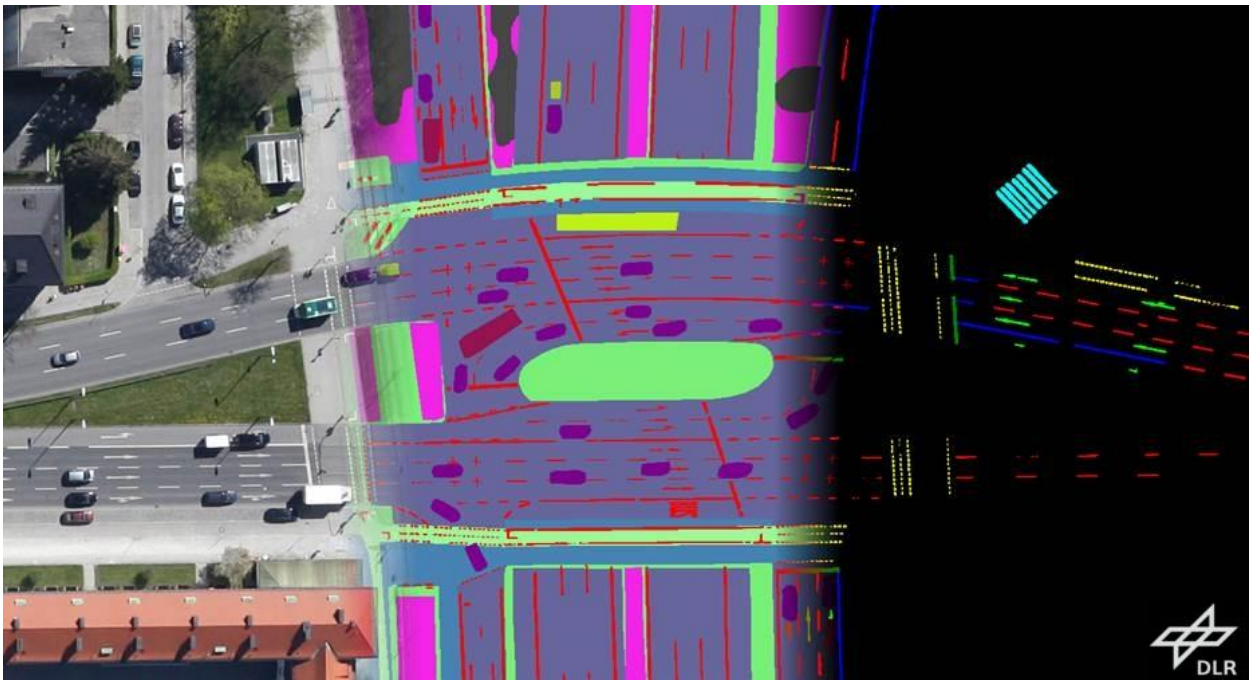


Figure 3.1: An example of the SkyScapes Dataset. Aerial image (left), semantic annotation (centre), and training image input (right) (Azimi et al., 2019b).

This dataset includes 31 carefully annotated semantic categories, with a primary focus on elements found in urban areas (Azimi et al., 2019b). These categories cover a variety of features such as low vegetation, different types of roads and parking places, bikeways, sidewalks, entrance/exit points, danger zones, buildings, various vehicle types including cars, trailers, vans, trucks, large trucks, buses, as well as clutter, impervious surfaces, trees, and 12 different lane-marking types. The lane marking types are specified as dash-line, long-line, small dash-line, turn sign, plus sign, other signs, crosswalk, stop-line, zebra zone, no parking zone, parking zone, and other lane-markings. These categories are selected for their direct relevance to real-world urban scenarios, with a particular emphasis on road-associated objects to support applications like urban planning and autonomous driving.

Annotations for the dataset were created through a meticulous manual process aimed at ensuring high precision, a necessity for applications such as autonomous vehicle navigation (Azimi et al., 2019b). The labor-intensive annotation required approximately 200 man-hours for each image, highlighting the dataset's detail and the accuracy of its

annotations.

Overall, the SkyScapes dataset comprises more than 70,000 instances spread across its 31 classes, demonstrating a wide variety of class sizes and complexities (Azimi et al., 2019b). Additionally, it offers an in-depth examination of lane markings through the specific SkyScapes-Lane task, enriching its utility for detailed urban analysis.

Table 3.1: Number of annotated pixels (filled) per class in SkyScapes, multi-lane.

Class	Pixel Count
Background	167,817,849
Dash line	74,502
Long line	237,217
Small dash line	10,523
Turn signs	4,915
Other signs	7,914
Crosswalk	4,422
Stop line	5,428
Zebra zone	29,290
No parking zone	2,960
Parking space	5,154
Other lane-marking	10,258

3.1.2 Waterloo Urban Scene Dataset

Waterloo is a city located in Ontario, Canada, is part of the Region of Waterloo metropolitan region, which features an extensive network of roads, including significant highways that improve its connection with the Greater Toronto Areas and nearby areas. The complex system of highways and city streets is essential for ensuring smooth traffic movement and safety, making Waterloo a suitable place for conducting research on road lane extraction.

Building on this, the Waterloo Urban Scene Dataset, extracted from the readily available Waterloo Building Dataset, presents high-resolution aerial ortho imagery of the

Kitchener-Waterloo region, as shown in Figure 3.2 and 3.3. With its extensive area coverage of 205.8 km² and a fine spatial resolution of 12 cm per pixel, this dataset serves as a perfect platform for urban and traffic semantic segmentation projects (He et al., 2022). The dataset's detailed imagery and broad accessibility have been crucial factors in its selection for this study, offering a detailed and expansive foundation for analytical endeavors.

Figure 3.2 An example of the Waterloo Urban Scene dataset raw image.

Figure 3.3 The annotation of 14 road lane marking class and background for Figure 3.2.

To adapt the Waterloo Urban Scene Dataset for this study, it was enriched with a series of manually added annotations to form a robust ground truth, crucial for evaluating the developed model's effectiveness across diverse datasets. Given the original dataset's lack of specific urban and traffic classifications, 14 classes were introduced, including road, vehicle, sidewalk, crosswalk, various lane markings, and traffic islands. This addition of

classes and annotations aligns the dataset with the practical demands of urban and traffic contexts, facilitating precise model testing and development.

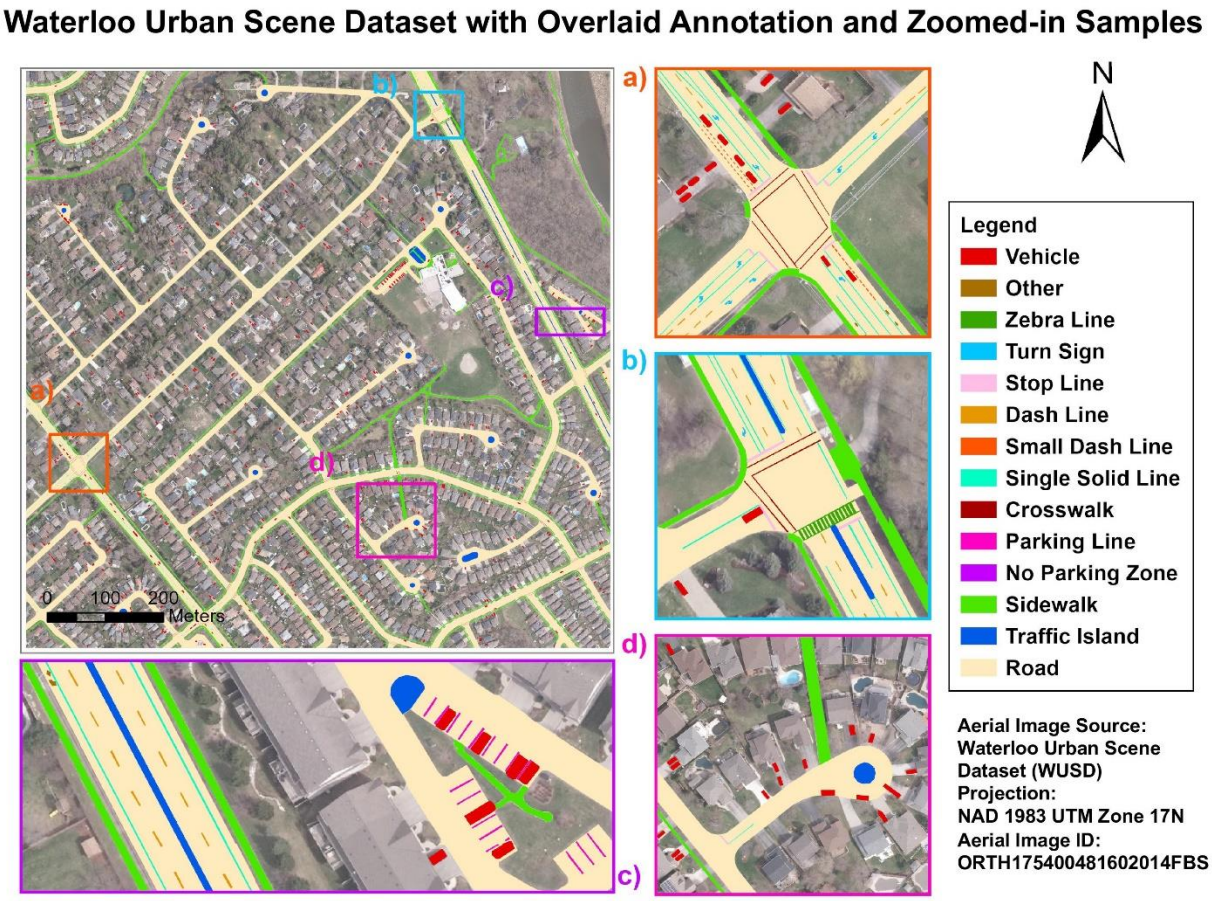


Figure 3.4: Illustrations from the Waterloo Urban Scene Dataset featuring annotated overlays.

These classes were organized into three main categories to enhance clarity and specificity. Facility types, including background, road, traffic island, and sidewalk, form the first group. The second group encompasses road lane markings, while the third is focused on vehicles. This organization mirrors real-world conditions for accurate model testing.

Table 3.2: Number of annotated pixels (filled) per class in the Waterloo Urban Scene Dataset.

Class	Pixel Count
Background	421,098,771
Road	58,305,344
Traffic Island	2,211,212
Sidewalk	15,421,199
Crosswalk	122,878
Dash line	138,418
Single solid line	833,854
Small dash line	14,050
Turn sign	16,990
Stop line	76,384
Zebra line	25,748
Parking line	477,782
No parking zone	300,028
Other	10,072
Vehicle	4,788,038

An essential aspect of our annotation process was establishing a priority system for overlapping classes in the imagery. This hierarchy was crucial for resolving cases where a pixel belonged to multiple classes, such as a vehicle overlapping a dash line on a road. The hierarchy ensuring clarity and consistency in the annotations:

Vehicles Group: Given top priority, this category includes all vehicle types, reflecting their non-occlusion by roads or lanes.

Road Lane and Signs Class: Occupying the second level. This group encompasses the following sequence: Dash Line, Single Solid Line, Small Dash Line, Turn Sign, Stop Line, Zebra Line, Parking Line, No Parking Zone, and others.

Facility Types: The lowest tier, consisting of infrastructure elements. This group includes the following classes: Road, Traffic Island/Roundabout, Sidewalk, Background (Undefined Features, e.g., trees, buildings, waterbody). Although these elements form the fundamental structure of urban landscapes, they are assigned the lowest priority in the

context of this specific research objective.

The annotation process posed significant challenges, particularly due to the intricacies of aerial imagery (Azimi et al., 2019b). Factors such as small object size, shadows, occlusion, and unclear boundaries added complexity to the task. To facilitate detailed annotation, all the images are annotated by the 14 research assistants and group members from the Waterloo Geospatial Intelligence and Mapping Laboratory (GIM Lab) with expertise in geomatics. We employed GIS fishnet techniques to segment large images. As illustrated in Figure 3.3, it displays a 17×17 grid structure. The numeration and directional arrow depict the methodology employed to segment the image using the fishnet approach, subsequently allocating a unique identifier to each subsection for future annotation purposes. The numbering of cells follows a serpentine pattern, beginning with 0 in the bottom left corner and concluding with 16 in the bottom right corner. The final cell, numbered 289, is in the top right corner. In the Waterloo region, for instance, each large image was divided into 86 smaller images with 8,350×8,350 pixels. Each larger image in the dataset is segmented into 289 smaller images, with most measuring 512×512 pixels, except the last row and column. After annotation, these images underwent three rounds of quality checks.

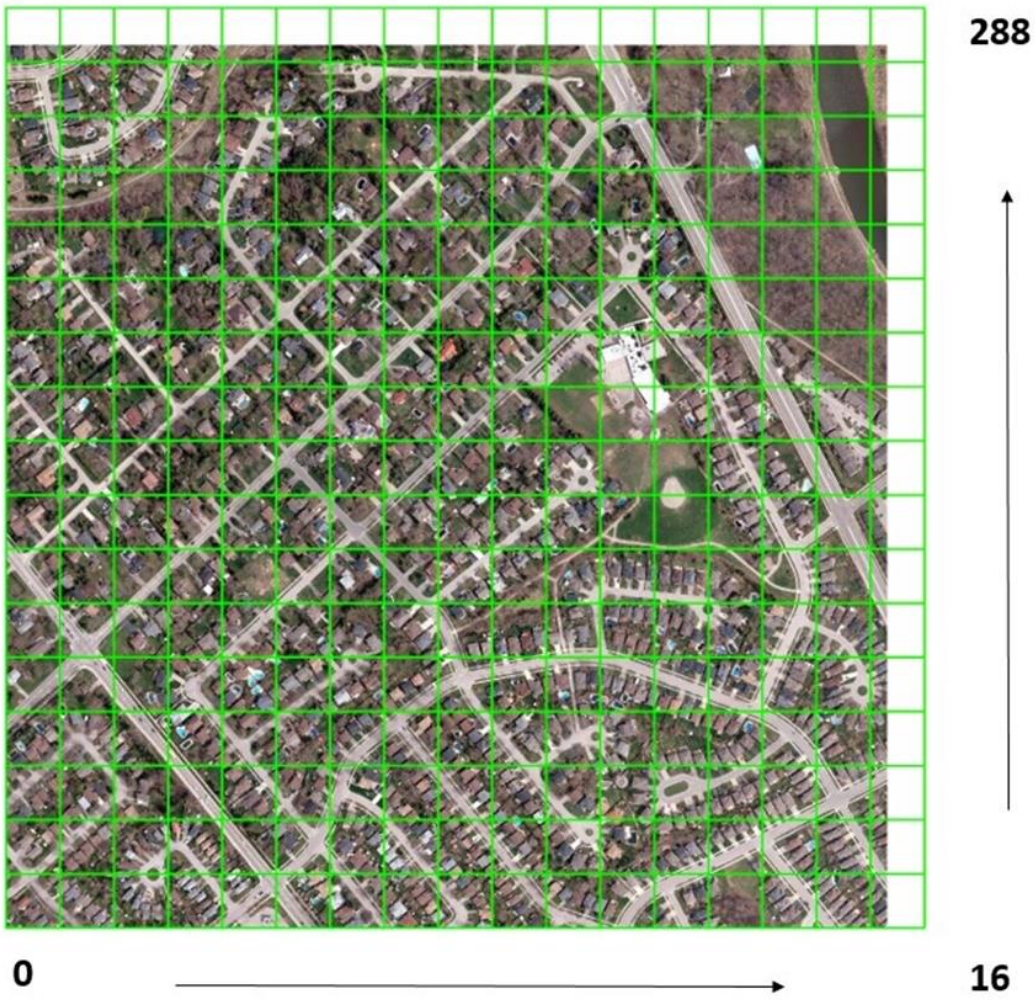


Figure 3.5: Aerial Perspective of an Urban Region with Fishnet Grid Overlay. This illustration displays a 17×17 grid structure with numeration and a directional arrow to depict the segmentation methodology.

3.2 Description of the General Workflow

As depicted in Figure 3.4 below, this section outlines the workflow of the experiment. Initially, there was preparation involving two datasets: the Skyscapes Dataset and the Waterloo Urban Scenes Dataset. Both datasets underwent data augmentation and parameter calculation before entering the training phase. In data augmentation, original images in Red Green Blue (RGB) format with three channels were flipped and cropped to increase the size of the dataset. For parameter calculation, the mean and standard deviation were computed on the training images to normalize the inputs for the training phase, which helps reduce training time and computational resources.

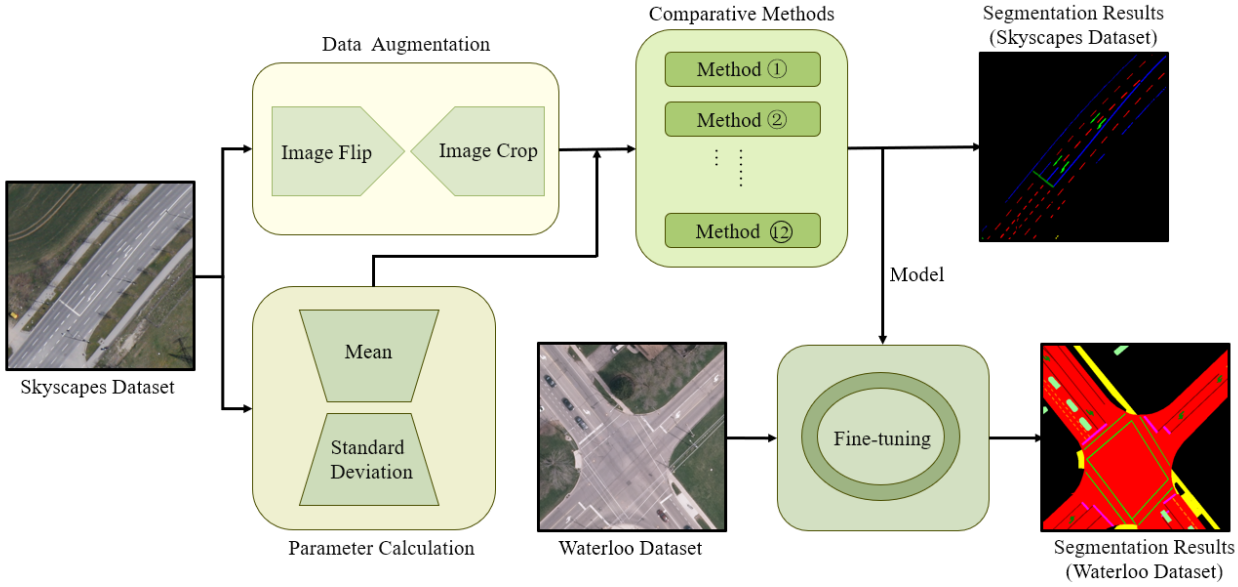


Figure 3.6: General workflow of the experiment

Following this, twelve comparative models were trained using the Skyscapes dataset and evaluated to produce performance metrics on the Skyscapes test dataset. Then, these twelve models, preloaded with weights from their training on the Skyscapes dataset, were fine-tuned using the Waterloo Urban Scenes dataset, after which their performance metrics were generated for this dataset.

3.3 Training Environment

The training was conducted using NVIDIA RTX 3080 GPU. Learning rate was chosen based on pretrained model setups with proven convergence from toolbox. Given the intensive computational requirements, a batch size of 2 was opted to most model training setup. The models were trained around 20 epochs, a duration determined to be sufficient for converging to a stable solution without overfitting, based on the validation performance.

3.4 Imaging Pre-processing

The comparative study utilized the SkyScapes and the Waterloo Urban Scene datasets, which are composed of high-resolution aerial imagery. The SkyScapes dataset includes a total of 10 images, divided into 8 training and 2 validation images, each with dimensions of 5,616×3,744 pixels (Azimi et al., 2019b). The Waterloo Urban Scene dataset, on the other hand, consists of images with a resolution of 8,350×8,350 pixels (He et al., 2021). To facilitate the analysis of ground truth and predictions, distinct palettes were developed to represent different classes. The grayscale ground truths were converted into palette-labeled images for ease of use.

Given the large size of the original images and the memory limitations of Graphics Processing Units (GPU), the raw images were segmented into smaller patches measuring 512×512 pixels. This patch size, chosen based on the SkyScapes study, represents a compromise between preserving adequate contextual information and ensuring a manageable computational load.

To enhance model robustness and account for various orientations, the datasets underwent augmentation through horizontal flips, vertical flips, and combined flips. These techniques expanded the diversity of the training data without necessitating additional labeled images. To further maximize data utility, adjacent patches were overlapped by 50% in both vertical and horizontal dimensions. This approach aimed to provide exhaustive coverage of the imagery and minimize information loss at patch

boundaries. Consequently, the training sets for both the SkyScapes dataset and the Waterloo Urban Scene dataset comprised approximately 8,000 images each.

Prior to training, the images were standardized to normalize the data distribution. The mean and standard deviation of the training set, which was calculated from the SkyScapes training dataset, was applied to both training and test datasets. Considering the Waterloo Urban Scene dataset was fine-tuned from the SkyScapes dataset, it was subjected to the same standardization parameters to ensure consistency in model training.

3.5 Models Training

This section outlines the enhanced procedures for training the model, detailing settings for parameters, strategies for optimization, and techniques for managing data. Modifications are made in response to initial experiments and the distinct features of the datasets being examined, specifically the SkyScapes and Waterloo Urban Scene dataset.

Parameter Settings

Models were initialized with parameters derived from two sources: pre-trained ImageNet 1K backbones and settings identified through preliminary experimentation. This hybrid approach ensured a solid foundation for feature extraction, supplemented by adjustments tailored to the unique demands of aerial imagery analysis. Learning rates were individually set for each model to accommodate their sensitivity to large adjustments. Here are detailed learning rate settings for each model: ANN and Twins used 0.0001; DeepLabV3, DeepLabV3+, FCN, FastFCN, MobileNetV3, PSPNet, U-Net used 0.01; SegFormer, SegNeXt, and Swin used 0.00006. A common strategy across all models was the employment of a warm-up phase, constituting approximately half an epoch, where learning rates were minimized to stabilize the models at the beginning. Subsequently, a Poly Learning Rate schedule with a power of 0.9 was applied to systematically decrease learning rate, optimizing the training progression. The batch size was set at 2 for most of the models except for MobileNetV3, which had a batch size of 4, a decision dictated by the large input size of 512x512 pixels and the constraints of memory usage. Smaller batch size necessitated careful management of learning dynamics to maintain model effectiveness.

Optimization Strategies

The training routine for the SkyScapes dataset extends across 20 epochs, a period during which most models typically achieve convergence, signifying optimal performance in feature extraction from the dataset. The training for the Waterloo Urban Scene requires fewer epochs because the models have already gained a general understanding of scenes from the SkyScapes dataset, and only need to adjust this knowledge to a similar domain through fine-tuning. The research employed AdamW and Stochastic Gradient Descent (SGD) optimizers to leverage their unique advantages in managing sparse gradients and momentum. Specifically, AdamW was the optimizer of choice for SegFormer, SegNeXt, Swin, and Twins, while the remaining models were optimized using SGD. A diverse suite of loss functions, including cross-entropy loss, soft IoU loss, dice loss, and focal loss, was combined with varying weights. Further information on the loss functions employed for both datasets across the 12 models can be found in Appendix C, specifically in Table C.1. This approach allowed for a detailed learning process, accommodating the complex spatial relationships and varying object scales present in aerial images.

Data Splits

The SkyScapes dataset, publicly available with predefined splits, comprises 8 training, 2 validation, and 6 test images. This split was adopted without modification to ensure consistency with public benchmarks. In contrast, the Waterloo Urban Scene dataset, featuring larger original images than the SkyScapes dataset, included 1 original image for both training and validation phases. These were further segmented, with a training-validation split of 80:20 applied to the cropped images, facilitating a focused examination of urban landscapes.

This detailed account of the model training methodology integrates specific parameter adjustments, optimization strategies, and a tailored approach to data management, reflecting the complexity and specificity of applying transfer learning to aerial imagery analysis. Through this rigorous and methodically adjusted training process, the study aims to provide a robust and reproducible framework for analyzing diverse landscapes

and urban scenes.

3.6 Transfer Learning Approaches

This study employs a strategic transfer learning approach to leverage pre-existing knowledge from a broad dataset to enhance the performance on specific urban and SkyScapes scenes. The methodology is structured around the application of transfer learning from a general dataset to the SkyScapes dataset and subsequently transferring the learned features to the Waterloo Urban Scene dataset.

Pre-Trained Backbone Models: At the core of our transfer learning strategy lie twelve models, each utilizing a backbone pre-trained on the ImageNet 1K dataset. These backbones include, but are not limited to, ResNet-50, ResNet-101, and Vision ViT. The choice of these models is predicated on their proven efficacy in various image recognition tasks and their ability to serve as robust feature extractors for diverse visual domains.

Training on SkyScapes Dataset: Initially, the models are fine-tuned on the SkyScapes dataset, a comprehensive collection of diverse SkyScapes images. To accommodate the unique learning capabilities and preferences of each model, a variety of loss functions were employed. A simple preliminary experiment was conducted to identify the optimal combination of loss functions and the learning rate tailored for each specific model. Furthermore, the models were trained over slightly different numbers of epochs to achieve the best equivalent effects, ensuring a customized and efficient learning process that aligns with the distinct characteristics of each model. This tailored approach ensures that the models not only learn effectively but also adapt to the specific details of SkyScapes imagery.

Fine-Tuning for Waterloo Urban Scene Dataset: Upon achieving satisfactory performance on the SkyScapes dataset, the models undergo a subsequent phase of fine-tuning for the Waterloo Urban Scene dataset. This dataset, focusing on urban scenes in Waterloo, presents a distinct set of challenges and characteristics compared to SkyScapes.

The fine-tuning process is meticulously designed to leverage the SkyScapes-based pre-training, allowing the models to transfer and adapt their learned features to urban scenes effectively. Fine-tuning on the Waterloo Urban Scene dataset involved a faster decline of the learning rate compared to the SkyScapes dataset, acknowledging the models' pre-existing familiarity with related scenes. In this stage, the training duration was controlled more flexibly by using iterations, which means batches, instead of being limited by a fixed number of batches in each epoch. The optimizers were reduced while the combination of loss functions from the initial training phase were retained from those in the SkyScapes dataset, ensuring continuity in the optimization logic while accommodating the datasets' differing characteristics.

This dual-stage transfer learning approach, from a generalized dataset to the SkyScapes dataset and subsequently to the Waterloo Urban Scene dataset, underscores the flexibility and effectiveness of leveraging pre-trained models for domain-specific tasks. By fine-tuning pre-existing models, powerful representational capabilities are utilized, significantly enhancing performance on targeted tasks within the areas of urban and SkyScapes scene analysis.

3.7 Evaluation Metrics

The evaluation of semantic segmentation models, particularly in remote sensing imagery, prioritizes segmentation accuracy due to its critical role in application effectiveness. A set of evaluation metrics is used to assess segmentation methodologies' performance, offering insights into the precision of pixel classification. The selected metrics include:

All Pixel Accuracy (aAcc): This metric calculates the overall accuracy of the segmentation across the entire dataset. It is defined as the ratio of correctly classified pixels to the total number of pixels. aAcc provides a high-level overview of model performance but may be biased towards dominant classes in unbalanced datasets.

$$All\ Pixel\ Accuracy = \frac{N_{Correct\ Pixels}}{N_{All\ Pixels}} \tag{3.1}$$

Mean Pixel Accuracy (mAcc): This metric calculates the accuracy for each class individually and then averages these accuracies. This metric offers insight into the model's consistency across different classes.

$$Mean\ Pixel\ Accuracy = \frac{\sum_1^{N_{Class}} Accuracy_{Class\ i}}{N_{Class}} \tag{3.2}$$

Mean Intersection over Union (mIoU): IoU is a metric that measures the overlap between the predicted segmentation and the ground truth, normalized by the union of these two areas. mIoU is calculated for each class and then averaged, providing a balanced view of the model's precision and recall.

True Positives (TP) refers to the instances where both predicted and actual values are positive. False Positives (FP) refers to the instances where predicted value is positive but actual value is negative. True Negatives (TN) refers to the instances where both predicted and actual values are negative. False Negatives (FN) refers to the instances where predicted value is negative but actual value is positive.

$$Intersection\ Over\ Union = \frac{TP}{TP + FP + FN} \tag{3.3}$$

$$Mean\ Intersection\ Over\ Union = \frac{\sum^{N_{Class}} IoU_{Class}}{N_{Class}} \tag{3.4}$$

Precision: Precision, or the positive predictive value, quantifies the ratio of true positive pixels to the sum of true positive and false positive pixels for each class, averaged across all classes. It reflects the model's ability to exclude false positives from the segmentation. Mean precision is the mean value over all classes' precisions.

$$Precision = \frac{TP}{TP + FP} \tag{3.5}$$

Recall: Recall, also known as sensitivity, measures the ratio of true positive pixels to the sum of true positive and false negative pixels for each class, averaged across classes. This metric highlights the model's capability to correctly identify all relevant pixels of a class. Mean recall is the mean value over all classes' recalls.

$$Recall = \frac{TP}{TP + FN} \tag{3.6}$$

F1-Score: The F1-score is the harmonic mean of precision and recall. It serves as a single metric that balances the trade-off between precision and recall, providing a comprehensive measure of the model's overall performance. The mean F1-score is computed by averaging the F1-scores obtained for each class.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.7}$$

Each of these metrics addresses different aspects of segmentation performance, from general accuracy to class-specific precision and recall. By utilizing this multifaceted evaluation framework, we aim to provide a thorough and detailed analysis of the segmentation models under study.

Chapter 4 Results and Discussion

This chapter presents the outcomes of the conducted experiments and engages in further discussions. Section 4.1 details the experimental results of two datasets. Section 4.2 focuses on the visualization of these results. In-depth discussions based on these findings are provided in Section 4.3.

4.1 Model Performance and Adaptation

4.1.1 SkyScapes Dataset

This section presents a comprehensive analysis of 12 models' performance on the SkyScapes dataset. The evaluation employs a suite of metrics, including mIoU, mAcc, overall Accuracy, mean Recall, mean Precision, and mean F1 score, to facilitate a detailed examination of each model's effectiveness. Key insights and noteworthy observations from the comparative assessment are highlighted, with additional detailed class-based quantitative results available in Appendix A for a deeper understanding of performance differences.

Based on Table 4.1 below, transformer-based models outperform CNN-based models. Within the transformer category, models such as SegFormer, Swin, and Twins achieve superior results compared to traditional CNN models, highlighting the effectiveness of attention mechanisms over non-attention-based approaches. Notably, SegNeXT, even without employing a transformer structure, achieves the second-best result in terms of MIoU among the 12 models through its unique CNN attention mechanism. Among the evaluated models, SegFormer leads in performance with a mIoU of 33.56%, mAcc of 99.92%, overall Accuracy of 99.59%, mean Precision of 64.33%, and an F1 score of 44.34%. The highest mean Recall is achieved by ANN at 66.00%.

Table 4.1: Benchmark of the state of the art on the SkyScapes-Lane task over all 12 classes (in %).

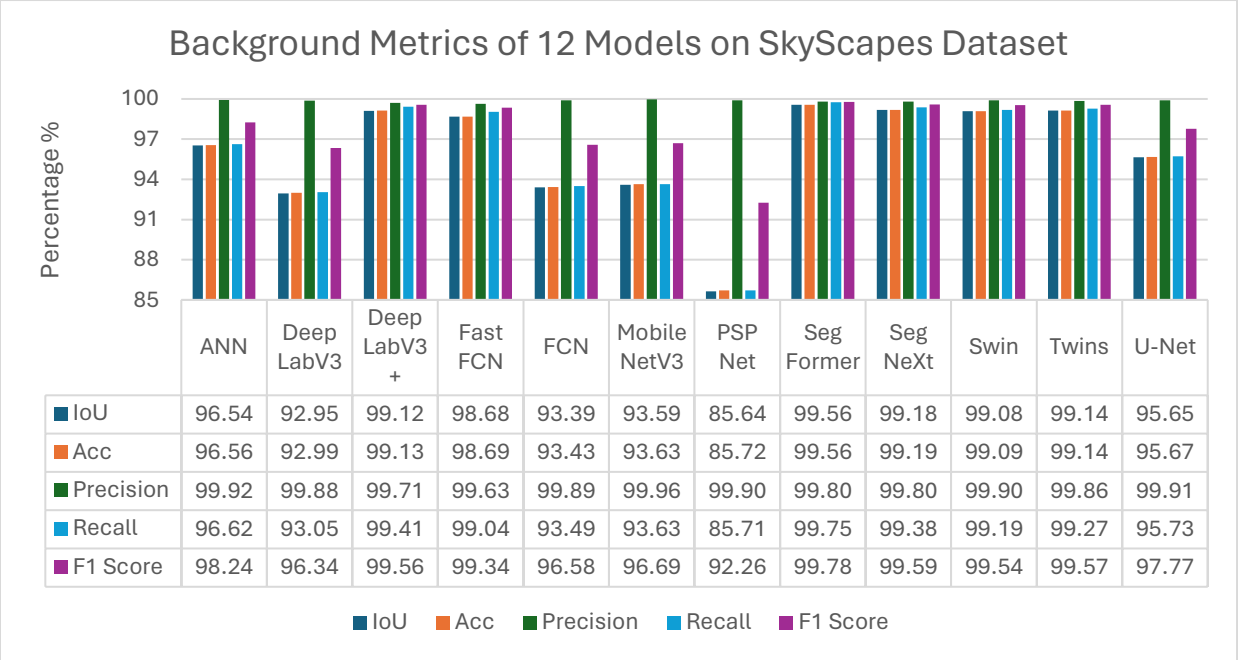
Method	Base	IoU	Accuracy		Confusion		F1
		mean	mAcc	aAcc	mRecall	mPrecision	
FCN	ResNet101	10.68	98.87	93.21	13.53	11.38	13.53
FastFCN	ResNet50	16.33	99.77	98.61	31.10	21.02	22.36
U-Net	-	14.98	99.24	95.43	50.45	18.97	21.80
DeepLabV3	ResNet101	10.24	98.79	92.72	35.32	11.12	12.69
DeepLabV3+	ResNet101	18.08	99.84	99.02	34.62	23.58	24.91
ANN	ResNet101	20.94	99.41	96.47	**66.00**	22.92	29.88
MobileNetV3	Large	11.74	98.91	93.47	54.55	12.45	15.25
PSPNet	ResNet101	8.68	97.55	85.29	32.24	10.09	10.65
SegNeXt	Base	32.26	99.86	99.14	49.96	48.98	44.20
Twins	Base	30.11	99.85	99.08	45.79	51.81	41.87
Swin	Base	30.51	99.84	99.02	60.03	40.27	42.97
SegFormer	Base	**33.56**	**99.92**	**99.50**	43.85	**64.33**	**44.34**

Generally, Recall exceeds Precision across the models, except for Twins and SegFormer, which display higher mean Precision than mean Recall. This suggests their predictions tended to minimize false negatives over false positive to achieve high recall, perhaps over predicting pixels as positives. The F1 score, a harmonic mean of Precision and Recall, serves as a balanced metric, emphasizing that both Precision and Recall need to improve proportionally to enhance the F1 score. It reflects the overall prediction quality without bias towards either Precision or Recall. The mIoU metric quantifies the overlap between predicted and actual class pixels, providing a direct measure of prediction accuracy in relation to the ground truth. Like the F1 score in its construction but with different coefficients, IoU and F1 scores generally exhibit parallel trends in model performance evaluation.

All models demonstrate exceptionally high Accuracy, with each model achieving at least 97.55% mAcc and some nearing 99.92%. This phenomenon is attributed to the dominant presence of background pixels, where accurate background prediction significantly influences the Accuracy metric, leading to scale imbalance and diminished result sensitivity. Further interpretation of these results will be elaborated in discussion chapter.

Within the class-based analysis of the SkyScapes dataset involving all 12 models from Table 4.2 below, it is observed that the background class universally achieves the highest results across all evaluation metrics. Detailed per-model per-class metrics can be found in appendix A.1. This trend is attributable to most pixels that the background class constitutes, enabling models to master its prediction more easily. This ease of learning for the background prediction is particularly pronounced when models encounter difficulties in discerning minor variations in loss attributed to predictions of other classes, indicating a general propensity for models to excel in background identification when faced with complex class distinctions.

Table 4.2: Evaluation metrics for 12 different models on the performance of background extraction using the SkyScapes Dataset.



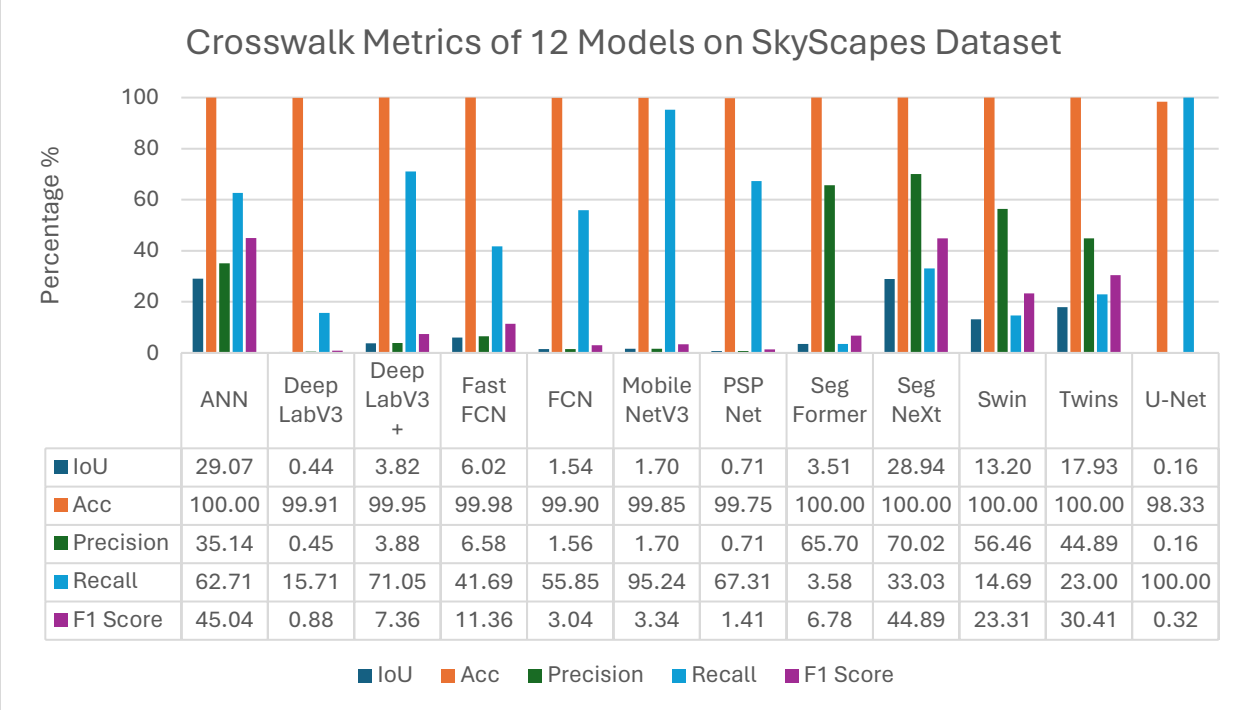
	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	96.54	92.95	99.12	98.68	93.39	93.59	85.64	99.56	99.18	99.08	99.14	95.65
Acc	96.56	92.99	99.13	98.69	93.43	93.63	85.72	99.56	99.19	99.09	99.14	95.67
Precision	99.92	99.88	99.71	99.63	99.89	99.96	99.90	99.80	99.80	99.90	99.86	99.91
Recall	96.62	93.05	99.41	99.04	93.49	93.63	85.71	99.75	99.38	99.19	99.27	95.73
F1 Score	98.24	96.34	99.56	99.34	96.58	96.69	92.26	99.78	99.59	99.54	99.57	97.77

Based on Table 4.2, we can see that PSPNet and Mobilenetv3 have mean IoUs of 85.64% and 93.59%, respectively. This deviation suggests a tendency for these models to misclassify more pixels as non-background classes, pointing towards a distinct behavior in handling class predictions that diverges from the norm observed in other models.

Furthermore, a detailed examination reveals that each model exhibits a distinct preference for certain classes beyond the background, manifesting in varied performance

metrics from Table 4.3 to 4.13. Referencing to Table 4.3, which details the metrics for crosswalk identification, it is noteworthy that several models, despite exhibiting lower scores across other metrics, achieve notable high recall rates, particularly U-Net, which reaches a recall of 100%. This phenomenon is primarily attributed to overestimation of overlap, coupled with a resultant decrease in precision. Due to the limitations inherent in the models' capabilities, they tend to predict a much broader area than the actual crosswalk, merging it into a unified block. This results in a significantly low number of False Positives, which is a critical factor in the recall's denominator, alongside the quantity of correct predictions. Among the models, those based on ANN secure the highest scores in recognizing crosswalks, suggesting that ANNs may possess structural advantages or have been more finely optimized for detecting crosswalk features compared to their transformer-based counterparts.

Table 4.3: Evaluation metrics for 12 different models on the performance of crosswalk extraction using the SkyScapes Dataset

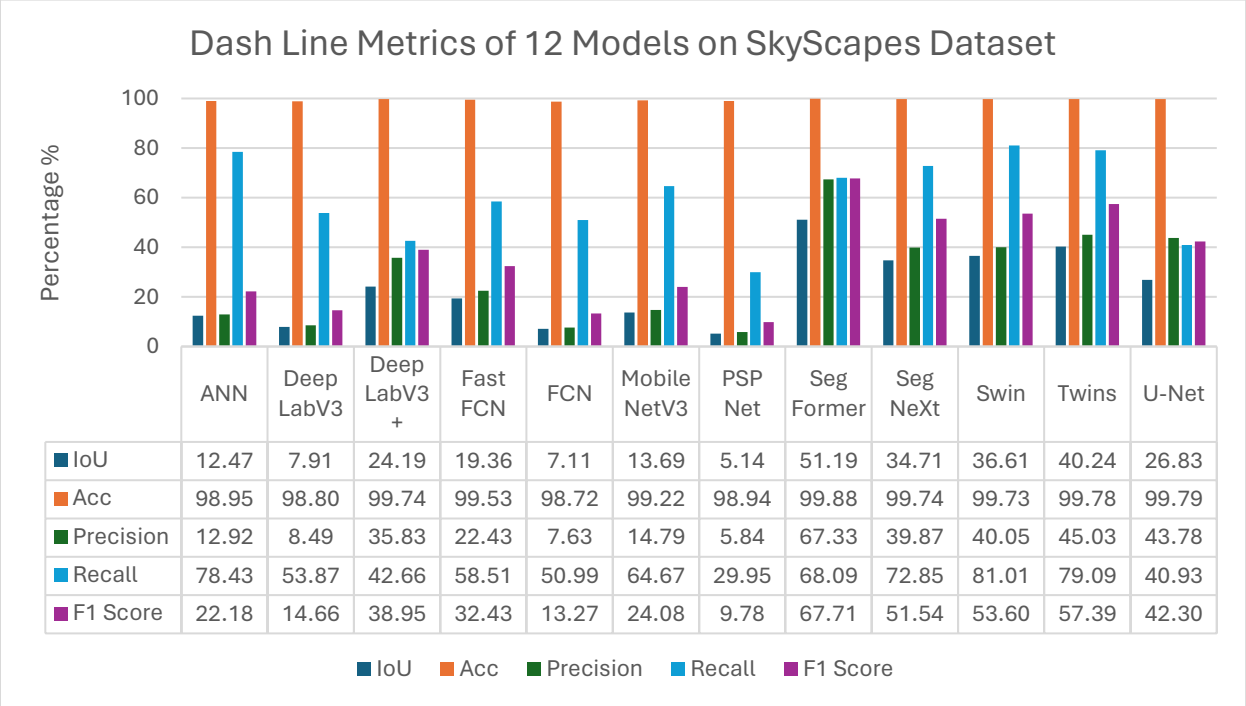


	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	29.07	0.44	3.82	6.02	1.54	1.70	0.71	3.51	28.94	13.20	17.93	0.16
Acc	100.00	99.91	99.95	99.98	99.90	99.85	99.75	100.00	100.00	100.00	100.00	98.33
Precision	35.14	0.45	3.88	6.58	1.56	1.70	0.71	65.70	70.02	56.46	44.89	0.16
Recall	62.71	15.71	71.05	41.69	55.85	95.24	67.31	3.58	33.03	14.69	23.00	100.00
F1 Score	45.04	0.88	7.36	11.36	3.04	3.34	1.41	6.78	44.89	23.31	30.41	0.32

Table 4.4 provides an in-depth analysis of the performance metrics related to the detection of dashed lines, indicating a notably high proficiency across models in this specific area, surpassing their effectiveness in several other categories. Models that

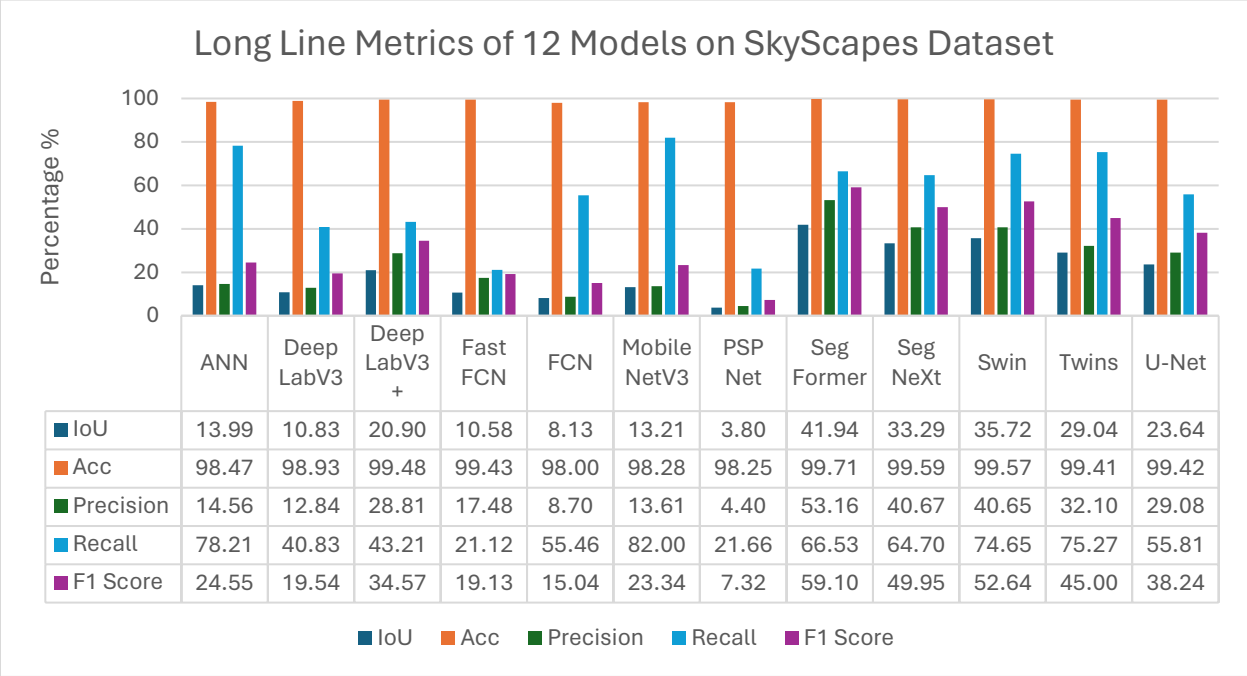
integrate attention mechanisms are particularly distinguished, invariably ranking within the top quartile for performance metrics. Notably, the U-Net model demonstrates exceptional performance, achieving consistently high average scores that position it just below the leading models. This achievement highlights U-Net's adeptness in capturing the intricate details of object contours accurately. Transitioning to the examination presented in Table 4.5, the analysis of the detection capabilities for longer lines corroborates the trends observed in dashed line detection, further evidencing the models' proficiency in identifying distinct and continuous line elements. This consistency evidences a comprehensive capability among the models, especially those akin to U-Net, in reliably discerning and precisely classifying various types of line-based road markings.

Table 4.4: Evaluation metrics for 12 different models on the performance of dash line extraction using the SkyScapes Dataset.



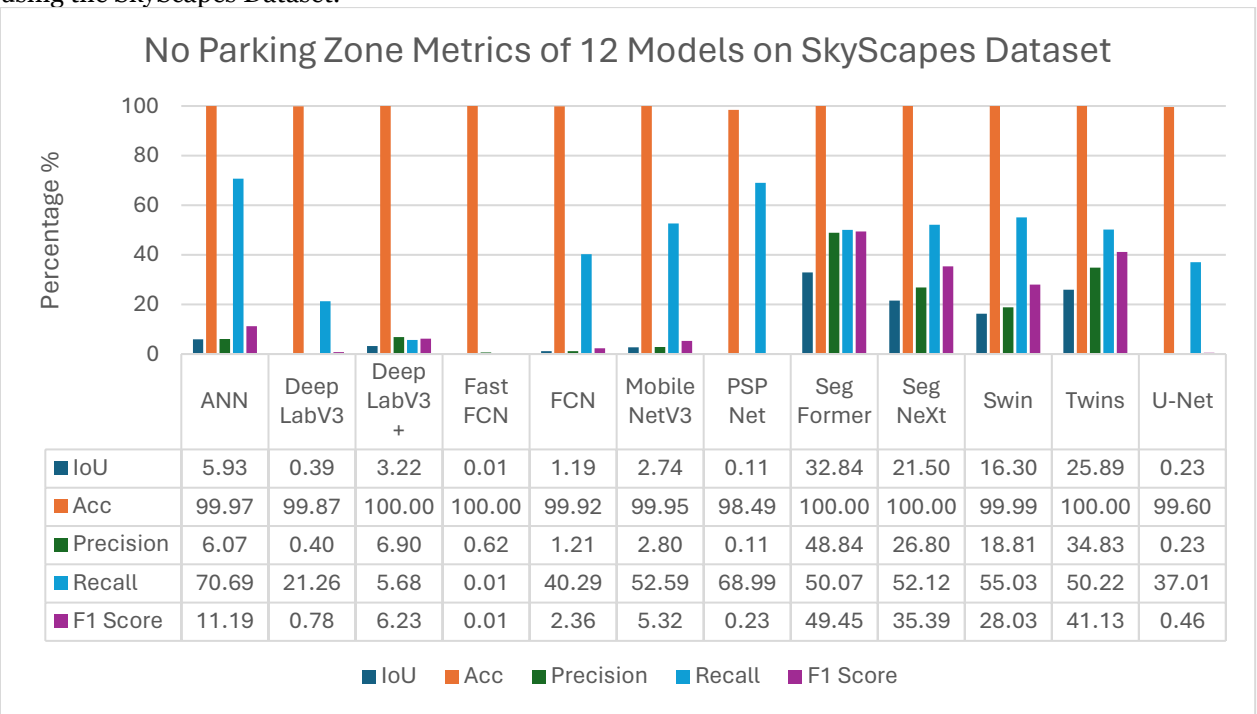
	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	12.47	7.91	24.19	19.36	7.11	13.69	5.14	51.19	34.71	36.61	40.24	26.83
Acc	98.95	98.80	99.74	99.53	98.72	99.22	98.94	99.88	99.74	99.73	99.78	99.79
Precision	12.92	8.49	35.83	22.43	7.63	14.79	5.84	67.33	39.87	40.05	45.03	43.78
Recall	78.43	53.87	42.66	58.51	50.99	64.67	29.95	68.09	72.85	81.01	79.09	40.93
F1 Score	22.18	14.66	38.95	32.43	13.27	24.08	9.78	67.71	51.54	53.60	57.39	42.30

Table 4.5: Evaluation metrics for 12 different models on the performance of long line extraction using the SkyScapes Dataset.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	13.99	10.83	20.90	10.58	8.13	13.21	3.80	41.94	33.29	35.72	29.04	23.64
Acc	98.47	98.93	99.48	99.43	98.00	98.28	98.25	99.71	99.59	99.57	99.41	99.42
Precision	14.56	12.84	28.81	17.48	8.70	13.61	4.40	53.16	40.67	40.65	32.10	29.08
Recall	78.21	40.83	43.21	21.12	55.46	82.00	21.66	66.53	64.70	74.65	75.27	55.81
F1 Score	24.55	19.54	34.57	19.13	15.04	23.34	7.32	59.10	49.95	52.64	45.00	38.24

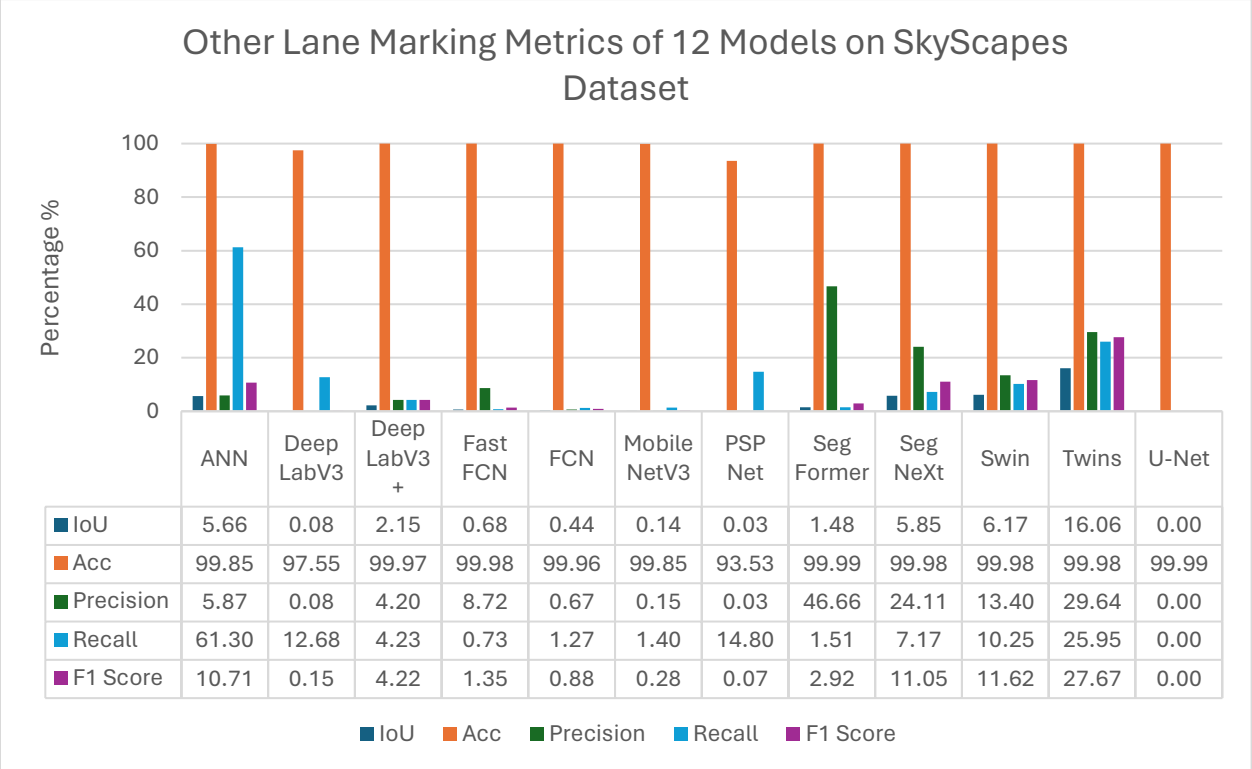
Table 4.6: Evaluation metrics for 12 different models on the performance of no parking zone extraction using the SkyScapes Dataset.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	5.93	0.39	3.22	0.01	1.19	2.74	0.11	32.84	21.50	16.30	25.89	0.23
Acc	99.97	99.87	100.00	100.00	99.92	99.95	98.49	100.00	100.00	99.99	100.00	99.60
Precision	6.07	0.40	6.90	0.62	1.21	2.80	0.11	48.84	26.80	18.81	34.83	0.23
Recall	70.69	21.26	5.68	0.01	40.29	52.59	68.99	50.07	52.12	55.03	50.22	37.01
F1 Score	11.19	0.78	6.23	0.01	2.36	5.32	0.23	49.45	35.39	28.03	41.13	0.46

Table 4.6 delves into the performance metrics for detecting no parking zones, revealing a pattern consistent with that observed in the detection of dash lines, although with a notable deviation where U-Net exhibits a diminished performance in this category. Moving forward to Table 4.7, which outlines the performance metrics for various other lane marking classes, a parallel trend to that observed with dash lines emerges, with Twins models outperforming others to secure the top position in overall performance. Furthermore, Table 4.8 focuses on the performance metrics for the 'other sign' category, highlighting SegNeXt as the leading model among those based on transformer architectures. Remarkably, ANN models also demonstrate formidable performance, with their metrics nearly matching, and in some cases closely competing, those achieved by transformer-based models, illustrating their efficacy and competitive edge in this domain.

Table 4.7: Evaluation metrics for 12 different models on the performance of other lane marking extraction using the SkyScapes Dataset.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	5.66	0.08	2.15	0.68	0.44	0.14	0.03	1.48	5.85	6.17	16.06	0.00
Acc	99.85	97.55	99.97	99.98	99.96	99.85	93.53	99.99	99.98	99.98	99.98	99.99
Precision	5.87	0.08	4.20	8.72	0.67	0.15	0.03	46.66	24.11	13.40	29.64	0.00
Recall	61.30	12.68	4.23	0.73	1.27	1.40	14.80	1.51	7.17	10.25	25.95	0.00
F1 Score	10.71	0.15	4.22	1.35	0.88	0.28	0.07	2.92	11.05	11.62	27.67	0.00

Table 4.8: Evaluation metrics for 12 different models on the performance of other signs extraction using the SkyScapes Dataset.

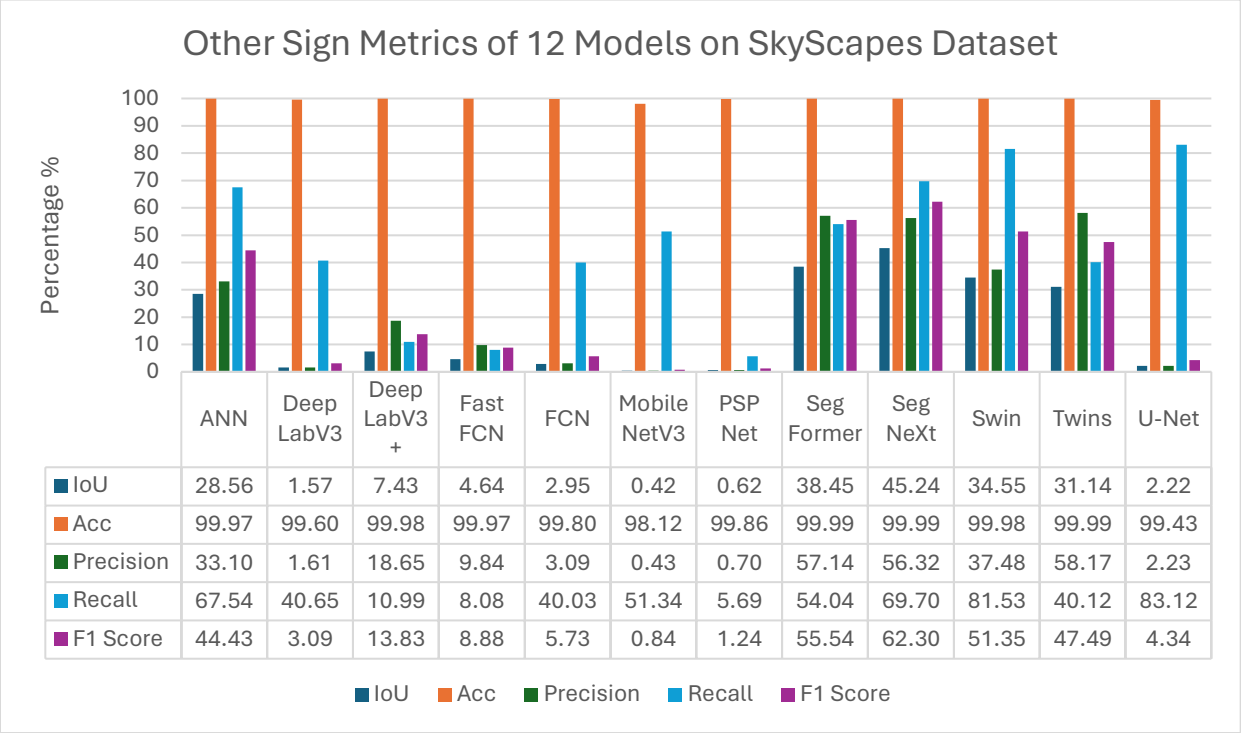


Table 4.9: Evaluation metrics for 12 different models on the performance of parking space extraction using the SkyScapes Dataset.

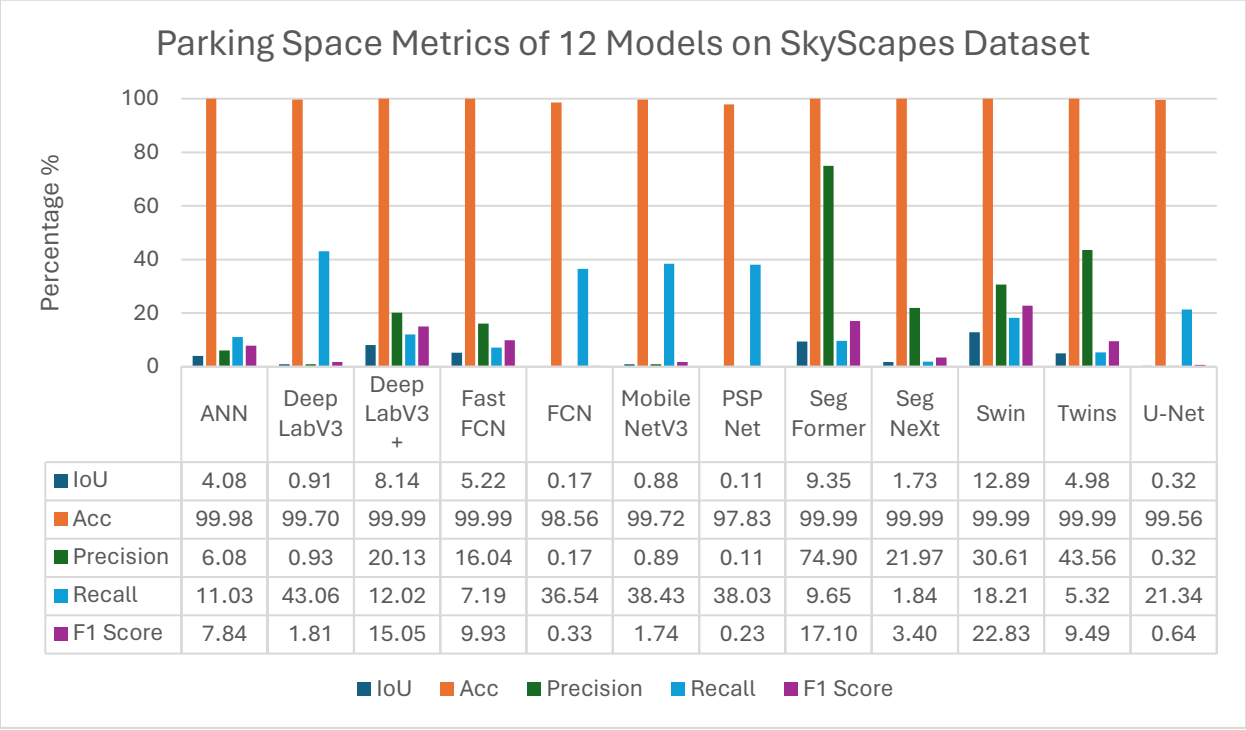
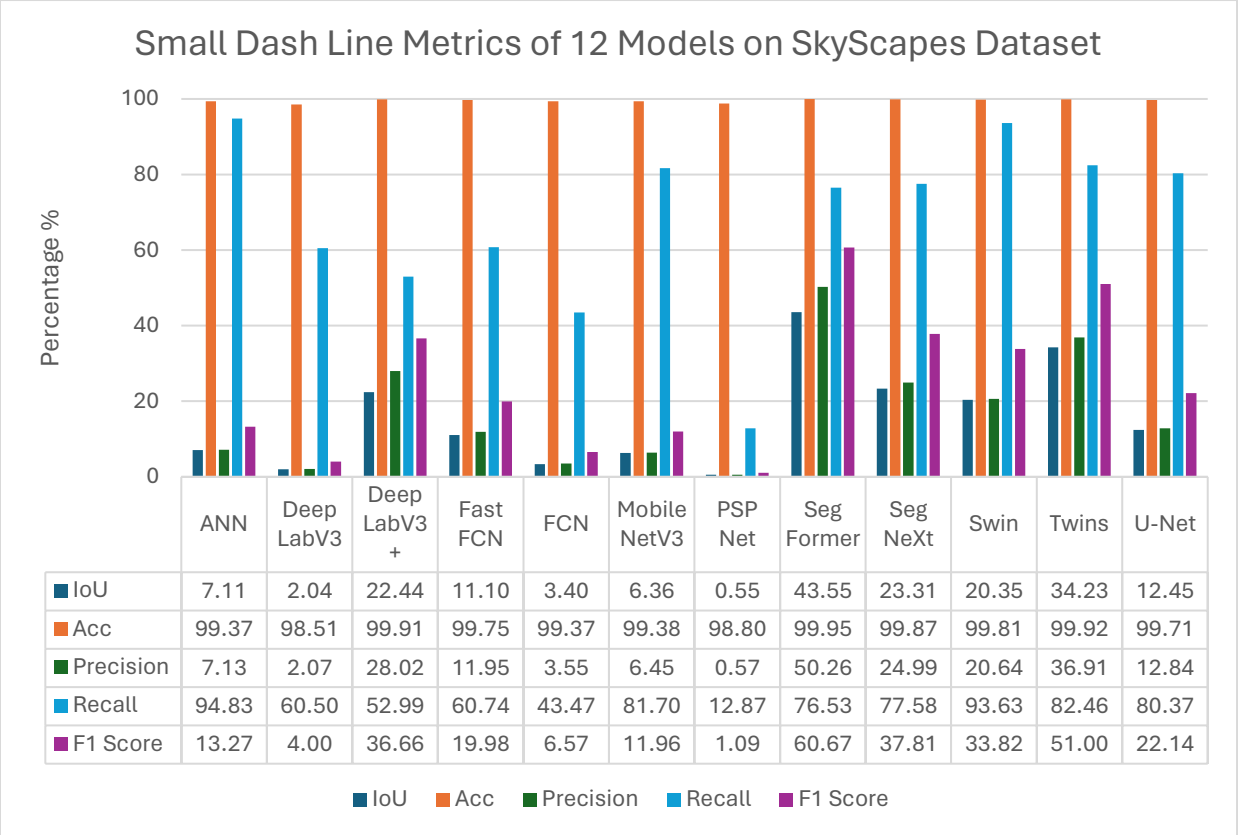


Table 4.9 focusing on the assessment metrics for parking space, reveals that transformer-based models significantly outperform others in terms of precision, where they secure the highest figures compared to all other models. In contrast, select CNN-based models distinguish themselves by obtaining higher recall values. Moving to Table 4.10, which examines the evaluation metrics for small dashed lines, it is observed that PSPNet registers the lowest recall rate, marked at 12.87%, a figure significantly below the next lowest recall value of 43.47% recorded by other models. In this context, SegFormer emerges as the leader in terms of scoring, with DeeplabV3+, a model rooted in CNN technology, trailing closely behind, showcasing competitive performance.

Table 4.10: Evaluation metrics for 12 different models on the performance of small dash line extraction using the SkyScapes Dataset.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	7.11	2.04	22.44	11.10	3.40	6.36	0.55	43.55	23.31	20.35	34.23	12.45
Acc	99.37	98.51	99.91	99.75	99.37	99.38	98.80	99.95	99.87	99.81	99.92	99.71
Precision	7.13	2.07	28.02	11.95	3.55	6.45	0.57	50.26	24.99	20.64	36.91	12.84
Recall	94.83	60.50	52.99	60.74	43.47	81.70	12.87	76.53	77.58	93.63	82.46	80.37
F1 Score	13.27	4.00	36.66	19.98	6.57	11.96	1.09	60.67	37.81	33.82	51.00	22.14

Table 4.11: Evaluation metrics for 12 different models on the performance of stop line extraction using the SkyScapes Dataset.

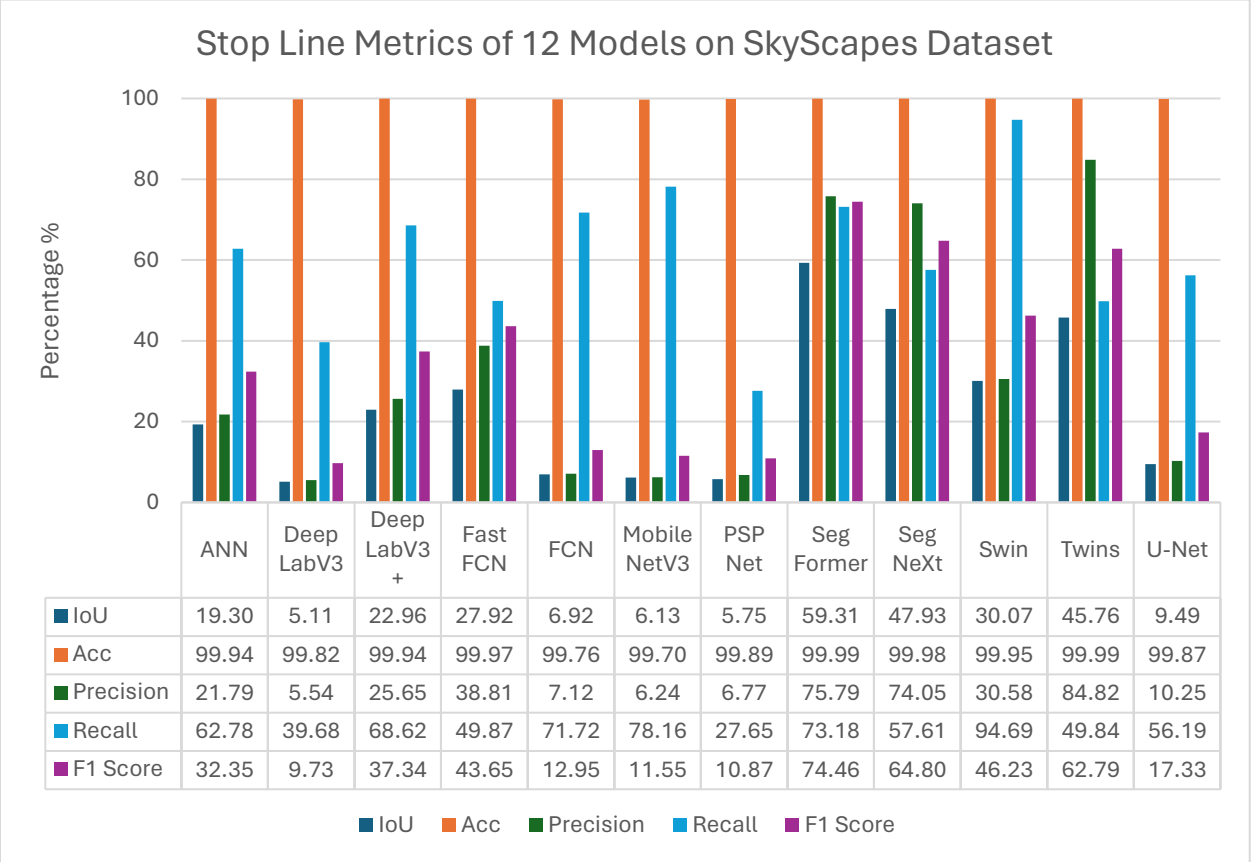
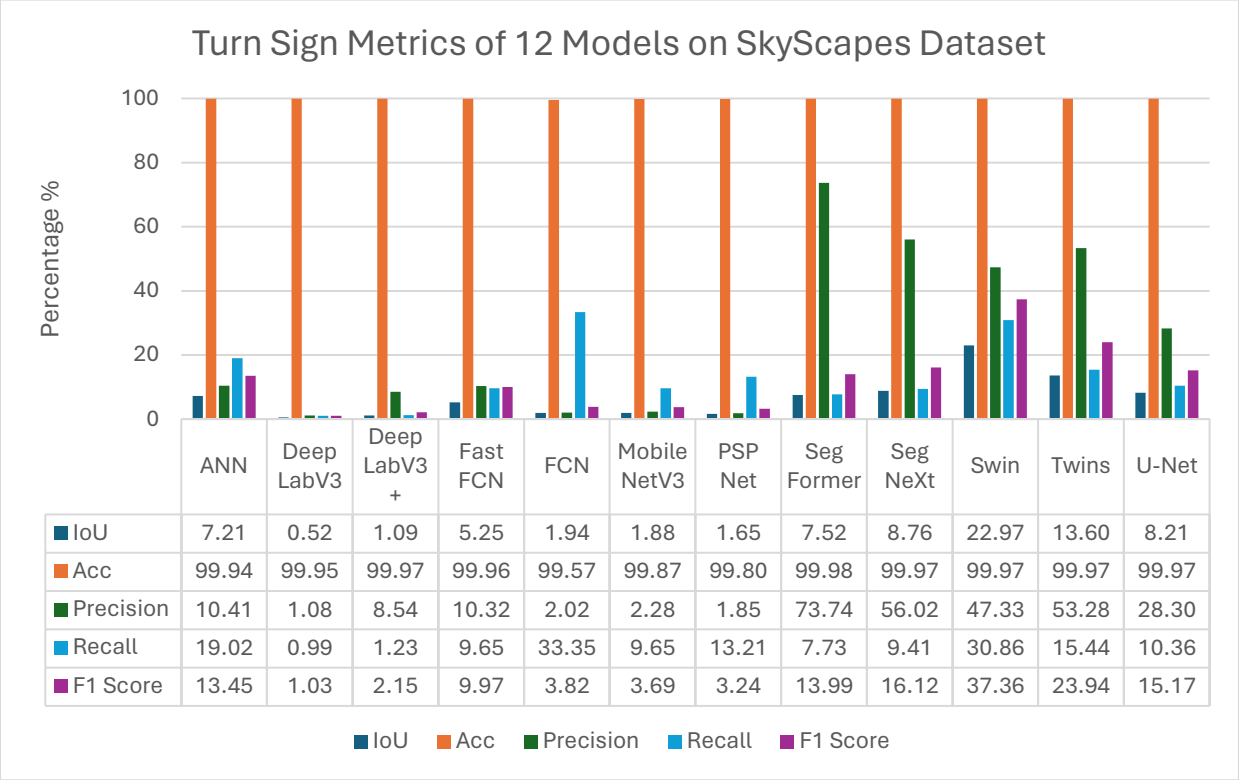


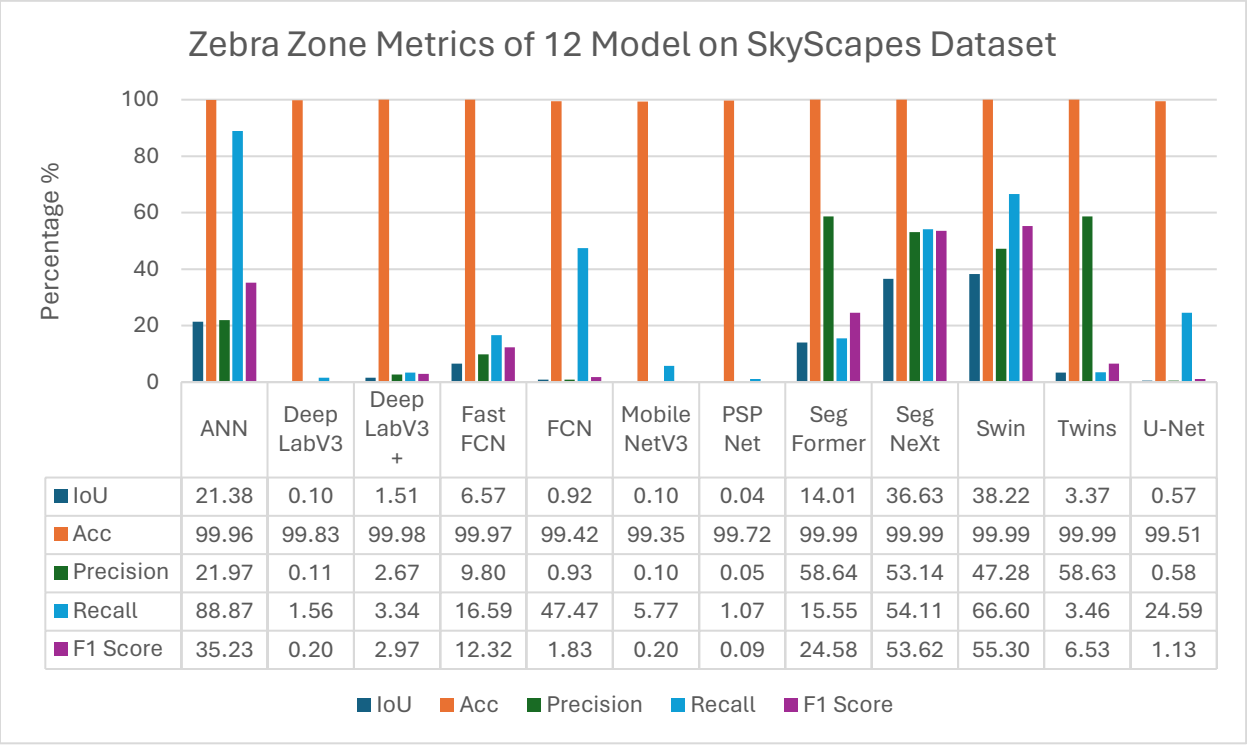
Table 4.11, presenting the evaluation metrics for stop line detection, indicates that models incorporating attention mechanisms consistently lead in performance. Notably, FastFCN achieves a ranking above DeeplabV3+, defying initial expectations. In Table 4.12, which examines the metrics for turn sign detection, it is evident that models leveraging attention mechanisms, along with U-Net, markedly surpass other models in precision. Swin distinguishes itself as the model with the highest comprehensive scores. Furthermore, Table 4.13, which outlines the metrics for zebra zone detection, demonstrates that SegNeXt and Swin perform similarly, with SegNeXt delivering a more evenly balanced performance. Twins are observed to be considerably less effective in comparison, whereas ANN secures the third position, surpassing SegFormer in performance.

Table 4.12: Evaluation metrics for 12 different models on the performance of turn sign extraction using the SkyScapes Dataset.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	7.21	0.52	1.09	5.25	1.94	1.88	1.65	7.52	8.76	22.97	13.60	8.21
Acc	99.94	99.95	99.97	99.96	99.57	99.87	99.80	99.98	99.97	99.97	99.97	99.97
Precision	10.41	1.08	8.54	10.32	2.02	2.28	1.85	73.74	56.02	47.33	53.28	28.30
Recall	19.02	0.99	1.23	9.65	33.35	9.65	13.21	7.73	9.41	30.86	15.44	10.36
F1 Score	13.45	1.03	2.15	9.97	3.82	3.69	3.24	13.99	16.12	37.36	23.94	15.17

Table 4.13: Evaluation metrics for 12 different models on the performance of zebra zone extraction using the SkyScapes Dataset.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	21.38	0.10	1.51	6.57	0.92	0.10	0.04	14.01	36.63	38.22	3.37	0.57
Acc	99.96	99.83	99.98	99.97	99.42	99.35	99.72	99.99	99.99	99.99	99.99	99.51
Precision	21.97	0.11	2.67	9.80	0.93	0.10	0.05	58.64	53.14	47.28	58.63	0.58
Recall	88.87	1.56	3.34	16.59	47.47	5.77	1.07	15.55	54.11	66.60	3.46	24.59
F1 Score	35.23	0.20	2.97	12.32	1.83	0.20	0.09	24.58	53.62	55.30	6.53	1.13

In conclusion, examining 12 models on the SkyScapes dataset provides valuable insights into how transformer-based and CNN-based models perform across various metrics. The analysis showed that predicting the background class was generally the easiest task for all models, mainly because it makes up most pixels. However, PSPNet and MobileNetV3 stood out for their lower mean IoU scores for the background, suggesting they might incorrectly classify more pixels as belonging to other classes. The study also revealed differences in how models perform in specific classes, which can be traced back to their architecture, the settings of their optimizers, and their loss functions. These results highlight the critical role of careful model selection and adjustment to meet the demands of specific tasks and the distinct advantages and challenges presented by different model architectures. The detailed examination of these differences and their reasons will be further discussed in the discussion section.

4.1.2 Waterloo Urban Scene Dataset

This section extends the comparative analysis of model performance from the SkyScapes dataset to the Waterloo Urban Scene dataset, aiming to understand how different models perform across diverse urban imaging domains. Employing uniform evaluation metrics, such as mIoU, mAcc, among others, the study presents tables showcasing the performance of 12 models based on these criteria. Furthermore, it elucidates and emphasizes notable outcomes derived from class-based table analyses of certain models, offering insights into their performance across specific classes.

The results section from Table 4.14 below showcases a notable improvement across all metrics on the Waterloo Urban Scene dataset, with mIoU scores now ranging between 33.56% to 76.11% and F1 scores spanning from 44.34% to 85.35%. This marks a significant enhancement in model performance compared to previous benchmarks, potentially attributed to the advantages of pretraining on the SkyScapes dataset and the unique characteristics of the Waterloo Urban Scene dataset itself.

Table 4.14: Benchmark of the state of the art on the Waterloo Urban Scene Dataset over all 15 classes (in %).

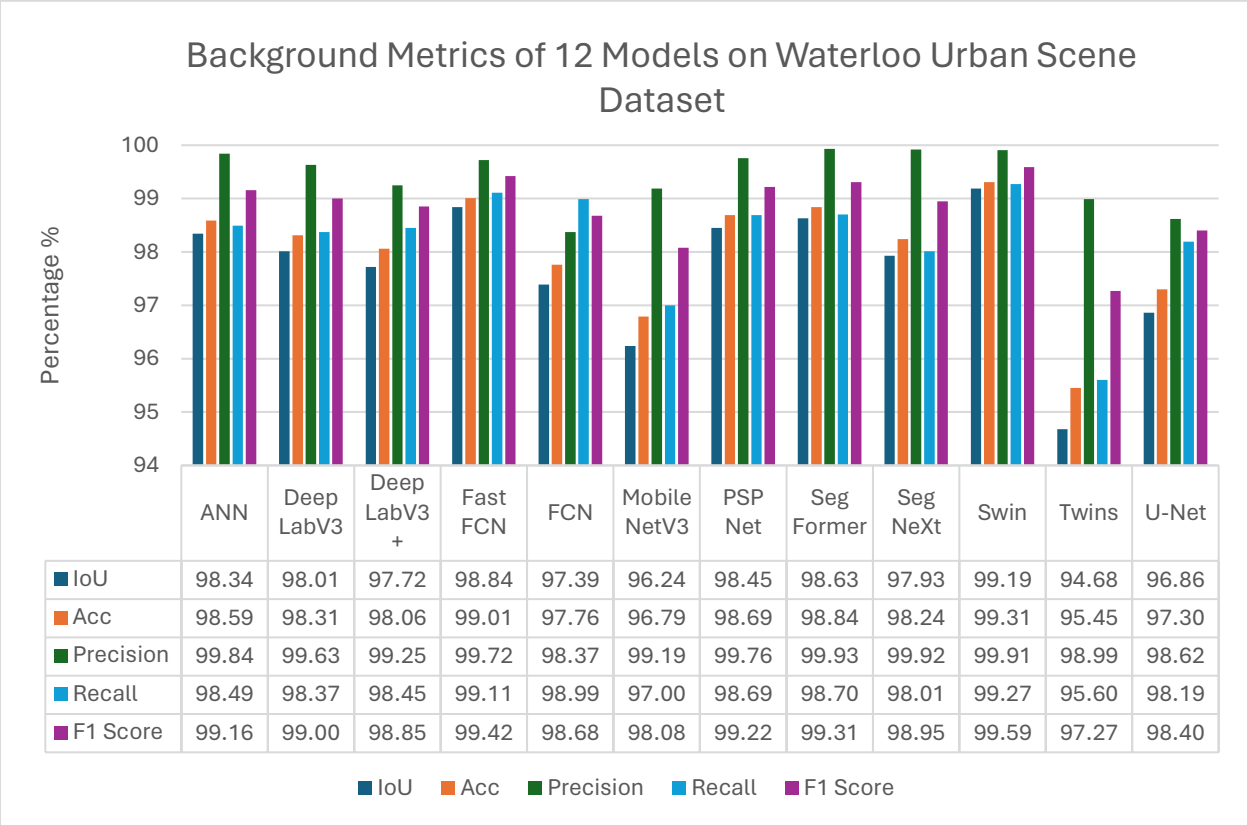
Method	Base	IoU	Accuracy		Confusion		F1
		mean	mAcc	aAcc	mRecall	mPrecision	
FCN	ResNet101	44.40	99.58	96.84	93.55	47.06	56.38
FastFCN	ResNet50	50.48	99.76	98.16	98.54	51.46	62.07
U-Net	-	44.64	99.52	96.37	91.14	46.95	56.62
DeepLabV3	ResNet101	47.55	99.63	97.22	93.97	49.44	60.31
DeepLabV3+	ResNet101	51.03	99.65	97.34	95.34	53.2	63.74
ANN	ResNet101	46.5	99.67	97.54	96.30	47.91	58.23
MobileNetV3	Large	34.94	99.37	95.26	89.89	37.44	46.03
PSPNet	ResNet101	50.48	99.72	97.87	96.73	51.90	62.29
SegNeXt	Base	65.77	99.66	97.45	96.81	67.24	77.60
Twins	Base	62.86	99.23	94.24	88.15	68.64	76.01
Swin	Base	60.22	**99.84**	**98.78**	**98.96**	60.84	72.48
SegFormer	Base	**76.11**	99.77	98.27	97.94	**77.44**	**85.35**

In line with the trends observed on the SkyScapes dataset, transformer-based models continue to outperform traditional CNN-based models on the Waterloo Urban Scene dataset based on Table 4.14. Specifically, SegNeXt, with its CNN attention mechanism, achieves the second highest mIoU, trailing only behind SegFormer. SegFormer leads in mIoU, mean precision, and F1 score, with impressive scores of 76.11%, 77.44%, and 85.34%, respectively. Swin, on the other hand, excels in mAcc, overall accuracy, and mean recall, recording the highest values at 99.84%, 98.74%, and 98.96%, respectively. Most models demonstrate exceptionally high recall values, exceeding 90%, with the exception of MobileNetV3 and Twins, which record slightly lower recalls at 89.89% and 88.15%, respectively. This trend suggests that models, in general, tend to predict more pixels outside the actual ground truth area than fewer pixels within it, as evidenced by the lower precision scores compared to recall scores.

Drawing on the insights from the SkyScapes dataset, the analysis of model performance on the Waterloo Urban Scene dataset reveals distinct trends. Detailed per-model per-class results can be found in appendix A.2. Contrary to the findings from the SkyScapes dataset, the analysis of models on the Waterloo Urban Scene dataset reveals enhanced performance in the sequence of Road, Traffic Island, Sidewalk, and Vehicle classes. This

shift in performance hierarchy suggests a different weighting of class importance within the Waterloo Urban Scene dataset, with these four classes playing a more significant role in overall model evaluation metrics.

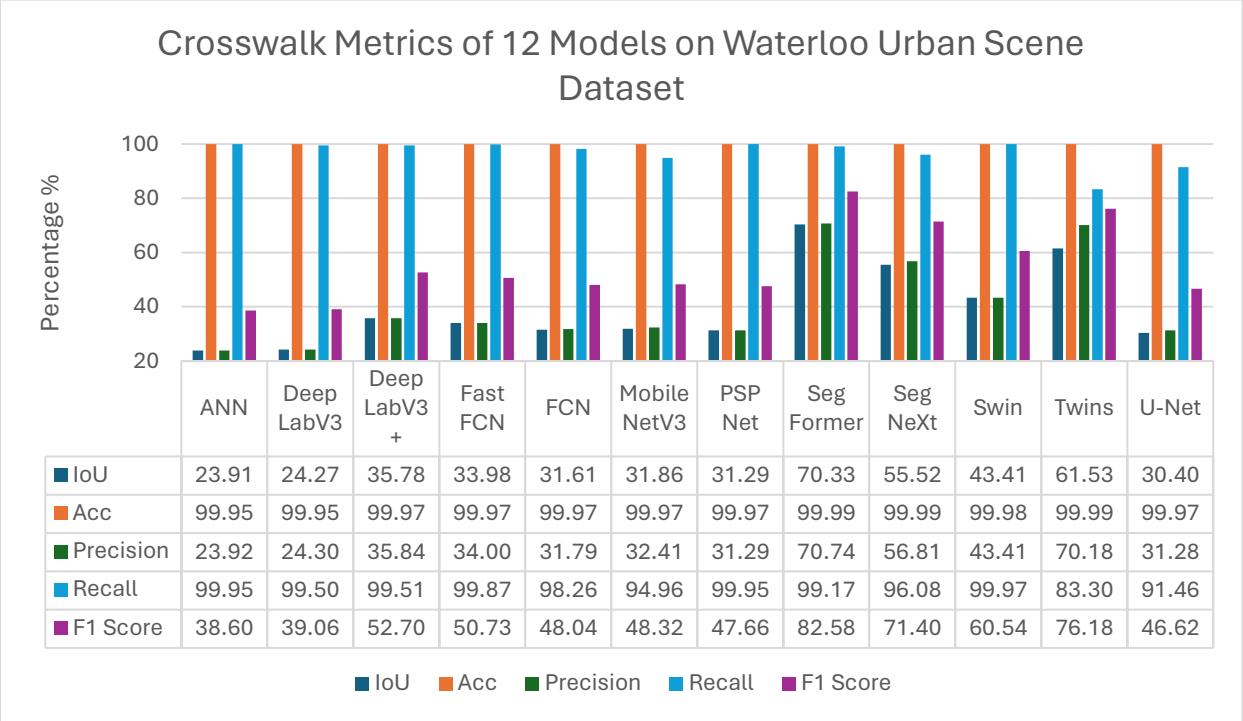
Table 4.15: Evaluation metrics for 12 different models on the performance of background extraction using the Waterloo Urban Scene Dataset after fine-tuning.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	98.34	98.01	97.72	98.84	97.39	96.24	98.45	98.63	97.93	99.19	94.68	96.86
Acc	98.59	98.31	98.06	99.01	97.76	96.79	98.69	98.84	98.24	99.31	95.45	97.30
Precision	99.84	99.63	99.25	99.72	98.37	99.19	99.76	99.93	99.92	99.91	98.99	98.62
Recall	98.49	98.37	98.45	99.11	98.99	97.00	98.69	98.70	98.01	99.27	95.60	98.19
F1 Score	99.16	99.00	98.85	99.42	98.68	98.08	99.22	99.31	98.95	99.59	97.27	98.40

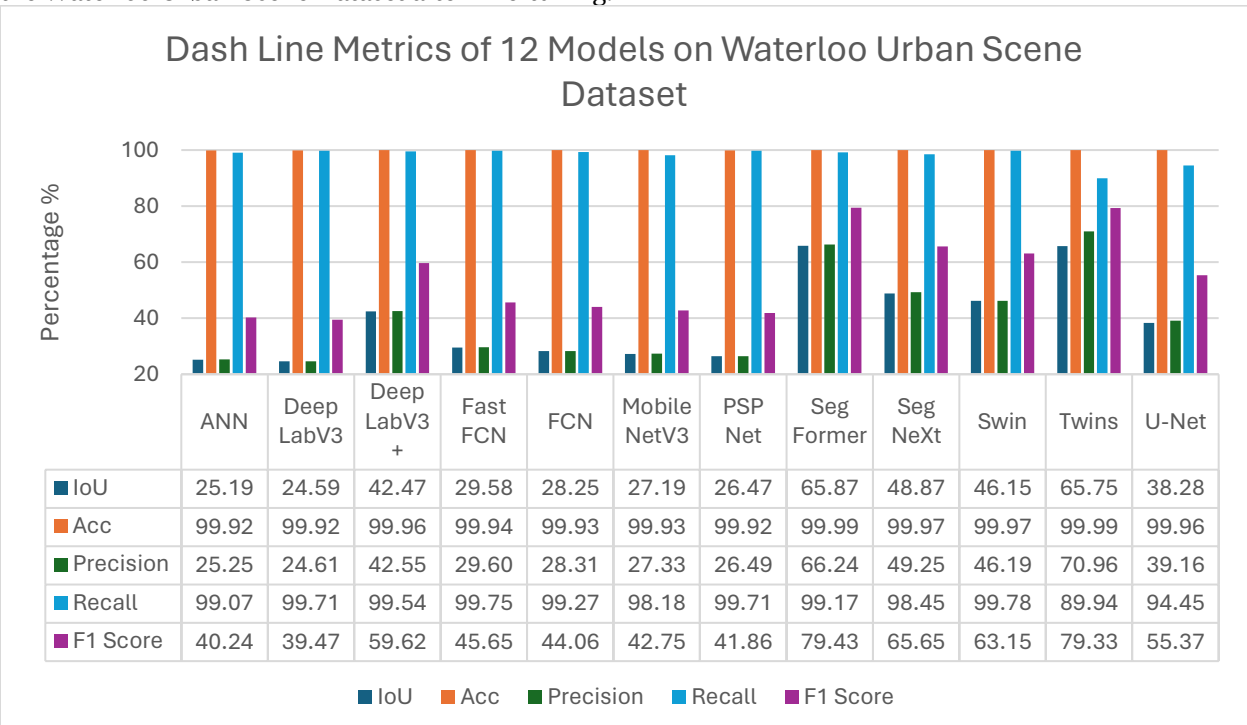
Referring to Table 4.15, which outlines the background metrics, it is observed that all models report satisfactory outcomes, except for Twins. This model exhibits noticeably poorer performance across all metrics compared to its counterparts. Table 4.16, detailing the evaluation metrics for crosswalks, indicates that attention-based models achieve superior outcomes, although Swin ranks as the least effective among these. Traditional CNN-based models display comparable performance. Notably, the recall rates across all models are significantly higher relative to other classes, a trend also observed in the crosswalk category of skyscape datasets. This could be attributed to the challenges posed by thin and densely packed dashed lines, leading to a tendency for models to overpredict.

Table 4.16: Evaluation metrics for 12 different models on the performance of crosswalk extraction using the Waterloo Urban Scene Dataset after fine-tuning.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	23.91	24.27	35.78	33.98	31.61	31.86	31.29	70.33	55.52	43.41	61.53	30.40
Acc	99.95	99.95	99.97	99.97	99.97	99.97	99.97	99.99	99.99	99.98	99.99	99.97
Precision	23.92	24.30	35.84	34.00	31.79	32.41	31.29	70.74	56.81	43.41	70.18	31.28
Recall	99.95	99.50	99.51	99.87	98.26	94.96	99.95	99.17	96.08	99.97	83.30	91.46
F1 Score	38.60	39.06	52.70	50.73	48.04	48.32	47.66	82.58	71.40	60.54	76.18	46.62

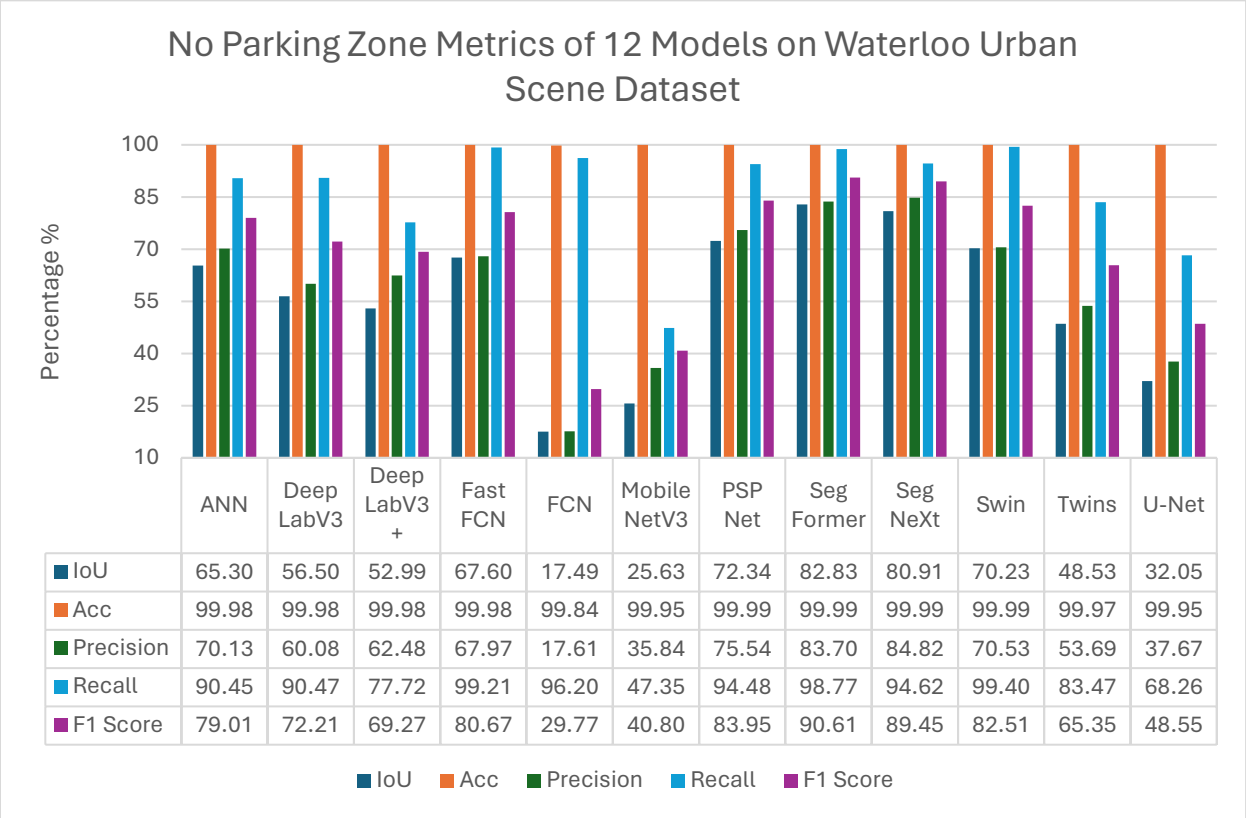
Table 4.17: Evaluation metrics for 12 different models on the performance of dash line extraction using the Waterloo Urban Scene Dataset after fine-tuning.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	25.19	24.59	42.47	29.58	28.25	27.19	26.47	65.87	48.87	46.15	65.75	38.28
Acc	99.92	99.92	99.96	99.94	99.93	99.93	99.92	99.99	99.97	99.97	99.99	99.96
Precision	25.25	24.61	42.55	29.60	28.31	27.33	26.49	66.24	49.25	46.19	70.96	39.16
Recall	99.07	99.71	99.54	99.75	99.27	98.18	99.71	99.17	98.45	99.78	89.94	94.45
F1 Score	40.24	39.47	59.62	45.65	44.06	42.75	41.86	79.43	65.65	63.15	79.33	55.37

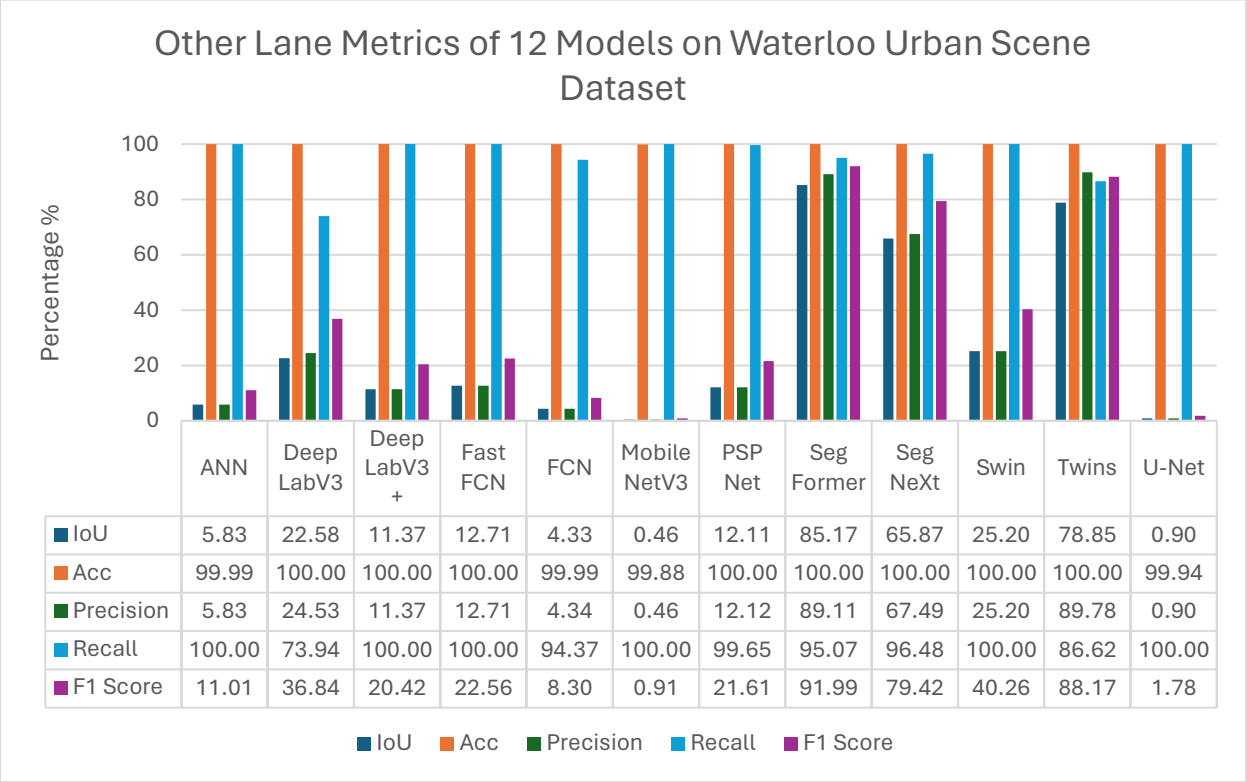
In Table 4.17, which shows the metrics for dash line detection, a consistent pattern emerges where models exhibit high recall rates, a trend that aligns with their performance in smaller classes. This suggests that the loss function might be influencing models to prioritize extensive coverage, possibly at the expense of precision. Moving on to Table 4.18, an examination of the no parking zone metrics displays uniform trends across models, despite differences in their scores. In this category, FCN, MobileNetV3, and U-Net are identified as the models with the lowest performance, whereas SegFormer, SegNeXt, and PSPNet stand out as the top performers. Lastly, Table 4.19, which focuses on metrics for other lane classes, indicates that models leveraging attention mechanisms once again achieve the highest scores. Contrary to expectations, Swin performs comparably to DeeplabV3, the latter being a prominent traditional CNN-based model and illustrating a deviation in performance for Swin from its anticipated outcomes.

Table 4.18: Evaluation metrics for 12 different models on the performance of no parking zone extraction using the Waterloo Urban Scene Dataset after fine-tuning.



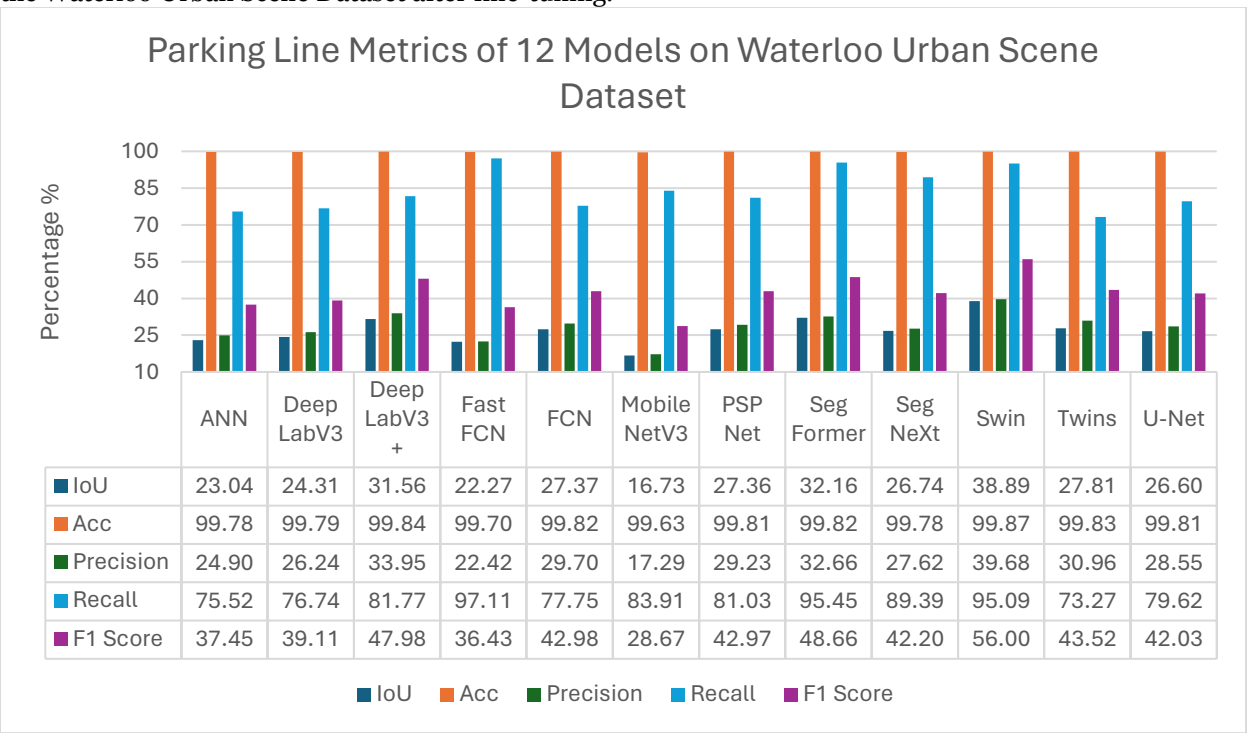
	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	65.30	56.50	52.99	67.60	17.49	25.63	72.34	82.83	80.91	70.23	48.53	32.05
Acc	99.98	99.98	99.98	99.98	99.84	99.95	99.99	99.99	99.99	99.99	99.97	99.95
Precision	70.13	60.08	62.48	67.97	17.61	35.84	75.54	83.70	84.82	70.53	53.69	37.67
Recall	90.45	90.47	77.72	99.21	96.20	47.35	94.48	98.77	94.62	99.40	83.47	68.26
F1 Score	79.01	72.21	69.27	80.67	29.77	40.80	83.95	90.61	89.45	82.51	65.35	48.55

Table 4.19: Evaluation metrics for 12 different models on the performance of other lane marking extraction using the Waterloo Urban Scene Dataset after fine-tuning.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	5.83	22.58	11.37	12.71	4.33	0.46	12.11	85.17	65.87	25.20	78.85	0.90
Acc	99.99	100.00	100.00	100.00	99.99	99.88	100.00	100.00	100.00	100.00	100.00	99.94
Precision	5.83	24.53	11.37	12.71	4.34	0.46	12.12	89.11	67.49	25.20	89.78	0.90
Recall	100.00	73.94	100.00	100.00	94.37	100.00	99.65	95.07	96.48	100.00	86.62	100.00
F1 Score	11.01	36.84	20.42	22.56	8.30	0.91	21.61	91.99	79.42	40.26	88.17	1.78

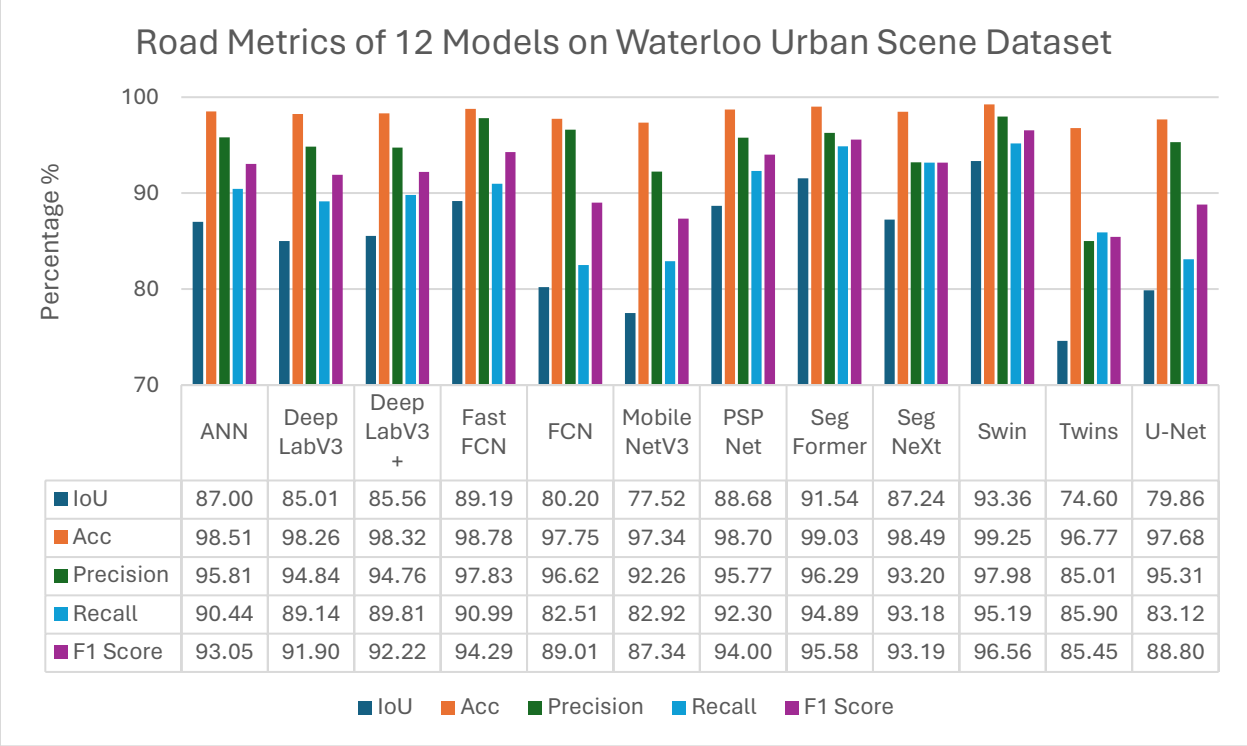
Table 4.20: Evaluation metrics for 12 different models on the performance of parking line extraction using the Waterloo Urban Scene Dataset after fine-tuning.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	23.04	24.31	31.56	22.27	27.37	16.73	27.36	32.16	26.74	38.89	27.81	26.60
Acc	99.78	99.79	99.84	99.70	99.82	99.63	99.81	99.82	99.78	99.87	99.83	99.81
Precision	24.90	26.24	33.95	22.42	29.70	17.29	29.23	32.66	27.62	39.68	30.96	28.55
Recall	75.52	76.74	81.77	97.11	77.75	83.91	81.03	95.45	89.39	95.09	73.27	79.62
F1 Score	37.45	39.11	47.98	36.43	42.98	28.67	42.97	48.66	42.20	56.00	43.52	42.03

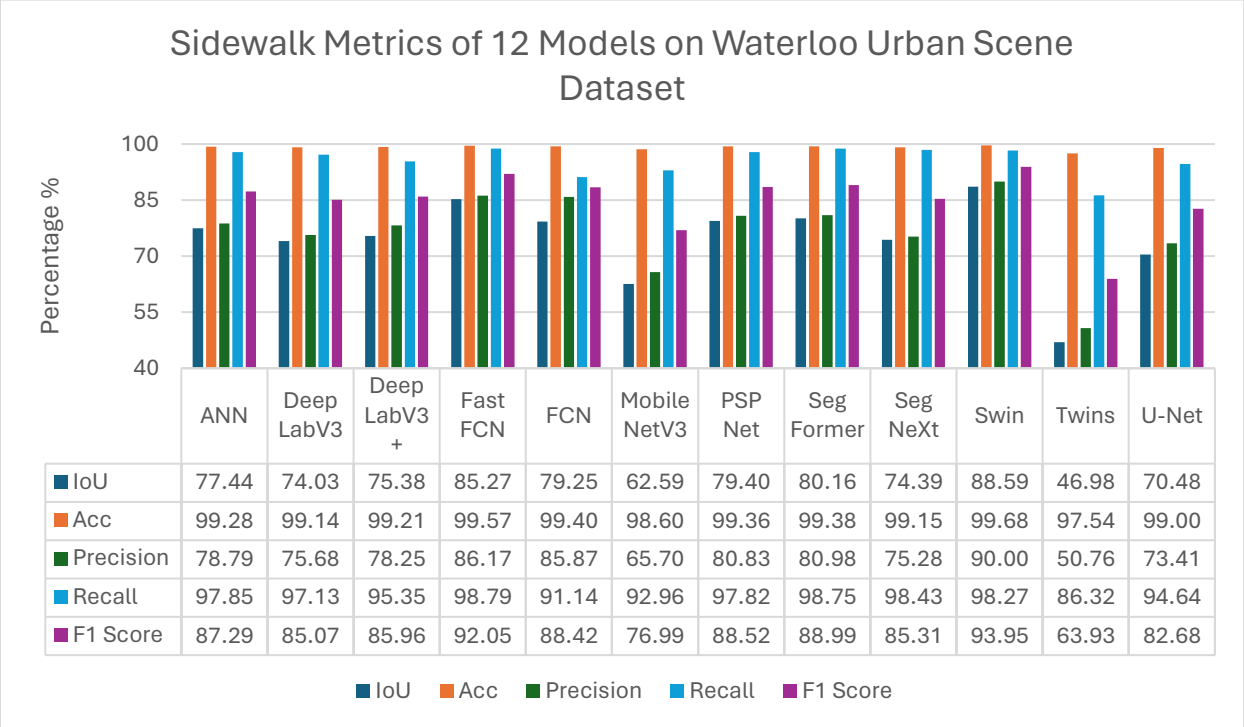
In Table 4.20, which assesses the performance metrics for parking lines, a uniform level of achievement is observed across all models, each characterized by exceptionally high recall rates. Moving to Table 4.21, the evaluation of road metrics demonstrates a uniformity in performance among all models as well, with each model exhibiting remarkably high precision values. This elevated precision suggests that model predictions are generally more cautious, often erring on the side of underestimating compared to the actual ground truth. Table 4.22, focusing on the evaluation of sidewalk metrics, reveals that the performances of the models are largely comparable and well-balanced, with the notable exception of the Twins model, which performs significantly worse than the rest.

Table 4.21: Evaluation metrics for 12 different models on the performance of road extraction using the Waterloo Urban Scene Dataset after fine-tuning.



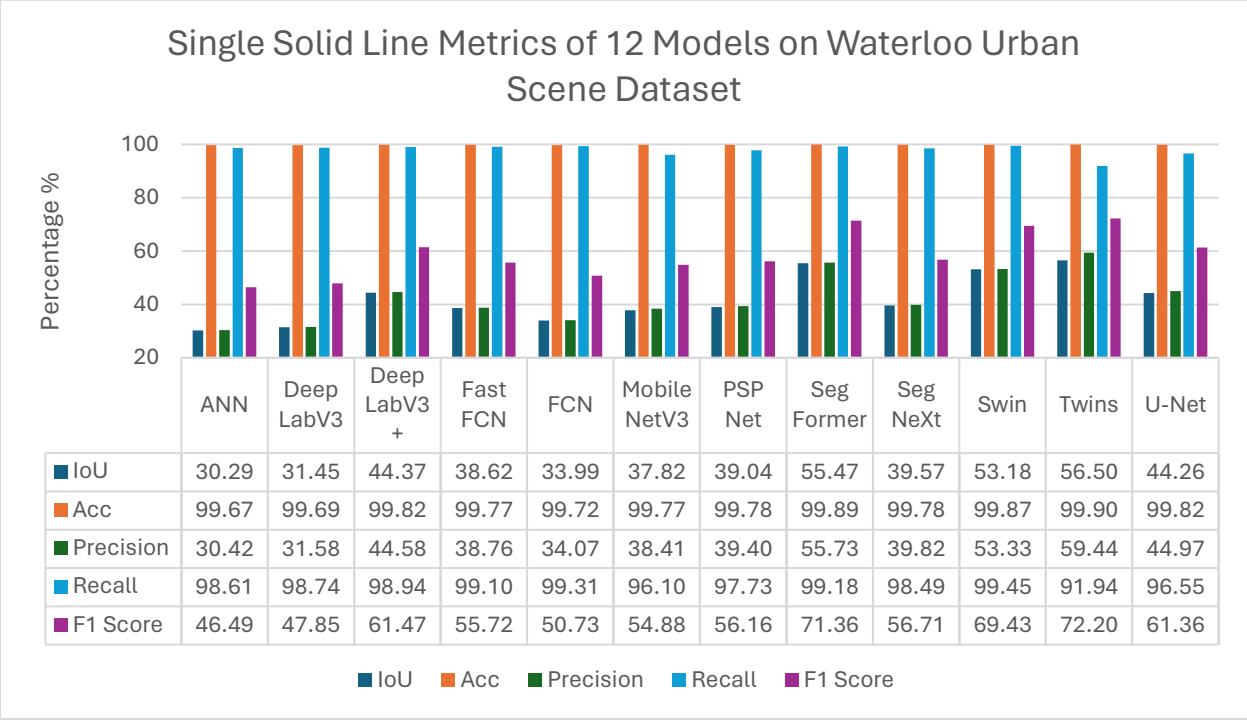
	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	87.00	85.01	85.56	89.19	80.20	77.52	88.68	91.54	87.24	93.36	74.60	79.86
Acc	98.51	98.26	98.32	98.78	97.75	97.34	98.70	99.03	98.49	99.25	96.77	97.68
Precision	95.81	94.84	94.76	97.83	96.62	92.26	95.77	96.29	93.20	97.98	85.01	95.31
Recall	90.44	89.14	89.81	90.99	82.51	82.92	92.30	94.89	93.18	95.19	85.90	83.12
F1 Score	93.05	91.90	92.22	94.29	89.01	87.34	94.00	95.58	93.19	96.56	85.45	88.80

Table 4.22: Evaluation metrics for 12 different models on the performance of sidewalk extraction using the Waterloo Urban Scene Dataset after fine-tuning.



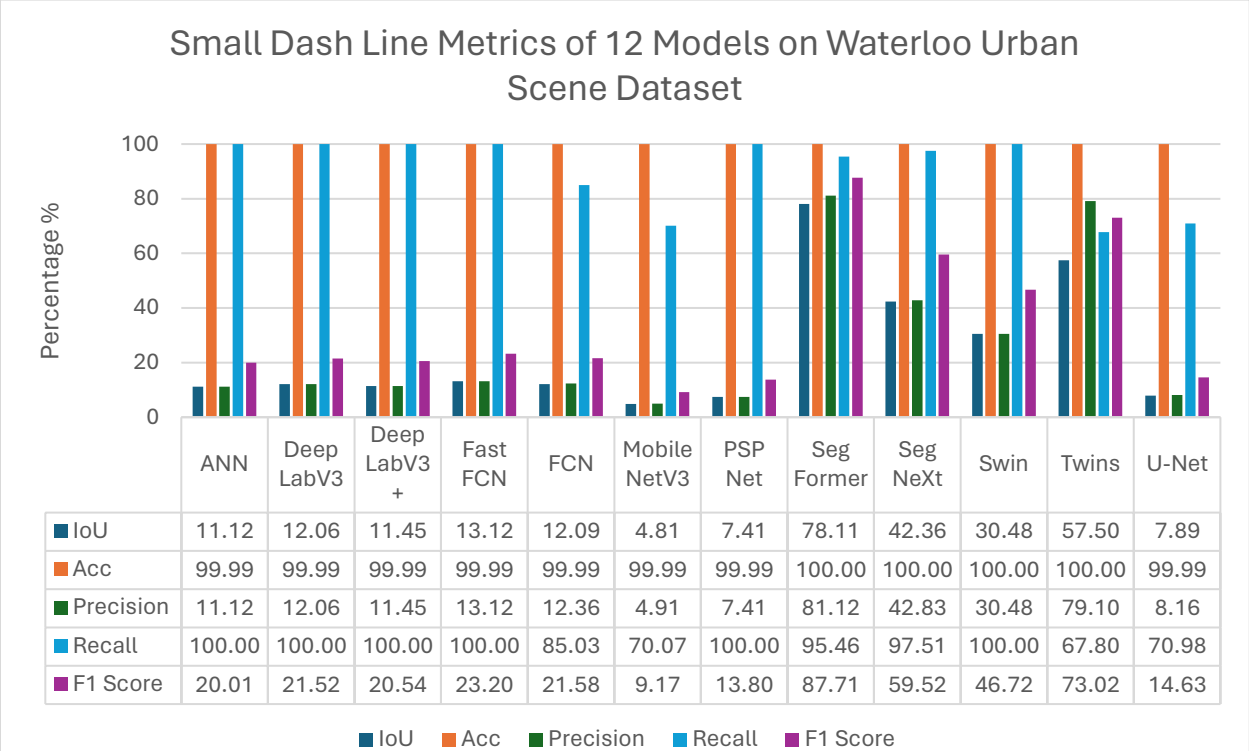
	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	77.44	74.03	75.38	85.27	79.25	62.59	79.40	80.16	74.39	88.59	46.98	70.48
Acc	99.28	99.14	99.21	99.57	99.40	98.60	99.36	99.38	99.15	99.68	97.54	99.00
Precision	78.79	75.68	78.25	86.17	85.87	65.70	80.83	80.98	75.28	90.00	50.76	73.41
Recall	97.85	97.13	95.35	98.79	91.14	92.96	97.82	98.75	98.43	98.27	86.32	94.64
F1 Score	87.29	85.07	85.96	92.05	88.42	76.99	88.52	88.99	85.31	93.95	63.93	82.68

Table 4.23: Evaluation metrics for 12 different models on the performance of single solid line extraction using the Waterloo Urban Scene Dataset after fine-tuning.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	30.29	31.45	44.37	38.62	33.99	37.82	39.04	55.47	39.57	53.18	56.50	44.26
Acc	99.67	99.69	99.82	99.77	99.72	99.77	99.78	99.89	99.78	99.87	99.90	99.82
Precision	30.42	31.58	44.58	38.76	34.07	38.41	39.40	55.73	39.82	53.33	59.44	44.97
Recall	98.61	98.74	98.94	99.10	99.31	96.10	97.73	99.18	98.49	99.45	91.94	96.55
F1 Score	46.49	47.85	61.47	55.72	50.73	54.88	56.16	71.36	56.71	69.43	72.20	61.36

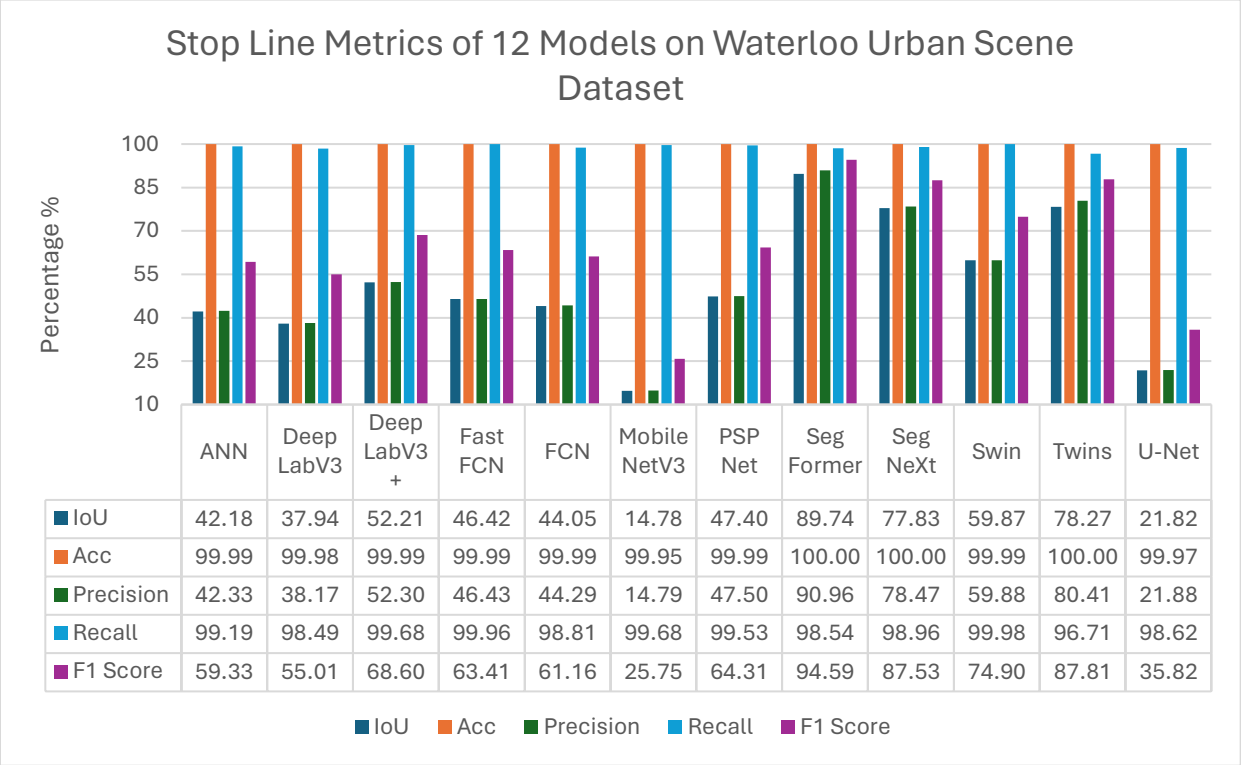
Table 4.24: Evaluation metrics for 12 different models on the performance of small dash line extraction using the Waterloo Urban Scene Dataset after fine-tuning.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	11.12	12.06	11.45	13.12	12.09	4.81	7.41	78.11	42.36	30.48	57.50	7.89
Acc	99.99	99.99	99.99	99.99	99.99	99.99	99.99	100.00	100.00	100.00	100.00	99.99
Precision	11.12	12.06	11.45	13.12	12.36	4.91	7.41	81.12	42.83	30.48	79.10	8.16
Recall	100.00	100.00	100.00	100.00	85.03	70.07	100.00	95.46	97.51	100.00	67.80	70.98
F1 Score	20.01	21.52	20.54	23.20	21.58	9.17	13.80	87.71	59.52	46.72	73.02	14.63

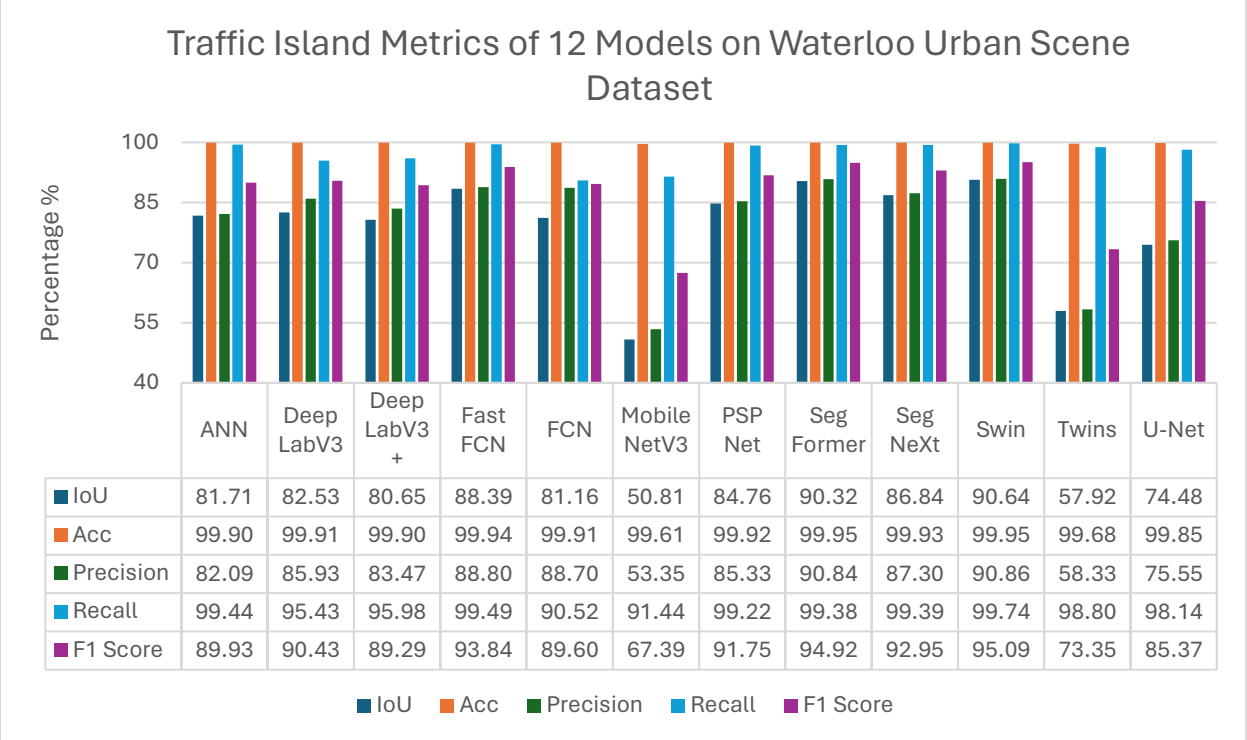
In Table 4.24, the focus is on the metrics for small dashed lines, showcasing that models with attention mechanisms excel, particularly SegFormer, which stands out by achieving outstanding metrics above all others. Transitioning to Table 4.25, which evaluates stop line detection, a pattern emerges where all models demonstrate exceptionally high recall rates. Within this framework, MobileNetV3 is identified as the least effective, with U-Net marginally better in terms of overall performance. This observation is consistent with a broader trend of high recall rates observed across various classes, indicating a general strength in model detection capabilities.

Table 4.25: Evaluation metrics for 12 different models on the performance of stop line extraction using the Waterloo Urban Scene Dataset after fine-tuning.



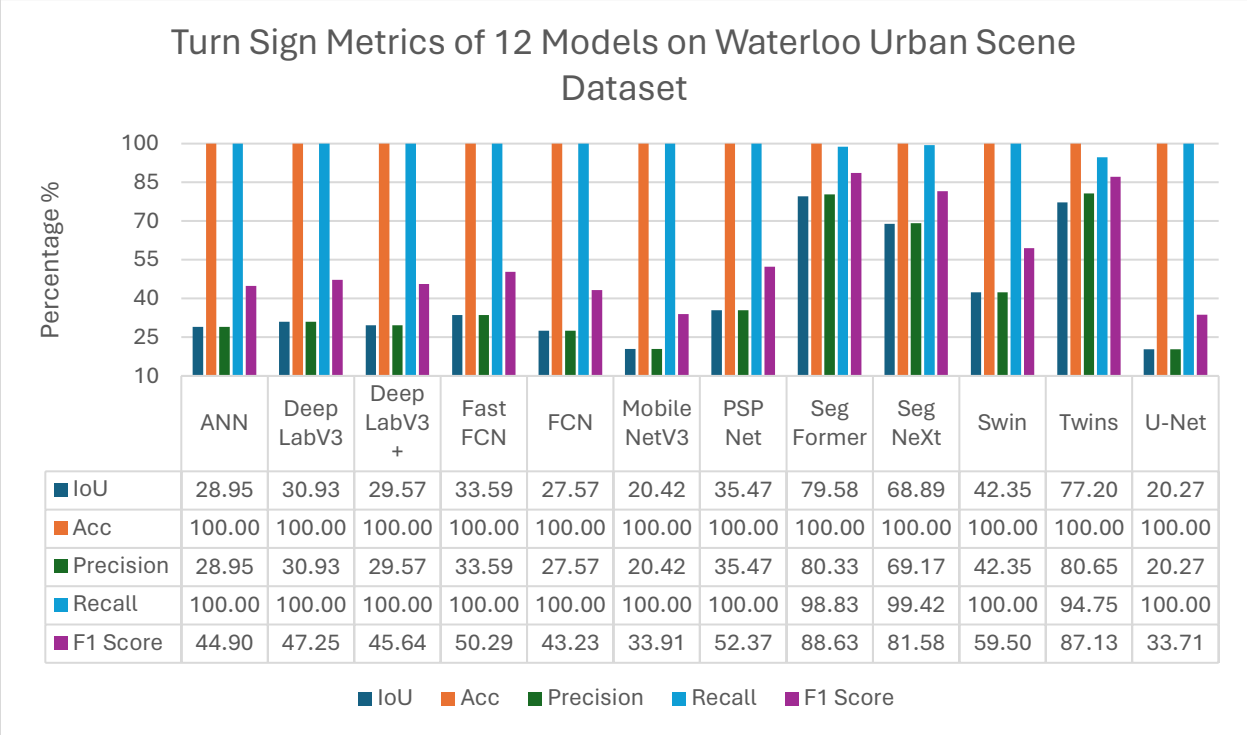
	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	42.18	37.94	52.21	46.42	44.05	14.78	47.40	89.74	77.83	59.87	78.27	21.82
Acc	99.99	99.98	99.99	99.99	99.99	99.95	99.99	100.00	100.00	99.99	100.00	99.97
Precision	42.33	38.17	52.30	46.43	44.29	14.79	47.50	90.96	78.47	59.88	80.41	21.88
Recall	99.19	98.49	99.68	99.96	98.81	99.68	99.53	98.54	98.96	99.98	96.71	98.62
F1 Score	59.33	55.01	68.60	63.41	61.16	25.75	64.31	94.59	87.53	74.90	87.81	35.82

Table 4.26: Evaluation metrics for 12 different models on the performance of traffic island extraction using the Waterloo Urban Scene Dataset after fine-tuning.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	81.71	82.53	80.65	88.39	81.16	50.81	84.76	90.32	86.84	90.64	57.92	74.48
Acc	99.90	99.91	99.90	99.94	99.91	99.61	99.92	99.95	99.93	99.95	99.68	99.85
Precision	82.09	85.93	83.47	88.80	88.70	53.35	85.33	90.84	87.30	90.86	58.33	75.55
Recall	99.44	95.43	95.98	99.49	90.52	91.44	99.22	99.38	99.39	99.74	98.80	98.14
F1 Score	89.93	90.43	89.29	93.84	89.60	67.39	91.75	94.92	92.95	95.09	73.35	85.37

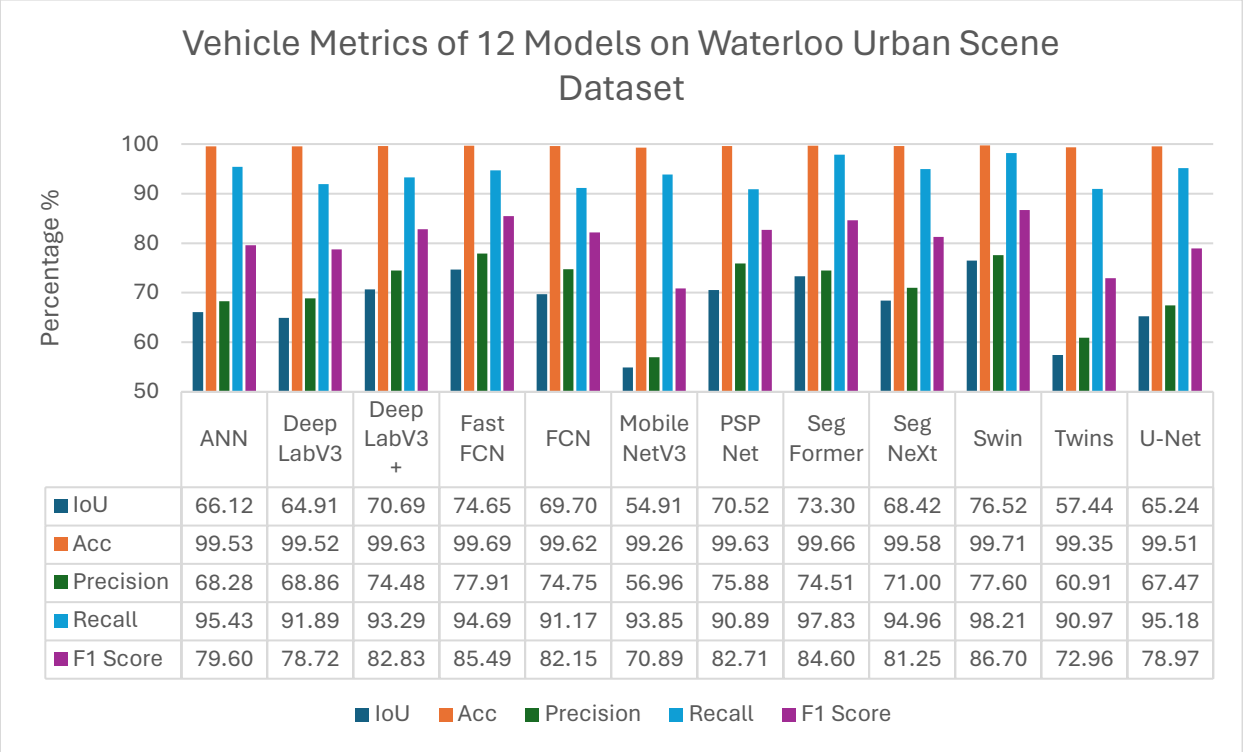
Table 4.27: Evaluation metrics for 12 different models on the performance of turn sign extraction using the Waterloo Urban Scene Dataset after fine-tuning.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	28.95	30.93	29.57	33.59	27.57	20.42	35.47	79.58	68.89	42.35	77.20	20.27
Acc	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Precision	28.95	30.93	29.57	33.59	27.57	20.42	35.47	80.33	69.17	42.35	80.65	20.27
Recall	100.00	100.00	100.00	100.00	100.00	100.00	100.00	98.83	99.42	100.00	94.75	100.00
F1 Score	44.90	47.25	45.64	50.29	43.23	33.91	52.37	88.63	81.58	59.50	87.13	33.71

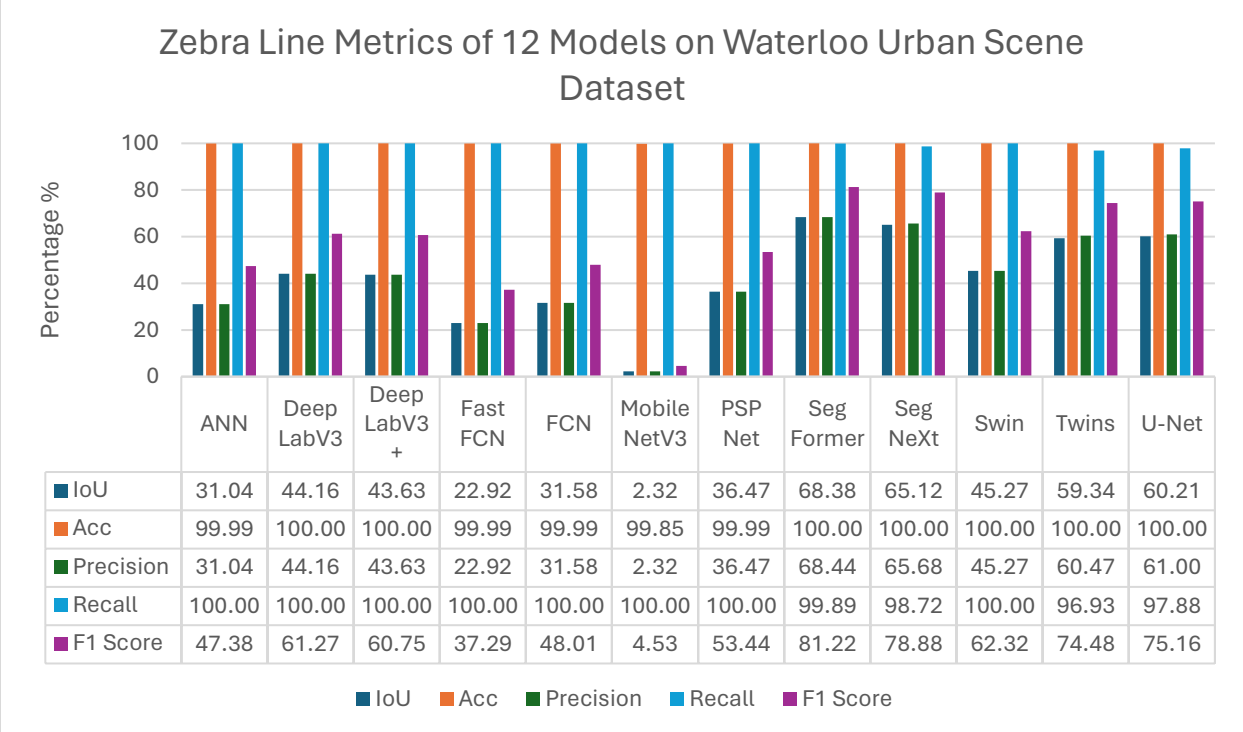
Table 4.27 examines the metrics for turn sign detection, showing that models employing attention mechanisms are at the forefront of performance, with Swin slightly surpassing PSPNet, the top performer among the conventional CNN-based models. Notably, all models achieve exceptionally high recall rates in this category. In Table 4.28, the analysis of vehicle metrics reveals a pattern like that seen with the traffic island class. However, MobileNetV3 and Twins display performances that align more closely with the rest of the models, marking a deviation from their usual standings. Table 4.29 evaluates the metrics for the zebra zone, where MobileNetV3 is identified as the weakest performer, ranking near the bottom. Contrary to expectations, U-Net achieves results comparable to those of leading models such as SegFormer, SegNeXt, and Twins, even exceeding Swin's performance, and is distinguished by its exceptionally high recall rate.

Table 4.28: Evaluation metrics for 12 different models on the performance of vehicle extraction using the Waterloo Urban Scene Dataset after fine-tuning.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	66.12	64.91	70.69	74.65	69.70	54.91	70.52	73.30	68.42	76.52	57.44	65.24
Acc	99.53	99.52	99.63	99.69	99.62	99.26	99.63	99.66	99.58	99.71	99.35	99.51
Precision	68.28	68.86	74.48	77.91	74.75	56.96	75.88	74.51	71.00	77.60	60.91	67.47
Recall	95.43	91.89	93.29	94.69	91.17	93.85	90.89	97.83	94.96	98.21	90.97	95.18
F1 Score	79.60	78.72	82.83	85.49	82.15	70.89	82.71	84.60	81.25	86.70	72.96	78.97

Table 4.29: Evaluation metrics for 12 different models on the performance of zebra line extraction using the Waterloo Urban Scene Dataset after fine-tuning.



	ANN	Deep LabV3	Deep LabV3 +	Fast FCN	FCN	Mobile NetV3	PSP Net	Seg Former	Seg NeXt	Swin	Twins	U-Net
IoU	31.04	44.16	43.63	22.92	31.58	2.32	36.47	68.38	65.12	45.27	59.34	60.21
Acc	99.99	100.00	100.00	99.99	99.99	99.85	99.99	100.00	100.00	100.00	100.00	100.00
Precision	31.04	44.16	43.63	22.92	31.58	2.32	36.47	68.44	65.68	45.27	60.47	61.00
Recall	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.89	98.72	100.00	96.93	97.88
F1 Score	47.38	61.27	60.75	37.29	48.01	4.53	53.44	81.22	78.88	62.32	74.48	75.16

The evaluation of the Waterloo Urban Scene dataset reveals nuanced differences in model performance across various urban imaging domains, which, when compared with previous findings from the SkyScapes dataset, highlight the importance of dataset diversity in understanding model behavior. The enhanced accuracy observed in classes with a larger presence in the imagery, such as Road, Traffic Island, Sidewalk, and Vehicle, can be directly attributed to their alignment with the core objectives of semantic segmentation models tailored for ground view scene analysis. This congruence leads to a notable improvement in model performance for these categories, underscoring the effectiveness of these models in recognizing and processing larger object classes within urban scenes.

Conversely, classes characterized by smaller or linear features, like Dash Lines and Long Lines, showcase a divergent trend, with less pronounced performance enhancements, suggesting a variability in model effectiveness across different class types within the Waterloo Urban Scene dataset. These observations not only complement the insights gained from the SkyScapes dataset but also set the stage for a deeper examination in the subsequent discussion chapter.

4.2 Visualization of Results

4.2.1 SkyScapes Dataset

This section provides visual representations of the predictions generated by 12 models on the SkyScapes dataset. Through side-by-side comparisons with input images and ground truth labels, this section offers insights into the accuracy and details of each model's semantic segmentation. Highlighting main characteristics and key findings, quantitative metrics with qualitative analysis are complemented. Additional visualization results are available in Appendix B.1.

Among the visualized predictions, the initial focus is drawn towards PSPNet and MobileNetV3, where misclassification errors are significant, particularly in the background class. This observation aligns with their previously noted lower accuracy on

background class segmentation. Conversely, despite SegNeXt achieving the second highest mIoU, it is noteworthy that transformer-based models such as SegFormer, Swin, and Twins exhibit superior visualization results. These transformer-based models demonstrate remarkable fidelity to the ground truth, exhibiting minimal distortion in lane thickness and negligible irrelevant errors. Furthermore, U-Net's performance is notable for its adeptness in delineating boundaries; however, it appears to weak in classifying multiple classes. This limitation may stem from U-Net's original design intended for binary classification tasks in medical imagery, suggesting its potential inadequacy in handling the complexity of multi-class semantic segmentation tasks.

These findings underscore the strengths and weaknesses inherent in each model's architecture and highlight the importance of considering the original design objectives when assessing their performance across diverse tasks. Such insights gained from the visual analysis offer valuable perspectives for further refinement and optimization of semantic segmentation models.

In the visual analysis depicted in Figure 4.1, the ground truth includes various line markings, such as dashed, solid, stop lines, and turn signs. Examination of the visual outputs from several models reveals that models like ANN, DeepLabV3, FastFCN, FCN, MobileNetV3, and PSPNet produce road lane markings that are significantly thicker than those in the ground truth. This observation is supported by data in Table 4.1, where a notable difference is observed between recall and precision for these models, with recall significantly higher than precision. Conversely, all transformer-based models, along with SegNeXt and U-Net, tend to create lane markings that are finer and closer to the ground truth. Most models are accurate in capturing both long and dashed lines. Additionally, ANN, DeepLabV3+, FastFCN, SegFormer, SegNeXt, Swin, and Twins are capable of correctly identifying stop lines. Only the transformer models and SegNeXt accurately identify turn signs. However, some models, particularly MobileNetV3 and PSPNet, struggle with a high rate of background misclassification, while DeepLabV3 and FCN also exhibit minor issues with background misclassification. This issue with background misclassification is further validated by Table 4.2, where these models' background mIoU is low.

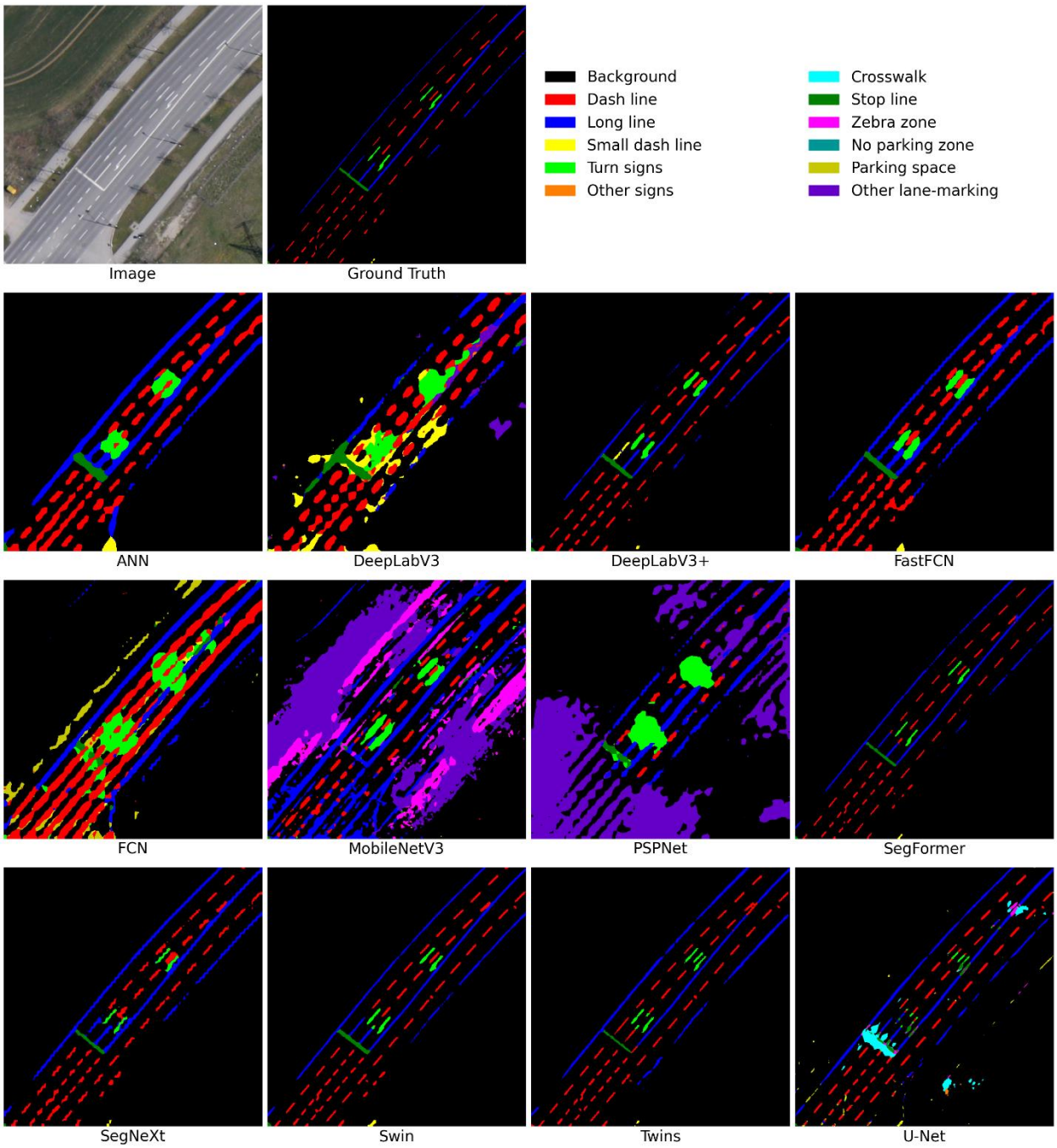


Figure 4.1 A comparative visualization of road lane detection by 12 different models on a SkyScapes dataset sample. Each model's output is showcased alongside the original image and the ground truth for reference.

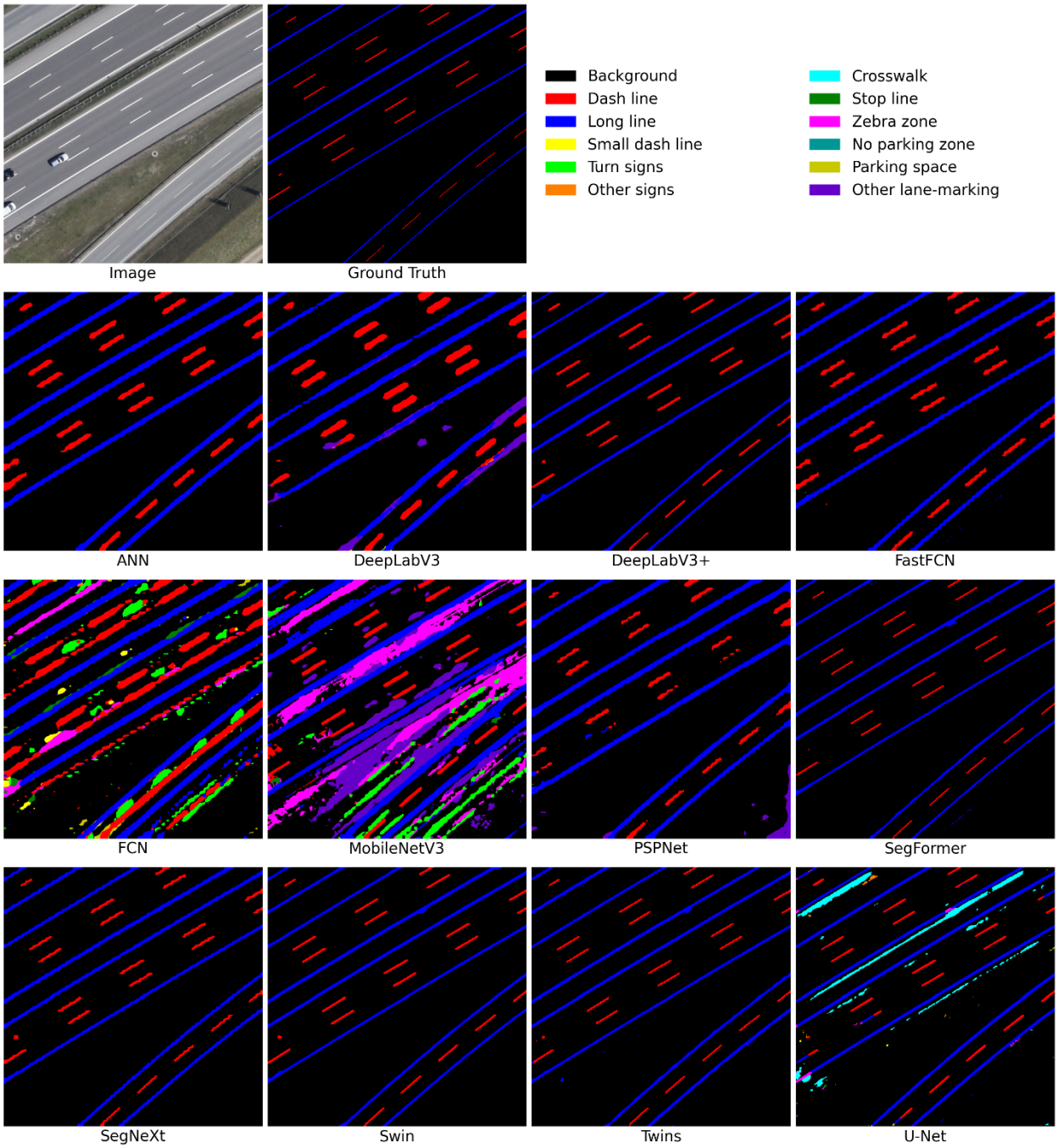


Figure 4.2: A comparative visualization of road lane detection at a highway by 12 different models on a SkyScapes dataset sample. Each model's output is showcased alongside the original image and the ground truth for reference.

In the highway scenario depicted in Figure 4.2, various roads feature both long and dashed lines. Most models accurately identify these lines, though models like ANN, DeepLabV3, and FCN generate lines that are noticeably thicker compared to the ground

truth markings. Additionally, some models, including MobileNetV3, FCN, and U-Net, incorrectly classify elements of the background. Among the twelve models evaluated, SegFormer, SegNeXt, Swin, and Twins outperform the others by accurately and finely mapping the road lanes. This superior performance is corroborated by data in Table 4.1, showing that the mIoU for these models exceeds 30%, significantly higher than that of the other models, which fall below this threshold.

Figure 4.3 depicts a highway scene featuring various lane markings, including zebra zones, dashed lines, and long lines. It is observed that most of the models successfully identify the long and dashed lines, although some models, like ANN, DeepLabV3, and FCN, depict these lines thicker than intended. The detection of zebra zones presents a challenge for several models; however, DeepLabV3+, SegFormer, SegNeXt, Swin, and Twins demonstrate the capability to accurately recognize the zebra zone.

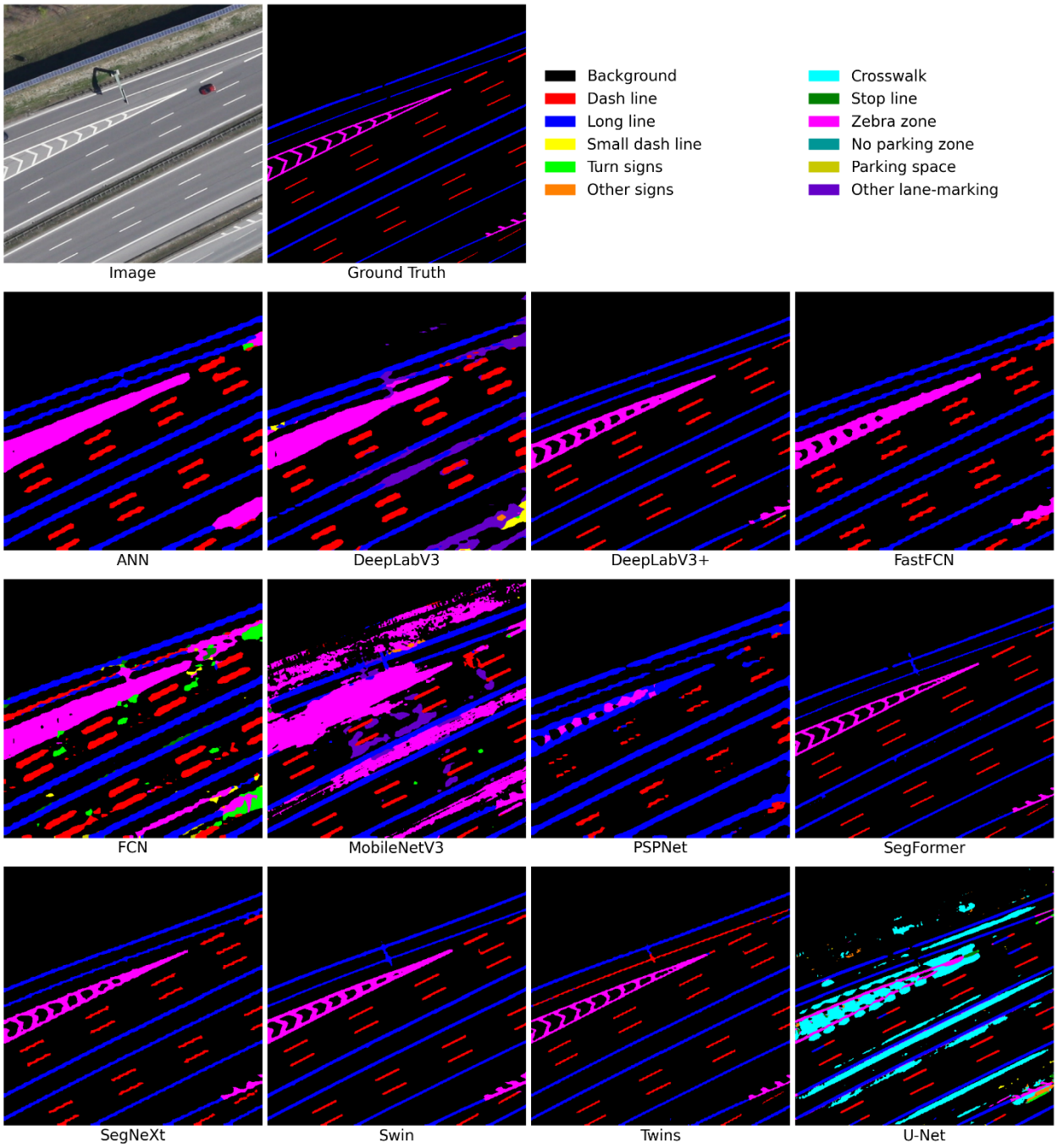


Figure 4.3 A comparative visualization of road lane detection at a highway with zebra zone by 12 different models on a SkyScapes dataset sample. Each model's output is showcased alongside the original image and the ground truth for reference.

4.2.2 Waterloo Urban Scene Dataset

In the complex intersection scenario shown in Figure 4.4 below, the road lane classifications include features such as stop lines, crosswalks, turn signs, single solid lines, dashed lines, and small dashed lines. The figure also presents classes for vehicles, sidewalks, and roads. Given the scene's complexity, while many models capture the overall appearance, certain elements like turn signs are depicted with excessive thickness, making them difficult to discern. Swin stands out by accurately classifying most of the scene without significant background misclassifications, although small dashed lines and turn signs are somewhat indistinct. SegFormer also performs well, effectively identifying road lanes despite some confusion in classifying certain background areas as roads. This is deemed acceptable given that these areas share similar color and texture with the road, as verified by the ground truth.

In the parking area depicted in Figure 4.5, parking lines are the primary feature within the road lane classification. Other identified classes include sidewalks, vehicles, traffic islands, and roads. Swin stands out among all the models for its performance, producing thin road lanes that are all accurately classified. In Figure 4.6, which depicts a partial intersection, the road lane classes include crosswalk, stop line, single solid line, dash line, and turn sign. Additional features present are the traffic island, sidewalk, and the road itself. Among the 12 models evaluated, Swin stands out for its performance, effectively distinguishing the thin lane against a clear background and accurately classifying all the mentioned classes.

In the road scene with parked vehicles depicted in Figure 4.8, the road lane classification primarily includes single solid lines. Additionally, the scene contains traffic islands, sidewalks, roads, and vehicles. Once again, Swin emerges as the top performer in road lane extraction among all twelve models. Some of the figures above clearly show significant noise in these aerial view images, particularly in Figure 4.7 where numerous vehicles parked on the road obscure the lane markings. Despite this, models like Swin are still able to successfully extract the road lane markings.

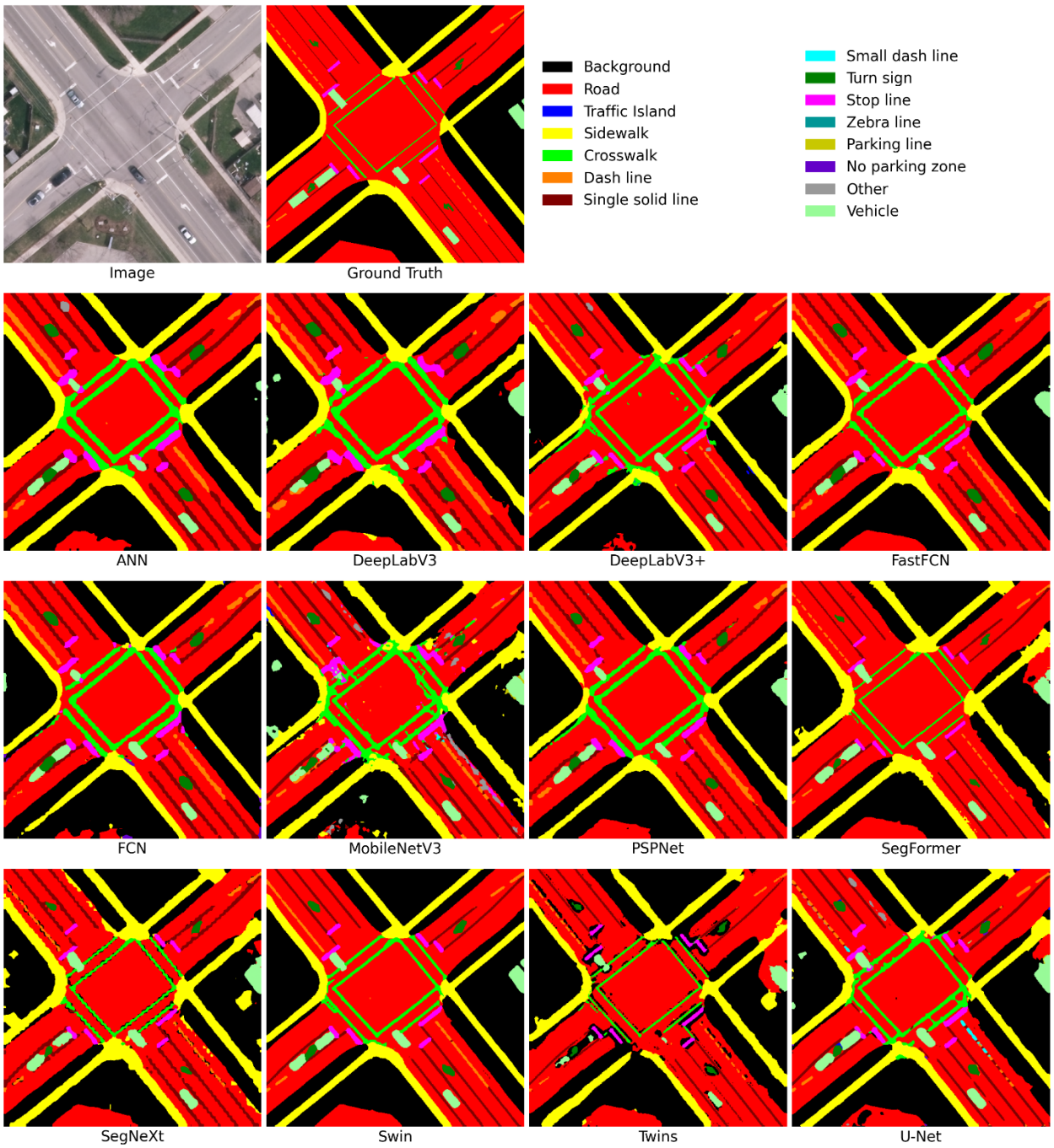


Figure 4.4: A comparative visualization of road lane detection at an intersection by 12 different models on a Waterloo Urban Scene dataset sample. Each model's output is showcased alongside the original image and the ground truth for reference.

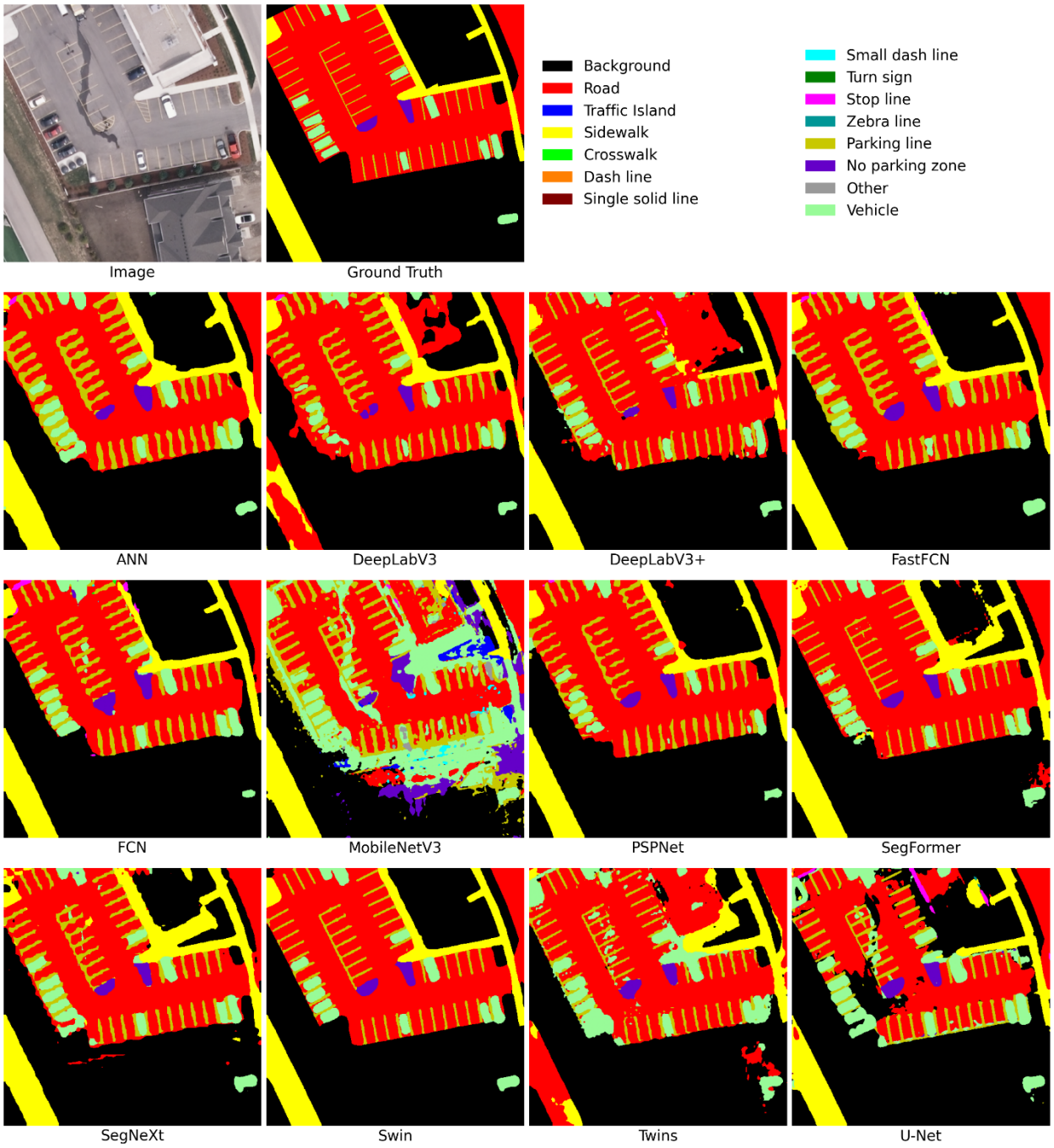


Figure 4.5: A comparative visualization of road lane detection at a parking zone by 12 different models on a Waterloo Urban Scene dataset sample. Each model's output is showcased alongside the original image and the ground truth for reference.

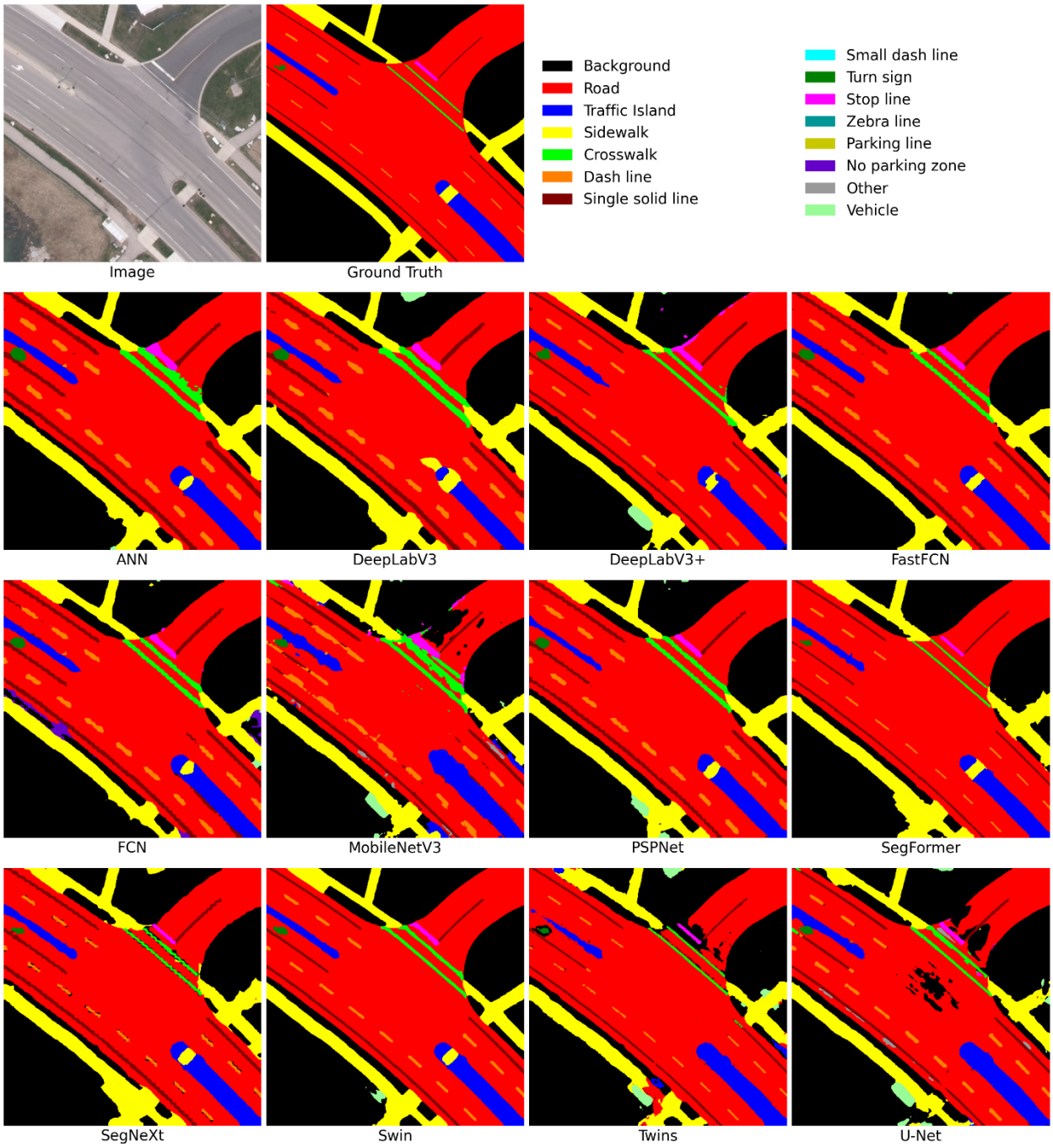


Figure 4.6: A comparative visualization of road lane detection at an intersection by 12 different models on a Waterloo Urban Scene dataset sample. Each model's output is showcased alongside the original image and the ground truth for reference.

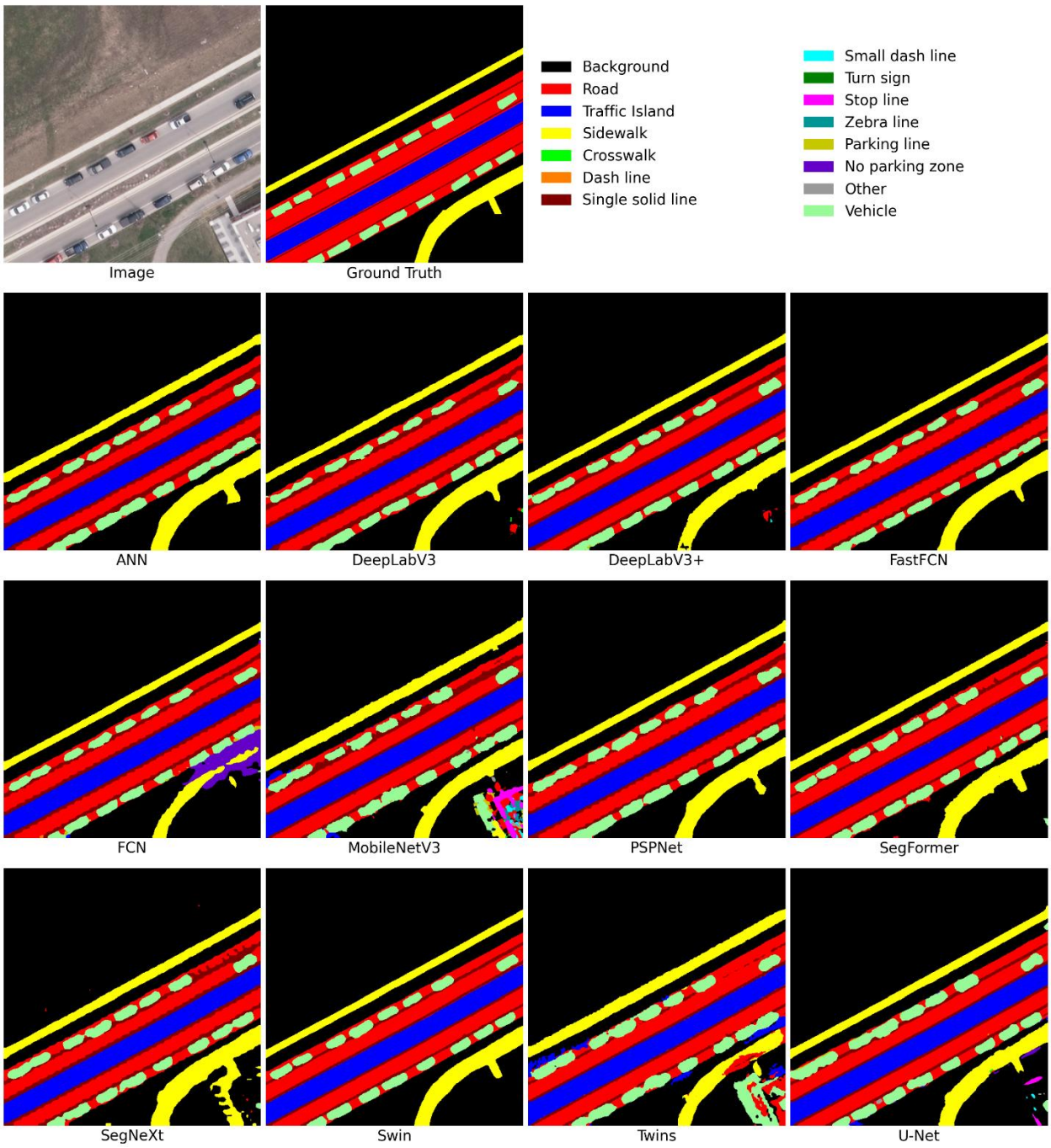


Figure 4.7: A comparative visualization of road lane detection at a road with parked vehicles by 12 different models on a Waterloo Urban Scene dataset sample. Each model's output is showcased alongside the original image and the ground truth for reference.

In contrast to the SkyScapes dataset, the visual analysis of the Waterloo Urban Scene dataset reveals different performance trends among the models. More visualization results are covered in appendix B.2. MobileNetV3, U-Net, and Twins appear to be the least effective models, showing misclassifications and inconsistencies in scene interpretation, especially when roads are covered by shadows or vegetation. On the other hand, SegFormer and Swin consistently perform well, accurately capturing scene details.

One notable observation is that SegNeXt sometimes produces strange background predictions around the Dash Line class, which is not seen in other models. This suggests the need for careful interpretation of model outputs. Moreover, CNN-based models generally perform better on the Waterloo Urban Scene dataset compared to the SkyScapes dataset, particularly in simpler scenes. These models show improved accuracy in identifying large objects like sidewalks, roads, and vehicles. These findings reinforce the importance of considering both quantitative metrics and qualitative visual assessments when evaluating model performance across different datasets and scene complexities.

4.3 Discussion

Transformer vs. CNN Based Models:

Transformer-based models exhibit an advantage in capturing long-range dependencies, which is crucial for understanding complex scenes in remote sensing imagery. Unlike CNNs, which primarily focus on local dependencies, transformers can efficiently learn relationships between distant pixels. However, transformers require intensive training to extract features effectively. The absence of inherent knowledge about pixel distributions and local relations necessitates pre-trained backbones for transformers to achieve robust performance, particularly in smaller datasets with simpler scenes.

Recent research has shown that CNNs can emulate the long-range dependency capturing capability of transformers through the integration of attention mechanisms (Liu et al., 2022). Models like SegNeXt demonstrate the efficacy of this approach, suggesting that CNNs can rival transformers in certain tasks (Guo et al., 2022). Additionally, incorporating special pooling layers after CNNs can further enhance performance while

reducing computational complexity and runtime.

Recall and Precision:

Recall and precision are fundamental metrics in evaluating semantic segmentation performance. As recall increases, precision tends to decrease, and vice versa. For example, in SkyScapes dataset, ANN demonstrates a recall of 66.0% and a precision of 22.92%, while SegFormer, a transformer-based model, exhibits a recall of 43.85% and a precision of 64.33%. Achieving a balance between these metrics is essential for accurate detection. Strategies to improve recall and precision include refining prediction boundaries to better match ground truth objects and minimizing over-predictions or under-predictions. Utilizing the F1 score, which combines both recall and precision, provides a comprehensive assessment of model performance, particularly in tasks where balancing these metrics is challenging.

IoU and F1 Score:

IoU and F1 score evaluate the overlap between predicted and ground truth regions, considering both shape and location. They provide nuanced insights into the accuracy of object detection algorithms. While both metrics incorporate true positives, false positives, and false negatives, they use different coefficients to weigh these components. In a comparative study of model performance on SkyScapes dataset, DeepLabV3, ANN, and Swin achieved mIoU scores of 10.24%, 20.94%, and 30.51%, respectively, with corresponding F1 scores of 12.96%, 29.88%, and 42.97%. A subsequent evaluation on Waterloo Urban Scene Dataset revealed mIoU scores of 60.31%, 58.23%, and 72.48%, with F1 scores 47.55%, 46.5%, and 72.48% paralleling with IoU changes, demonstrating the consistency in both mIoU and F1 metrics across models. Understanding these differences aids in interpreting the performance of models across various datasets and scenarios.

Accuracy:

Accuracy metrics may be skewed by class imbalances, particularly in datasets where certain classes dominate, such as backgrounds in remote sensing imagery. This dominance inflates accuracy scores, potentially masking performance issues in other

classes. For instance, in SkyScapes dataset analysis, the SegFormer model achieved a mAcc of 99.92%, with an all accuracy of 99.50%. In contrast, evaluations on the Waterloo Urban Scene Dataset yielded accuracies of 99.77% and 98.27%, respectively. The variance can be attributed to SkyScapes' larger background pixel proportion, leading to a higher rate of correct background predictions. The Waterloo Urban Scene Dataset, with a lower background pixel proportion, exhibits reduced accuracy, highlighting the impact of class distribution on accuracy metrics. Addressing class imbalances is essential for obtaining meaningful accuracy assessments and ensuring that models generalize well across diverse classes.

Distribution Differences:

Differences in datasets, as observed between the Waterloo Urban Scene and SkyScapes datasets, can significantly influence model performance. A more balanced distribution, as seen in the Waterloo Urban Scene dataset, facilitates better learning by the model, leading to improved performance. Additionally, higher-resolution datasets like Waterloo Urban Scene may present larger class objects, simplifying the learning task for models designed for ground-view applications. Understanding dataset characteristics is crucial for optimizing model training and performance evaluation in remote sensing tasks. When examining pixel counts across datasets, it's evident that the Waterloo Urban Scene Dataset encompasses a wider range of classes, and less skewed class distribution contributing to the observed differences.

Noise in Aerial View Imagery:

Aerial images frequently include various obstructions such as trees, vehicles, and shadows, which can obscure critical details. Figure 4.8 illustrates how background features in aerial views, such as road lane markings, often become obscured by trees and utility poles. Consequently, the SkyScapes dataset adopts a questionable annotation practice that only marks road lane markings that are visible in the aerial images, ignoring their actual existence. This method leads to ground truth annotations and training data where road lane markings appear fragmented, as depicted in Figure 4.9.

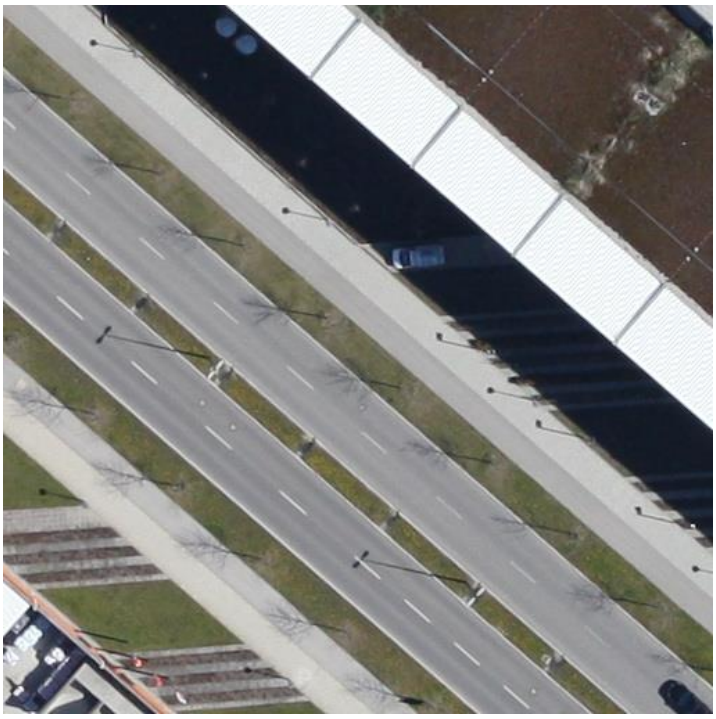


Figure 4.8 A raw aerial image from the SkyScapes dataset where trees and utility poles obstruct the visibility of road lane markings.

Despite these challenges, some models have demonstrated an ability to overcome such data limitations. Figure 4.10 showcases how the Swin model, for example, effectively discerns the spatial relationships between adjacent road lane marking pixels. This capability allows the model to reconstruct the continuity of road lane markings, thus compensating for the gaps and discontinuities present in the source aerial images. This adaptability highlights the potential of advanced deep learning models to mitigate the effects of noise in aerial imagery, thereby enhancing the reliability of the data derived from these images for various practical applications.

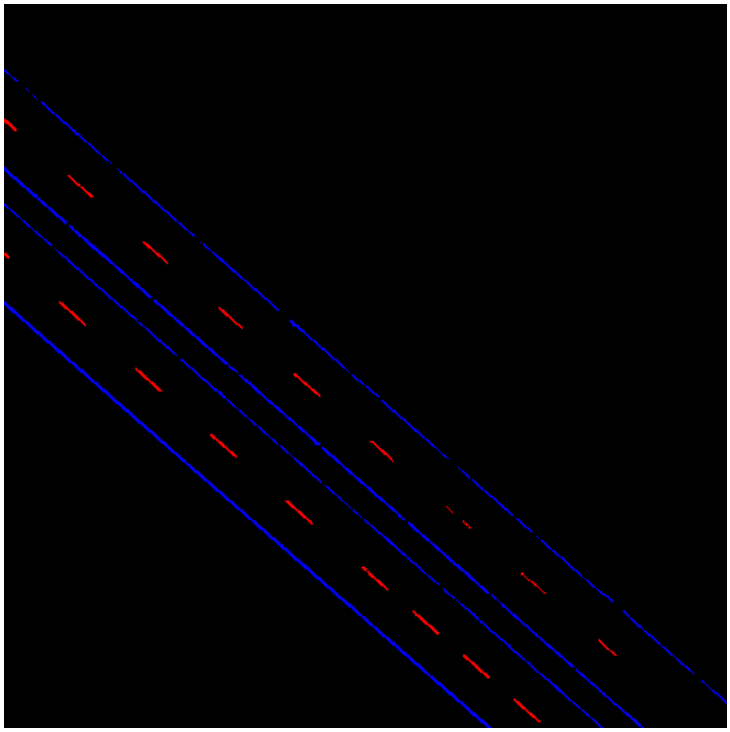


Figure 4.9 The ground truth image from Figure 4.8 of the SkyScapes dataset, illustrating that road lane markings are annotated as discontinuous rather than continuous due to obstructions.

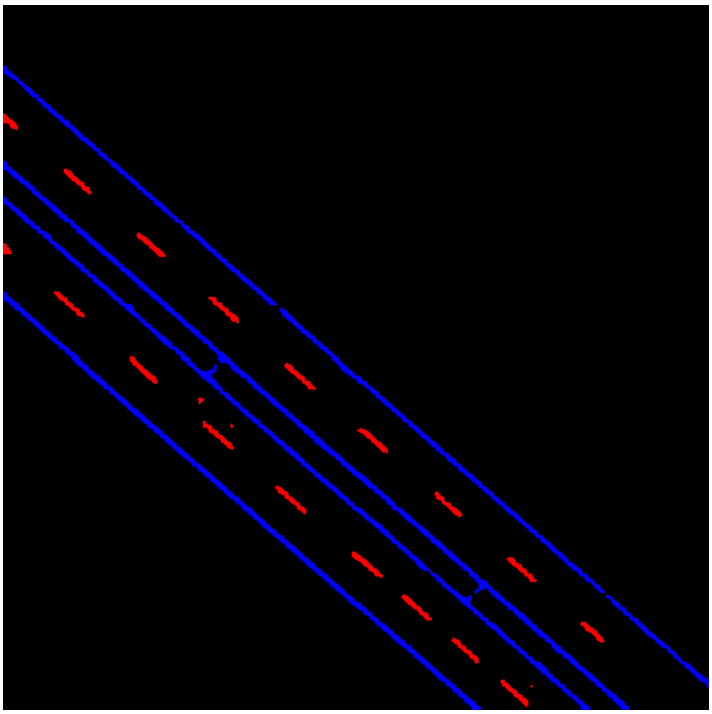


Figure 4.10 Prediction result from the Swin model on the SkyScapes test dataset, showing relatively continuous road lane markings.

Chapter 5 Conclusions and Recommendations

5.1 Conclusion

The rapid progress of AVs underscores the critical necessity for HD maps, where precise road lane detection is fundamental for their navigation. To generate HD maps, utilizing aerial imagery combined with deep learning techniques presents an effective method for the lane marking extraction task. This thesis has conducted an extensive comparative analysis to explore the performance of both CNN-based and transformer-based state-of-the-art models in semantic segmentation for road lane extraction tasks. A total of 12 models were evaluated using various performance metrics on the SkyScapes Dataset, and further fine-tuning was undertaken on the Waterloo Urban Scene Dataset.

Through thorough analysis of metric results across different models and specific categories within each model, it was found that the SegFormer model achieved the highest mIoU of 76.11% and an F1 score of 85.35% in the Waterloo Urban Scene dataset post-fine-tuning. Additionally, visual examination of prediction outputs provided substantial insights. It was noted that transformer-based models, such as Swin and SegFormer, perform well in road lane extraction tasks, particularly with more commonly occurring lane marking classes such as single solid lines or dashed lines. Despite challenges such as noise interference from trees or shadows in aerial imagery, some models demonstrated the ability to discern the continuous relationships between pixels, thereby ensuring the continuity of road lane marking predictions.

This study contributes to a more profound comprehension of the strengths and limitations of various semantic segmentation models in the context of aerial view road lane segmentation. It also suggests directions for future research, particularly the need to address issues arising from unbalanced datasets. Additionally, this research guides subsequent improvements and developments in model design and specificity, as well as the adoption of transfer learning procedures. Furthermore, the results will inform the design and annotation of training datasets, ensuring the provision of high-quality data to foster advancements in aerial view road lane extraction.

In conclusion, this comparative analysis clarifies the performance capabilities of semantic segmentation models and establishes a foundation for future research dedicated to enhancing road lane extraction algorithms in aerial view scenarios. This research contributes to developing more comprehensive and effective solutions in autonomous navigation and HD mapping, thus improving the functionality of HD maps for AV navigation in aerial imagery.

5.2 Recommendations for Future Research

In this section, recommendations are provided based on the limitations and challenges discussed in the preceding discussion section, which specifically addresses the task of road lane marking extraction from aerial imagery using deep learning methodologies. Considering AI's rapid evolution, additional potential research directions are suggested to enhance this task, with examples drawn from diffusion models and generative AI.

Unbalanced Class Distribution

Unbalanced class distribution is a significant challenge in semantic segmentation tasks, particularly in remote sensing where most of the data often comprises the background. In the SkyScapes dataset, the background class dominates significantly, comprising 167,817,849 of the total 168,210,432 pixels, which accounts for a staggering 99.77% of the dataset, as shown in Figure 5.1. Conversely, in the Waterloo Urban Scene dataset, the background constitutes 421,098,771 out of 503,840,768 total pixels, representing 83.58%, as illustrated in Figure 5.2. This imbalance can skew the model's learning process, leading to a bias towards the majority class and potentially ignoring the minority classes, which are often of greater interest.

To address this issue, advanced techniques such as weighted loss functions and oversampling of minority classes can be employed to balance the class distribution and ensure that the model pays equal attention to all classes. For instance, in SkyScapes, the weighting assigned to the background class was set at 0.08, while more critical, under-represented classes such as the no parking zone, dash line, and long line received

significantly higher weights of 4,371.37, 173.68, and 54.55, respectively. Similarly, in the Waterloo Urban Scene Dataset, weights were adjusted to reflect the skewed distribution, with background at 0.08, road at 0.58, traffic island at 15.19, and no parking zone at 111.95. These methods can help mitigate the dominance of background classes and improve the detection of finer details in the datasets.

The phenomenon of models predicting more adjacent pixels for minority classes reflects a significant challenge in semantic segmentation, particularly in the context of remote sensing and similar fields where the precision of spatial predictions is crucial. The underlying issue stems from the differential impact of accuracy improvements across classes due to the weighted loss. Specifically, enhancing accuracy for a minority class, which carries a higher weight, results in a more substantial reduction in the loss score compared to similar improvements in the majority class. Consequently, models are incentivized to increase the prediction area of minority classes, as doing so can decrease the overall loss more effectively than by correctly predicting the majority class's extent. This dynamic leads to an undesirable trade-off: the model achieves lower loss scores which suggests better performance at the expense of spatial accuracy, especially concerning the delineation of minority class boundaries.

To mitigate this issue, it's essential to explore alternative strategies that maintain the balance between addressing class imbalance and preserving spatial accuracy. Approaches such as focal loss, which modulates the loss contribution from each sample based on the correctness of the prediction, could offer a more detailed way to handle class imbalance by reducing the incentive for the overprediction of minority classes. Additionally, incorporating spatial context into the loss function, either through post-processing techniques like Conditional Random Fields (CRFs) or through novel loss functions that explicitly penalize spatial inaccuracies, could help in aligning the model's predictions more closely with the true spatial distribution of classes. These strategies aim to refine the model's learning process, ensuring that it not only balances the representation of classes but also accurately captures their spatial characteristics, leading to more precise and realistic segmentation results.

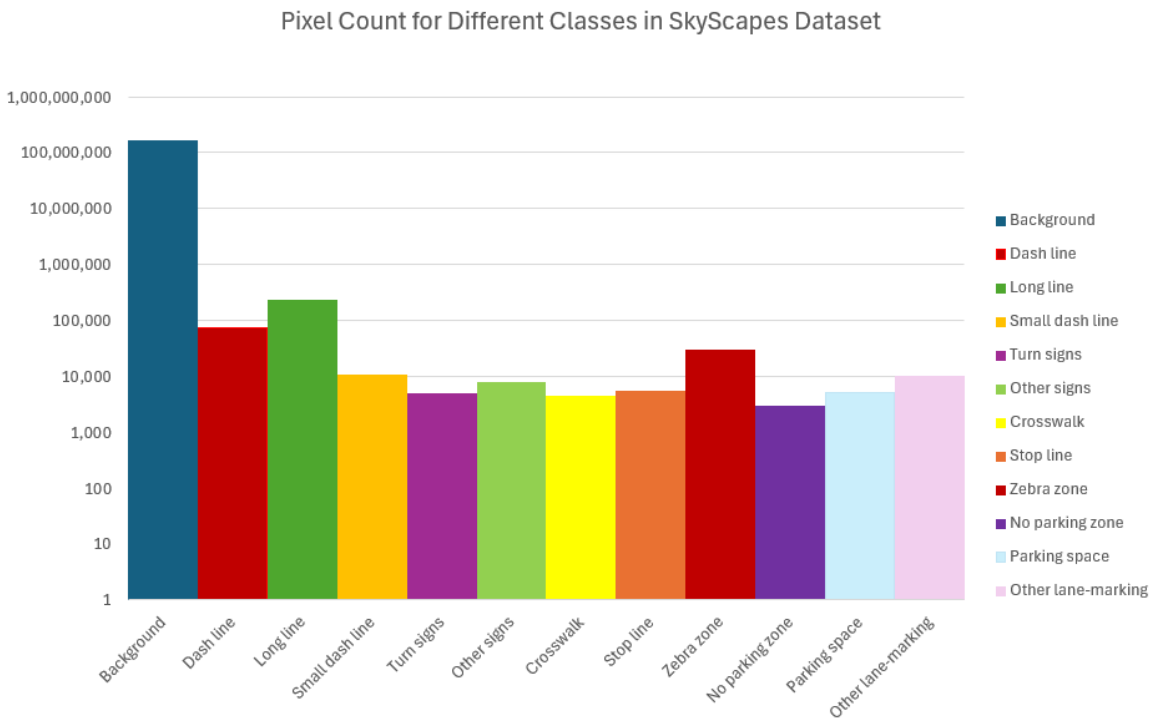


Figure 5.1 Logarithmic scale pixel count for various classes in the SkyScapes dataset.

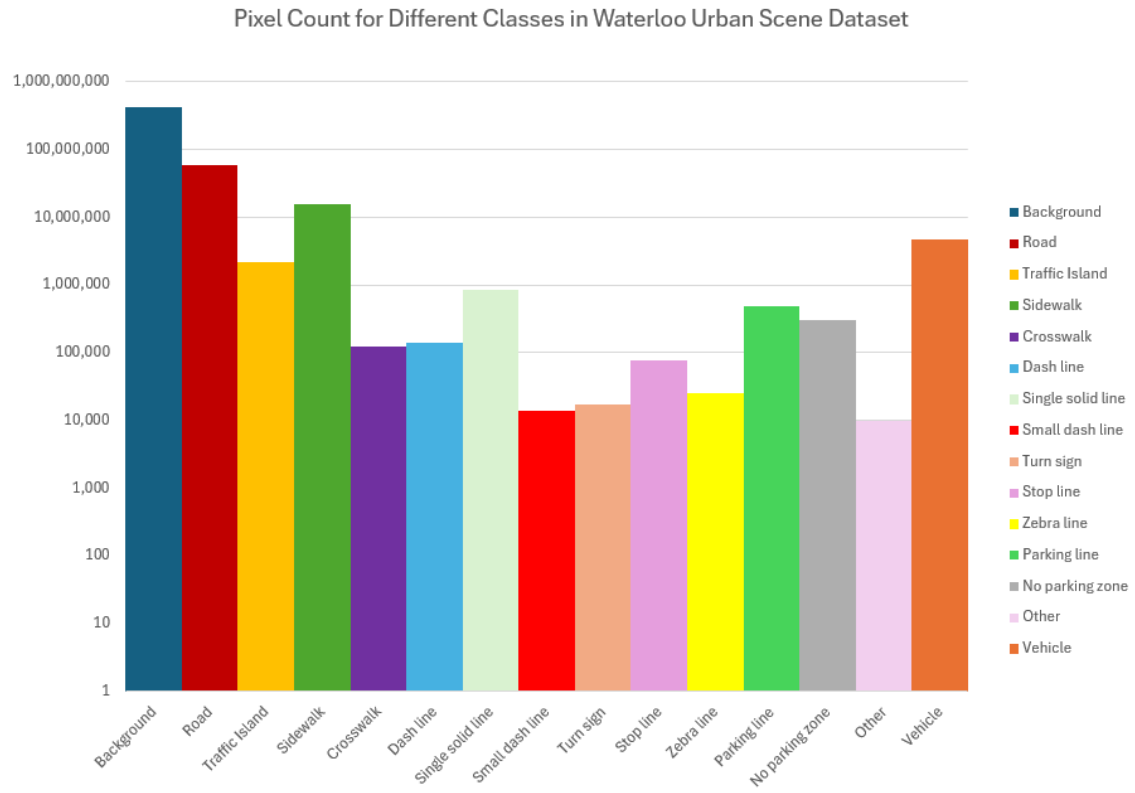


Figure 5.2 Logarithmic scale pixel count for various classes in the Waterloo Urban Scene dataset.

Transfer Learning Procedure

The observation that SkyScapes and Waterloo Urban Scene datasets have a similar distribution suggests an opportunity for leveraging transfer learning more effectively. Instead of treating fine-tuning on Waterloo Urban Scene datasets as a mere interpolation task, considering the distinct background percentages (99.77% and 83.58%) and weighted classes between SkyScapes and Waterloo Urban Scene Dataset indicates a nuanced relationship; they share similarities in their disproportionate background representation but also differ in the specific weights assigned to other classes, suggesting a complex landscape for model adaptation rather than a straightforward one. More sophisticated transfer learning strategies can be explored that explicitly account for the similarities and differences between datasets. Techniques such as domain adaptation or meta-learning can be particularly useful here. These methods can help in adjusting the model's parameters to better capture the unique characteristics of each dataset, thus improving the model's ability to generalize from one domain to another.

Model Design and Specificity

The limitation of general semantic segmentation models in handling fine, thin, and long objects points to a need for specialized model architectures or enhancements tailored to the unique requirements of remote sensing data. Incorporating modules designed to capture fine-grained details or employing architectures that better model spatial relationships can improve performance on such datasets. Additionally, exploring novel neural network architectures, such as transformers that are inherently better at capturing long-range dependencies, might offer significant improvements for the specific challenges of remote sensing imagery.

Labeling Errors and Spatial Information

The presence of labeling errors in the SkyScapes dataset, such as discontinuous labeling of objects obscured by obstacles, as shown in Figure 5.3 and 5.4, underscores the importance of high-quality, accurate labeling for training robust models. Developing semi-automated labeling tools that leverage model predictions to suggest labels, which are then verified and corrected by human annotators, can improve label accuracy.

Furthermore, enhancing models with mechanisms to infer spatial continuity and context can help in overcoming the limitations posed by incomplete or incorrect labels. Techniques such as CRFs integrated into deep learning frameworks, or attention mechanisms that allow models to learn spatial relationships, can enable models to better understand the scene rather than treating it as a mere pixel-to-pixel mapping task.

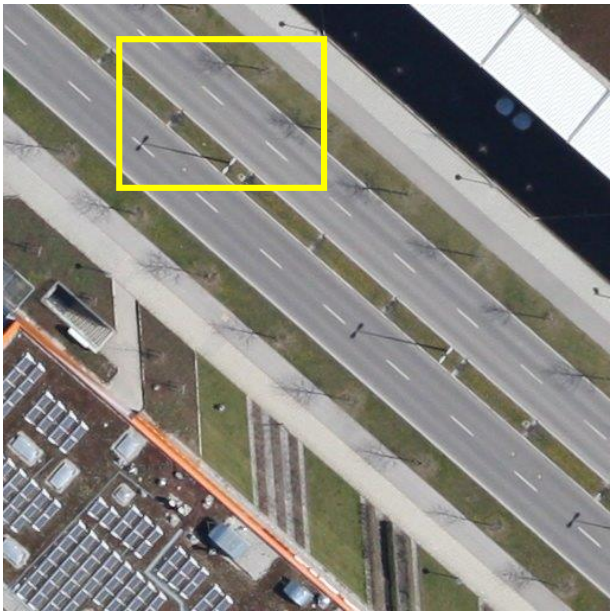


Figure 5.3 Obstructions covering road lane markings in the SkyScapes dataset.

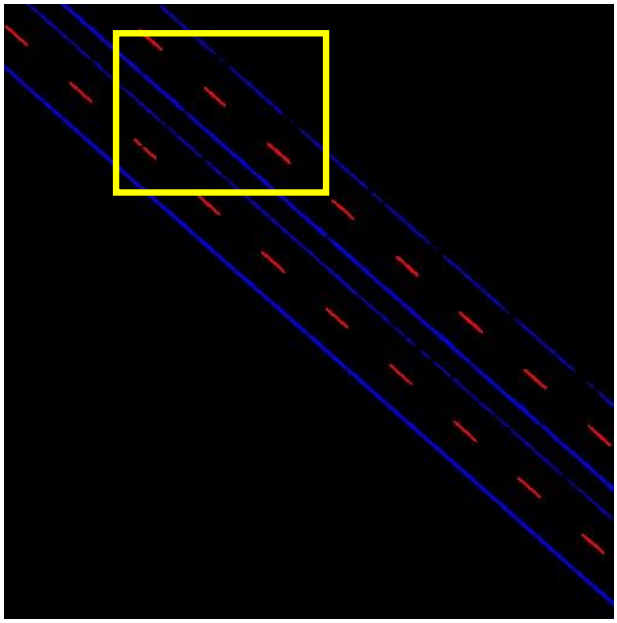


Figure 5.4 The annotation error of discontinuous road lane markings in the SkyScapes Dataset.

Addressing these challenges requires a multifaceted approach that combines advanced machine learning techniques, tailored model architectures, and improved data processing and labeling strategies. By focusing on these areas, researchers can develop more accurate, robust, and generalizable models for semantic segmentation tasks in remote sensing, thereby overcoming the limitations identified in this study.

Diffusion models and Generative 3D AI

In the ongoing evolution of artificial intelligence, the integration of diffusion models and generative AI could possibly be highly beneficial for enhancing the detection and extraction of lane markings, a pivotal aspect in the development of AVs' navigation systems. These advanced AI techniques hold the promise to refine the training of both 2D and 3D models, facilitating a more detailed understanding of complex environments. Specifically, diffusion models have the potential to transform the training process for 2D models in lane detection, by synthesizing/adapting training images with different conditions, such as fluctuations in lighting, weather changes, and the presence of dynamic obstacles.

Concurrently, 3D generative AI is promising for crafting sophisticated simulations of urban environments, supporting the thorough training of autonomous vehicles through the provision of realistic, intricate 3D city models and maps. These maps could include crucial information beyond mere topography, like obscured views and elevation profiles, thereby offering a more comprehensive dataset for AV systems to navigate safely and efficiently.

Further exploration into the cooperative application of these AI technologies could lead to significant advancements in how AVs interpret and interact with their surroundings. Generating detailed 3D representations of urban landscapes from aerial views, these AI models not only enhance the environmental awareness of AVs but also hold potential for urban planning and the establishment of smart cities. The creation of more precise and informative 3D maps, capturing real-time conditions and environmental variables, could

transform urban mobility and infrastructure management. Therefore, the continuous research and deployment of diffusion models and generative AI for lane markings detection and urban modeling stand to make significant contributions to the domains of autonomous driving and smart city initiatives.

References

Azimi, S. M., Fischer, P., Korner, M., & Reinartz, P. (2019a). Aerial LaneNet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing, 57(5), 2920–2938.

Azimi, S. M., Henry, C., Sommer, L., Schumann, A., & Vig, E. (2019b). SkyScapes - fine-grained semantic understanding of aerial scenes. 2019 IEEE/CVF International Conference on Computer Vision, 7393-7403.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12), 2481-2495.

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

Chao, F., Yu-Pei, S., & Ya-Jie, J. (2019). Multi-lane detection based on deep convolutional neural network. IEEE Access, 7, 150833-150841.

Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., ... & Shen, C. (2021). Twins: Revisiting the Design of Spatial Attention in Vision Transformers. arXiv preprint arXiv:2104.13840.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Gao, L., Shi, W., Miao, Z., & Lv, Z. (2018). Method based on edge constraint and fast marching for road centerline extraction from very high-resolution remote sensing images. Remote Sensing, 10(6), 900.

Gao, L., Song, W., Dai, J., & Chen, Y. (2019). Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network. Remote Sensing, 11(5), 552.

Gao, X. W., Podladchikova, L., Shaposhnikov, D., Hong, K., & Shevtsova, N. (2006). Recognition of traffic signs based on their colour and shape features extracted using human vision models. Journal of Visual Communication and Image Representation, 17(4), 675-685.

Fritsch, J., Kuehnl, T., & Geiger, A. (2013, October). A new performance measure and evaluation benchmark for road detection algorithms. In 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013) (pp. 1693-1700).

Guo, M. H., Lu, C. Z., Hou, Q., Liu, Z., Cheng, M. M., & Hu, S. M. (2022). Segnext: Rethinking convolutional attention design for semantic segmentation. Advances in Neural Information Processing Systems, 35, 1140-1156.

Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. International Journal of Multimedia Information Retrieval, 7, 87-93.

He, H., Jiang, Z., Gao, K., Narges Fatholahi, S., Tan, W., Hu, B., ... & Li, J. (2022). Waterloo building dataset: A city-scale vector building dataset for mapping building footprints using aerial orthoimagery. Geomatica, 75(3), 99-115.

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1314-1324).

Li, J., Jiang, F., Yang, J., Kong, B., Gogate, M., Dashtipour, K., & Hussain, A. (2021). Lane-deeplab: Lane semantic segmentation in automatic driving scenarios for high-definition maps. Neurocomputing, 465, 15-25.

Liu, X., Wang, G., Liao, J., Li, B., He, Q., & Meng, M. (2012). Detection of geometric shape for traffic lane and mark. In 2012 IEEE International Conference on Information and Automation (pp. 395-399).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (pp. 9992-10002).

Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11976-11986).

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 (pp. 3431-3440).

Long, Y., Xia, G.-S., Li, S., Yang, W., Yang, M. Y., Zhu, X. X., Zhang, L., & Li, D. (2021). On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 4205–4230.

Por, E., van Kooten, M., & Sarkovic, V. (2019). Nyquist–Shannon sampling theorem. Leiden University, 1(1), 5.

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 12179-12188).

Rehman, I., Ghous, H., Malik, M., & Ismail, M. (2023). Artificial Intelligence based Lane Detection using Satellite Images. International Journal of Information Systems and Computer Technologies, 2(1).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI, 2015, Proceedings, Part III 18 (pp. 234-241).

Rottensteiner, F., Sohn, G., Gerke, M., & Wegner, J. D. (2014). ISPRS semantic labeling contest. ISPRS: Leopoldshöhe, Germany, 1(4).

Shinar, D., Dewar, R., Summala, H., & Żakowska, L. (2003). Traffic sign symbol comprehension: A cross-cultural study. Ergonomics, 46(15), 1549-1565.

Sonka, M., Hlavac, V., & Boyle, R. (2013). Image Processing, Analysis, and Machine Vision (4th ed.). Springer. DOI: 10.1007/978-1-4899-3216-7.

Strudel, R., Garcia, R., Laptev, I., & Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7262-7272).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (Vol. 27).

Van Etten, A., Hogan, D., Manso, J. M., Shermeyer, J., Weir, N., & Lewis, R. (2021). The multi-temporal urban development spacenet dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6398-6407).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30).

Wei, Y., Zhang, K., & Ji, S. (2020). Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-based segmentation and tracing. IEEE Transactions on Geoscience and Remote Sensing, 58(12), 8919-8931.

Weinland, D., Ronfard, R., & Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation, and recognition. Computer Vision and Image Understanding, 115(2), 224-241.

Wu, H., Zhang, J., Huang, K., Liang, K., & Yu, Y. (2019). Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. arXiv preprint arXiv:1903.11816.

Wu, P.-C., Chang, C.-Y., & Lin, C.-H. (2014). Lane-mark extraction for automobiles under complex conditions. Pattern Recognition, 47, 2756-2767.

Xie, E., Luo, P., Alvarez, J. M., Anandkumar, A., Yu, Z., & Wang, W. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. In Advances in Neural Information Processing Systems. arXiv preprint arXiv:2105.15203.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., ... & Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2636-2645).

Yoo, S., Seok Lee, H., Myeong, H., Yun, S., Park, H., Cho, J., & Hoon Kim, D. (2020). End-to-end lane marker detection via row-wise classification. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. arXiv preprint arXiv:2005.08630.

Zhang, X., Han, X., Li, C., Tang, X., Zhou, H., & Jiao, L. (2019). Aerial image road extraction based on an improved generative adversarial network. Remote Sensing, 11(8), 930.

Zhang, Y., Xiong, Z., Zang, Y., Wang, C., Li, J., & Li, X. (2019). Topology-aware road network extraction via multi-supervised generative adversarial networks. Remote Sensing, 11(9), 1017.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2881-2890).

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., ... & Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6881-6890).

Zhou, Y., Takeda, Y., Tomizuka, M., & Zhan, W. (2021). Automatic construction of lane-level hd maps for urban scenes. In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 6649-6656).

Zhu, Z., Xu, M., Bai, S., Huang, T., & Bai, X. (2019). Asymmetric non-local neural networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 593-602).

Appendices

Appendix A. Benchmark of all 12 models on the SkyScapes Lane Dataset over all 12 classes [4.3]

Table A.1: Class-specific performance outcomes of the FCN model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	93.39	93.43	96.58	99.89	93.49
Dash line	7.11	98.72	13.27	7.63	50.99
Long line	8.13	98.00	15.04	8.70	55.46
Small dash line	3.40	99.37	6.57	3.55	43.47
Turn signs	1.94	99.57	3.82	2.02	33.35
Other signs	2.95	99.80	5.73	3.09	40.03
Crosswalk	1.54	99.90	3.04	1.56	55.85
Stop line	6.92	99.76	12.95	7.12	71.72
Zebra zone	0.92	99.42	1.83	0.93	47.47
No parking zone	1.19	99.92	2.36	1.21	40.29
Parking space	0.17	98.56	0.33	0.17	36.54
Other lane-marking	0.44	99.96	0.88	0.67	1.27

Table A.2: Class-specific performance outcomes of the FastFCN model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	98.68	98.69	99.34	99.63	99.04
Dash line	19.36	99.53	32.43	22.43	58.51
Long line	10.58	99.43	19.13	17.48	21.12
Small dash line	11.10	99.75	19.98	11.95	60.74
Turn signs	5.25	99.96	9.97	10.32	9.65
Other signs	4.64	99.97	8.88	9.84	8.08
Crosswalk	6.02	99.98	11.36	6.58	41.69
Stop line	27.92	99.97	43.65	38.81	49.87
Zebra zone	6.57	99.97	12.32	9.80	16.59
No parking zone	0.01	100.00	0.01	0.62	0.01
Parking space	5.22	99.99	9.93	16.04	7.19
Other lane-marking	0.68	99.98	1.35	8.72	0.73

Table A.3: Class-specific performance outcomes of the U-Net model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	95.65	95.67	97.77	99.91	95.73
Dash line	26.83	99.79	42.30	43.78	40.93
Long line	23.64	99.42	38.24	29.08	55.81
Small dash line	12.45	99.71	22.14	12.84	80.37
Turn signs	8.21	99.97	15.17	28.30	10.36
Other signs	2.22	99.43	4.34	2.23	83.12
Crosswalk	0.16	98.33	0.32	0.16	100.00
Stop line	9.49	99.87	17.33	10.25	56.19
Zebra zone	0.57	99.51	1.13	0.58	24.59
No parking zone	0.23	99.6	0.46	0.23	37.01
Parking space	0.32	99.56	0.64	0.32	21.34
Other lane-marking	0.00	99.99	0.00	0.00	0.00

Table A.4: Class-specific performance outcomes of the DeepLabV3 model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	92.95	92.99	96.34	99.88	93.05
Dash line	7.91	98.80	14.66	8.49	53.87
Long line	10.83	98.93	19.54	12.84	40.83
Small dash line	2.04	98.51	4.00	2.07	60.50
Turn signs	0.52	99.95	1.03	1.08	0.99
Other signs	1.57	99.60	3.09	1.61	40.65
Crosswalk	0.44	99.91	0.88	0.45	15.71
Stop line	5.11	99.82	9.73	5.54	39.68
Zebra zone	0.10	99.83	0.20	0.11	1.56
No parking zone	0.39	99.87	0.78	0.40	21.26
Parking space	0.91	99.70	1.81	0.93	43.06
Other lane-marking	0.08	97.55	0.15	0.08	12.68

Table A.5: Class-specific performance outcomes of the DeepLabV3+ model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	99.12	99.13	99.56	99.71	99.41
Dash line	24.19	99.74	38.95	35.83	42.66
Long line	20.90	99.48	34.57	28.81	43.21
Small dash line	22.44	99.91	36.66	28.02	52.99
Turn signs	1.09	99.97	2.15	8.54	1.23
Other signs	7.43	99.98	13.83	18.65	10.99
Crosswalk	3.82	99.95	7.36	3.88	71.05
Stop line	22.96	99.94	37.34	25.65	68.62
Zebra zone	1.51	99.98	2.97	2.67	3.34
No parking zone	3.22	100.00	6.23	6.90	5.68
Parking space	8.14	99.99	15.05	20.13	12.02
Other lane-marking	2.15	99.97	4.22	4.20	4.23

Table A.6: Class-specific performance outcomes of the ANN model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	96.54	96.56	98.24	99.92	96.62
Dash line	12.47	98.95	22.18	12.92	78.43
Long line	13.99	98.47	24.55	14.56	78.21
Small dash line	7.11	99.37	13.27	7.13	94.83
Turn signs	7.21	99.94	13.45	10.41	19.02
Other signs	28.56	99.97	44.43	33.10	67.54
Crosswalk	29.07	100.00	45.04	35.14	62.71
Stop line	19.30	99.94	32.35	21.79	62.78
Zebra zone	21.38	99.96	35.23	21.97	88.87
No parking zone	5.93	99.97	11.19	6.07	70.69
Parking space	4.08	99.98	7.84	6.08	11.03
Other lane-marking	5.66	99.85	10.71	5.87	61.30

Table A.7: Class-specific performance outcomes of the MobileNetV3 model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	93.59	93.63	96.69	99.96	93.63
Dash line	13.69	99.22	24.08	14.79	64.67
Long line	13.21	98.28	23.34	13.61	82.00
Small dash line	6.36	99.38	11.96	6.45	81.70
Turn signs	1.88	99.87	3.69	2.28	9.65
Other signs	0.42	98.12	0.84	0.43	51.34
Crosswalk	1.70	99.85	3.34	1.70	95.24
Stop line	6.13	99.70	11.55	6.24	78.16
Zebra zone	0.10	99.35	0.20	0.10	5.77
No parking zone	2.74	99.95	5.32	2.80	52.59
Parking space	0.88	99.72	1.74	0.89	38.43
Other lane-marking	0.14	99.85	0.28	0.15	1.40

Table A.8: Class-specific performance outcomes of the PSPNet model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	85.64	85.72	92.26	99.90	85.71
Dash line	5.14	98.94	9.78	5.84	29.95
Long line	3.80	98.25	7.32	4.40	21.66
Small dash line	0.55	98.80	1.09	0.57	12.87
Turn signs	1.65	99.80	3.24	1.85	13.21
Other signs	0.62	99.86	1.24	0.70	5.69
Crosswalk	0.71	99.75	1.41	0.71	67.31
Stop line	5.75	99.89	10.87	6.77	27.65
Zebra zone	0.04	99.72	0.09	0.05	1.07
No parking zone	0.11	98.49	0.23	0.11	68.99
Parking space	0.11	97.83	0.23	0.11	38.03
Other lane-marking	0.03	93.53	0.07	0.03	14.80

Table A.9: Class-specific performance outcomes of the SegNeXt model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	99.18	99.19	99.59	99.80	99.38
Dash line	34.71	99.74	51.54	39.87	72.85
Long line	33.29	99.59	49.95	40.67	64.70
Small dash line	23.31	99.87	37.81	24.99	77.58
Turn signs	8.76	99.97	16.12	56.02	9.41
Other signs	45.24	99.99	62.30	56.32	69.70
Crosswalk	28.94	100.00	44.89	70.02	33.03
Stop line	47.93	99.98	64.8	74.05	57.61
Zebra zone	36.63	99.99	53.62	53.14	54.11
No parking zone	21.50	100.00	35.39	26.80	52.12
Parking space	1.73	99.99	3.40	21.97	1.84
Other lane-marking	5.85	99.98	11.05	24.11	7.17

Table A.10: Class-specific performance outcomes of the Twins model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	99.14	99.14	99.57	99.86	99.27
Dash line	40.24	99.78	57.39	45.03	79.09
Long line	29.04	99.41	45.00	32.10	75.27
Small dash line	34.23	99.92	51.00	36.91	82.46
Turn signs	13.60	99.97	23.94	53.28	15.44
Other signs	31.14	99.99	47.49	58.17	40.12
Crosswalk	17.93	100.00	30.41	44.89	23.00
Stop line	45.76	99.99	62.79	84.82	49.84
Zebra zone	3.37	99.99	6.53	58.63	3.46
No parking zone	25.89	100.00	41.13	34.83	50.22
Parking space	4.98	99.99	9.49	43.56	5.32
Other lane-marking	16.06	99.98	27.67	29.64	25.95

Table A.11: Class-specific performance outcomes of the Swin model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	99.08	99.09	99.54	99.90	99.19
Dash line	36.61	99.73	53.60	40.05	81.01
Long line	35.72	99.57	52.64	40.65	74.65
Small dash line	20.35	99.81	33.82	20.64	93.63
Turn signs	22.97	99.97	37.36	47.33	30.86
Other signs	34.55	99.98	51.35	37.48	81.53
Crosswalk	13.20	100.00	23.31	56.46	14.69
Stop line	30.07	99.95	46.23	30.58	94.69
Zebra zone	38.22	99.99	55.30	47.28	66.60
No parking zone	16.30	99.99	28.03	18.81	55.03
Parking space	12.89	99.99	22.83	30.61	18.21
Other lane-marking	6.17	99.98	11.62	13.40	10.25

Table A.12: Class-specific performance outcomes of the SegFormer model trained on the SkyScapes Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	99.56	99.56	99.78	99.80	99.75
Dash line	51.19	99.88	67.71	67.33	68.09
Long line	41.94	99.71	59.10	53.16	66.53
Small dash line	43.55	99.95	60.67	50.26	76.53
Turn signs	7.52	99.98	13.99	73.74	7.73
Other signs	38.45	99.99	55.54	57.14	54.04
Crosswalk	3.51	100.00	6.78	65.70	3.58
Stop line	59.31	99.99	74.46	75.79	73.18
Zebra zone	14.01	99.99	24.58	58.64	15.55
No parking zone	32.84	100.00	49.45	48.84	50.07
Parking space	9.35	99.99	17.10	74.90	9.65
Other lane-marking	1.48	99.99	2.92	46.66	1.51

Appendix B. Benchmark of all 12 models on the Waterloo Urban Scene Dataset over all 15 classes [4.4]

Table B.1: Class-specific performance outcomes of the FCN model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	97.39	97.76	98.68	98.37	98.99
Road	80.20	97.75	89.01	96.62	82.51
Traffic Island	81.16	99.91	89.60	88.70	90.52
Sidewalk	79.25	99.40	88.42	85.87	91.14
Crosswalk	31.61	99.97	48.04	31.79	98.26
Dash line	28.25	99.93	44.06	28.31	99.27
Single solid line	33.99	99.72	50.73	34.07	99.31
Small dash line	12.09	99.99	21.58	12.36	85.03
Turn sign	27.57	100.00	43.23	27.57	100.00
Stop line	44.05	99.99	61.16	44.29	98.81
Zebra line	31.58	99.99	48.01	31.58	100.00
Parking line	27.37	99.82	42.98	29.70	77.75
No parking zone	17.49	99.84	29.77	17.61	96.20
Other	4.33	99.99	8.30	4.34	94.37
Vehicle	69.70	99.62	82.15	74.75	91.17

Table B.2: Class-specific performance outcomes of the FastFCN model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	98.84	99.01	99.42	99.72	99.11
Road	89.19	98.78	94.29	97.83	90.99
Traffic Island	88.39	99.94	93.84	88.80	99.49
Sidewalk	85.27	99.57	92.05	86.17	98.79
Crosswalk	33.98	99.97	50.73	34.00	99.87
Dash line	29.58	99.94	45.65	29.60	99.75
Single solid line	38.62	99.77	55.72	38.76	99.10
Small dash line	13.12	99.99	23.20	13.12	100.00
Turn sign	33.59	100.00	50.29	33.59	100.00
Stop line	46.42	99.99	63.41	46.43	99.96
Zebra line	22.92	99.99	37.29	22.92	100.00
Parking line	22.27	99.70	36.43	22.42	97.11
No parking zone	67.60	99.98	80.67	67.97	99.21
Other	12.71	100.00	22.56	12.71	100.00
Vehicle	74.65	99.69	85.49	77.91	94.69

Table B.3: Class-specific performance outcomes of the U-Net model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	96.86	97.30	98.40	98.62	98.19
Road	79.86	97.68	88.80	95.31	83.12
Traffic Island	74.48	99.85	85.37	75.55	98.14
Sidewalk	70.48	99.00	82.68	73.41	94.64
Crosswalk	30.40	99.97	46.62	31.28	91.46
Dash line	38.28	99.96	55.37	39.16	94.45
Single solid line	44.26	99.82	61.36	44.97	96.55
Small dash line	7.89	99.99	14.63	8.16	70.98
Turn sign	20.27	100.00	33.71	20.27	100.00
Stop line	21.82	99.97	35.82	21.88	98.62
Zebra line	60.21	100.00	75.16	61.00	97.88
Parking line	26.60	99.81	42.03	28.55	79.62
No parking zone	32.05	99.95	48.55	37.67	68.26
Other	0.90	99.94	1.78	0.90	100.00
Vehicle	65.24	99.51	78.97	67.47	95.18

Table B.4: Class-specific performance outcomes of the DeepLabV3 model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	98.01	98.31	99.00	99.63	98.37
Road	85.01	98.26	91.90	94.84	89.14
Traffic Island	82.53	99.91	90.43	85.93	95.43
Sidewalk	74.03	99.14	85.07	75.68	97.13
Crosswalk	24.27	99.95	39.06	24.30	99.50
Dash line	24.59	99.92	39.47	24.61	99.71
Single solid line	31.45	99.69	47.85	31.58	98.74
Small dash line	12.06	99.99	21.52	12.06	100.00
Turn sign	30.93	100.00	47.25	30.93	100.00
Stop line	37.94	99.98	55.01	38.17	98.49
Zebra line	44.16	100.00	61.27	44.16	100.00
Parking line	24.31	99.79	39.11	26.24	76.74
No parking zone	56.50	99.98	72.21	60.08	90.47
Other	22.58	100.00	36.84	24.53	73.94
Vehicle	64.91	99.52	78.72	68.86	91.89

Table B.5: Class-specific performance outcomes of the DeepLabV3+ model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	97.72	98.06	98.85	99.25	98.45
Road	85.56	98.32	92.22	94.76	89.81
Traffic Island	80.65	99.90	89.29	83.47	95.98
Sidewalk	75.38	99.21	85.96	78.25	95.35
Crosswalk	35.78	99.97	52.70	35.84	99.51
Dash line	42.47	99.96	59.62	42.55	99.54
Single solid line	44.37	99.82	61.47	44.58	98.94
Small dash line	11.45	99.99	20.54	11.45	100.00
Turn sign	29.57	100.00	45.64	29.57	100.00
Stop line	52.21	99.99	68.60	52.30	99.68
Zebra line	43.63	100.00	60.75	43.63	100.00
Parking line	31.56	99.84	47.98	33.95	81.77
No parking zone	52.99	99.98	69.27	62.48	77.72
Other	11.37	100.00	20.42	11.37	100.00
Vehicle	70.69	99.63	82.83	74.48	93.29

Table B.6: Class-specific performance outcomes of the ANN model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	98.34	98.59	99.16	99.84	98.49
Road	87.00	98.51	93.05	95.81	90.44
Traffic Island	81.71	99.90	89.93	82.09	99.44
Sidewalk	77.44	99.28	87.29	78.79	97.85
Crosswalk	23.91	99.95	38.60	23.92	99.95
Dash line	25.19	99.92	40.24	25.25	99.07
Single solid line	30.29	99.67	46.49	30.42	98.61
Small dash line	11.12	99.99	20.01	11.12	100.00
Turn sign	28.95	100.00	44.90	28.95	100.00
Stop line	42.18	99.99	59.33	42.33	99.19
Zebra line	31.04	99.99	47.38	31.04	100.00
Parking line	23.04	99.78	37.45	24.90	75.52
No parking zone	65.30	99.98	79.01	70.13	90.45
Other	5.83	99.99	11.01	5.83	100.00
Vehicle	66.12	99.53	79.60	68.28	95.43

Table B.7: Class-specific performance outcomes of the MobileNetV3 model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	96.24	96.79	98.08	99.19	97.00
Road	77.52	97.34	87.34	92.26	82.92
Traffic Island	50.81	99.61	67.39	53.35	91.44
Sidewalk	62.59	98.60	76.99	65.70	92.96
Crosswalk	31.86	99.97	48.32	32.41	94.96
Dash line	27.19	99.93	42.75	27.33	98.18
Single solid line	37.82	99.77	54.88	38.41	96.10
Small dash line	4.81	99.99	9.17	4.91	70.07
Turn sign	20.42	100.00	33.91	20.42	100.00
Stop line	14.78	99.95	25.75	14.79	99.68
Zebra line	2.32	99.85	4.53	2.32	100.00
Parking line	16.73	99.63	28.67	17.29	83.91
No parking zone	25.63	99.95	40.80	35.84	47.35
Other	0.46	99.88	0.91	0.46	100.00
Vehicle	54.91	99.26	70.89	56.96	93.85

Table B.8: Class-specific performance outcomes of the PSPNet model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	98.45	98.69	99.22	99.76	98.69
Road	88.68	98.70	94.00	95.77	92.30
Traffic Island	84.76	99.92	91.75	85.33	99.22
Sidewalk	79.40	99.36	88.52	80.83	97.82
Crosswalk	31.29	99.97	47.66	31.29	99.95
Dash line	26.47	99.92	41.86	26.49	99.71
Single solid line	39.04	99.78	56.16	39.40	97.73
Small dash line	7.41	99.99	13.80	7.41	100.00
Turn sign	35.47	100.00	52.37	35.47	100.00
Stop line	47.40	99.99	64.31	47.50	99.53
Zebra line	36.47	99.99	53.44	36.47	100.00
Parking line	27.36	99.81	42.97	29.23	81.03
No parking zone	72.34	99.99	83.95	75.54	94.48
Other	12.11	100.00	21.61	12.12	99.65
Vehicle	70.52	99.63	82.71	75.88	90.89

Table B.9: Class-specific performance outcomes of the SegNeXt model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	97.93	98.24	98.95	99.92	98.01
Road	87.24	98.49	93.19	93.20	93.18
Traffic Island	86.84	99.93	92.95	87.30	99.39
Sidewalk	74.39	99.15	85.31	75.28	98.43
Crosswalk	55.52	99.99	71.40	56.81	96.08
Dash line	48.87	99.97	65.65	49.25	98.45
Single solid line	39.57	99.78	56.71	39.82	98.49
Small dash line	42.36	100.00	59.52	42.83	97.51
Turn sign	68.89	100.00	81.58	69.17	99.42
Stop line	77.83	100.00	87.53	78.47	98.96
Zebra line	65.12	100.00	78.88	65.68	98.72
Parking line	26.74	99.78	42.20	27.62	89.39
No parking zone	80.91	99.99	89.45	84.82	94.62
Other	65.87	100.00	79.42	67.49	96.48
Vehicle	68.42	99.58	81.25	71.00	94.96

Table B.10: Class-specific performance outcomes of the Twins model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	94.68	95.45	97.27	98.99	95.60
Road	74.60	96.77	85.45	85.01	85.90
Traffic Island	57.92	99.68	73.35	58.33	98.80
Sidewalk	46.98	97.54	63.93	50.76	86.32
Crosswalk	61.53	99.99	76.18	70.18	83.30
Dash line	65.75	99.99	79.33	70.96	89.94
Single solid line	56.50	99.90	72.20	59.44	91.94
Small dash line	57.50	100.00	73.02	79.10	67.80
Turn sign	77.20	100.00	87.13	80.65	94.75
Stop line	78.27	100.00	87.81	80.41	96.71
Zebra line	59.34	100.00	74.48	60.47	96.93
Parking line	27.81	99.83	43.52	30.96	73.27
No parking zone	48.53	99.97	65.35	53.69	83.47
Other	78.85	100.00	88.17	89.78	86.62
Vehicle	57.44	99.35	72.96	60.91	90.97

Table B.11: Class-specific performance outcomes of the Swin model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	99.19	99.31	99.59	99.91	99.27
Road	93.36	99.25	96.56	97.98	95.19
Traffic Island	90.64	99.95	95.09	90.86	99.74
Sidewalk	88.59	99.68	93.95	90.00	98.27
Crosswalk	43.41	99.98	60.54	43.41	99.97
Dash line	46.15	99.97	63.15	46.19	99.78
Single solid line	53.18	99.87	69.43	53.33	99.45
Small dash line	30.48	100.00	46.72	30.48	100.00
Turn sign	42.35	100.00	59.50	42.35	100.00
Stop line	59.87	99.99	74.90	59.88	99.98
Zebra line	45.27	100.00	62.32	45.27	100.00
Parking line	38.89	99.87	56.00	39.68	95.09
No parking zone	70.23	99.99	82.51	70.53	99.40
Other	25.20	100.00	40.26	25.20	100.00
Vehicle	76.52	99.71	86.70	77.60	98.21

Table B.12: Class-specific performance outcomes of the SegFormer model trained on the Waterloo Urban Scene Dataset and assessed through a range of evaluation metrics.

Class	IoU (%)	Acc (%)	F1-score (%)	Precision (%)	Recall (%)
Background	98.63	98.84	99.31	99.93	98.70
Road	91.54	99.03	95.58	96.29	94.89
Traffic Island	90.32	99.95	94.92	90.84	99.38
Sidewalk	80.16	99.38	88.99	80.98	98.75
Crosswalk	70.33	99.99	82.58	70.74	99.17
Dash line	65.87	99.99	79.43	66.24	99.17
Single solid line	55.47	99.89	71.36	55.73	99.18
Small dash line	78.11	100.00	87.71	81.12	95.46
Turn sign	79.58	100.00	88.63	80.33	98.83
Stop line	89.74	100.00	94.59	90.96	98.54
Zebra line	68.38	100.00	81.22	68.44	99.89
Parking line	32.16	99.82	48.66	32.66	95.45
No parking zone	82.83	99.99	90.61	83.70	98.77
Other	85.17	100.00	91.99	89.11	95.07
Vehicle	73.30	99.66	84.60	74.51	97.83

Appendix C. Loss Functions for All 12 Models

Table C.1: Loss functions utilized across all 12 models.

Loss / Model	Cross Entropy	Dice	Focal	Tversky
FCN	✓	✓		
FastFCN	✓	✓		
U-Net	✓	✓		
DeepLabV3	✓	✓		
DeepLabV3⁺	✓	✓		
ANN	✓	✓	✓	
MobileNetV3	✓	✓		
PSPNet	✓	✓		
SegNeXt	✓			✓
Twins	✓			✓
Swin	✓	✓	✓	
SegFormer	✓			✓

Formula for per-pixel loss: The loss is calculated by summing across different classes, represented by the index c. For each class, yc is the binary ground truth indicator for a pixel, where 1 indicates the pixel belongs to class c and 0 indicates it does not. $y\hat{}c$ represents the predicted probability that the pixel belongs to class c. These individual pixel losses are then averaged across the entire image to determine the overall loss.

Cross-Entropy Loss: This function assesses the precision with which a classification model predicts probabilities for each class. It involves a comparison between these probabilistic predictions and the actual class labels. A lower value of the cross-entropy loss signifies enhanced model performance, rendering this function particularly advantageous for tasks centered on category prediction.

$$L_{CE} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{C.1}$$

Dice Loss: This function is particularly relevant for image segmentation tasks, as it measures the overlap between predicted and actual segments. It is highly valued for applications that require precise segmentations.

$$L_{Dice} = 1 - \frac{2\sum_{c=1}^{C} y_c \hat{y}_c}{\sum_{c=1}^{C} y_c^2 + \sum_{c=1}^{C} \hat{y}_c^2} \tag{C.2}$$

Focal Loss: This variant of cross-entropy loss increases focus on correcting misclassified examples, making it particularly effective for training on imbalanced datasets. It is often utilized in object detection scenarios where certain objects may appear infrequently.

$$L_{Focal} = -\sum_{c=1}^{C} \alpha_c y_c (1 - \hat{y}_c)^{\gamma} \log(\hat{y}_c) \tag{C.3}$$

Where αc is a weighting factor for class cc, which helps in adjusting the importance of each class in the loss function. γ is the focusing parameter, which adjusts the rate at which easy examples are down weighted. The higher the value of γ, the more the focus is on hard, misclassified examples.

Tversky Loss: Serving as a replacement for the frequently cited Soft IoU Loss, Tversky loss emphasizes the overlap between predicted and actual areas, particularly in object detection tasks. Its smooth and differentiable nature makes it ideal for gradient-based optimization, thereby enhancing object localization accuracy.

$$L_{Tversky} = 1 - \frac{\sum_{c=1}^{C} y_c \hat{y}_c}{\sum_{c=1}^{C} y_c \hat{y}_c + \alpha \sum_{c=1}^{C} (1 - y_c) \hat{y}_c + \beta \sum_{c=1}^{C} y_c (1 - \hat{y}_c)} \quad \text{(C.4)}$$

Where α weights the false positives (the penalty for predicting a class when it is not present), and β weights the false negatives (the penalty for not predicting a class when it is present).