

UNDERGRADUATE PROJECT REPORT

|  |  |
| --- | --- |
| **Project Title:** | **Semantic Segmentation on Deep Learning Application**  **--Segmentation of Urban Landscapes** |
| **Surname:** | **Feng** |
| **First Name:** | **Xiang** |
| **Student Number:** | **202018010119** |
| **Supervisor Name:** | **Dr Grace Ugochi Nneji** |
| **Module Code:** | **CHC 6096** |
| **Module Name:** | **Project** |
| **Date Submitted:** | **May 5, 2023** |

# **Declaration**

**Student** **Conduct** **Regulations:**

Please ensure you are familiar with the regulations in relation to Academic Integrity. The University takes this issue very seriously and students have been expelled or had their degrees withheld for cheating in assessment. It is important that students having difficulties with their work should seek help from their tutors rather than be tempted to use unfair means to gain marks. Students should not risk losing their degree and undermining all the work they have done towards it. You are expected to have familiarised yourself with these regulations.

<https://www.brookes.ac.uk/regulations/current/appeals-complaints-and-conduct/c1-1/>

Guidance on the correct use of references can be found on www.brookes.ac.uk/services/library, and also in a handout in the Library.

The full regulations may be accessed on-line at <https://www.brookes.ac.uk/students/sirt/student-conduct/> . If you do not understand what any of these terms mean, you should ask your Project Supervisor to clarify them for you.

**I** **declare** **that** **I** **have** **read** **and** **understood** **Regulations** **C1.1.4** **of** **the** **Regulations** **governing** **Academic** **Misconduct,** **and** **that** **the** **work** **I** **submit** **is** **fully** **in** **accordance** **with** **them**.

Signature: Data: 22/04/2024

REGULATIONS GOVERNING THE DEPOSIT AND USE OF OXFORD BROOKES UNIVERSITY MODULAR PROGRAMME PROJECTS AND DISSERTATIONS

Copies of projects/dissertations, submitted in fulfilment of Modular Programme requirements and achieving marks of 60% or above, shall normally be kept by the Library.

**|** **agree** **that** **this** **dissertation** **may** **be** **available** **for** **reading** **and** **photocopying** **in** **accordance** **with** **the** **Regulations** **governing** **use** **of** **the** **Library.**

Signature: Data: 22/04/2024

# **Acknowledgment**

I want to express my heartfelt gratitude to my Supervisor, Dr. Grace Ugochi Nneji, for her guidance and support throughout my final undergraduate project. Her expertise, patience, and encouragement have been invaluable in shaping this project. Also, I would like to thank the module leader Joojo Walker and other teachers throughout my undergraduate study time for their teaching and advice.

More so, I would like to acknowledge the resources and facilities provided by Oxford Brookes University in collaboration with the Chengdu University of Technology for an outstanding opportunity.

And to my family and friends, your unending love and encouragement have been with me. I will always be grateful.

# **Table of Contents**

**[Declaration](#_Toc129605831)** [i](#_Toc129605831)

**[Acknowledgment](#_Toc129605832)** [ii](#_Toc129605832)

**[Table of Contents](#_Toc129605833)** [iii](#_Toc129605833)

**[Abstract](#_Toc129605834)** [iv](#_Toc129605834)

**[Abbreviations](#_Toc129605835)** [v](#_Toc129605835)

**[Glossary](#_Toc129605836)** [vi](#_Toc129605836)

**[Chapter 1 Introduction](#_Toc129605837)** [1](#_Toc129605837)

**[1.1](#_Toc129605838)****[Background](#_Toc129605838)** [1](#_Toc129605838)

**[1.2](#_Toc129605839)****[Aim](#_Toc129605839)** [1](#_Toc129605839)

**[1.3](#_Toc129605840)****[Objectives](#_Toc129605840)** [1](#_Toc129605840)

**[1.4](#_Toc129605841)****[Project Overview](#_Toc129605841)** [1](#_Toc129605841)

**[1.4.1](#_Toc129605842)****[Scope](#_Toc129605842)** [1](#_Toc129605842)

**[1.4.2](#_Toc129605843)****[Audience](#_Toc129605843)** [1](#_Toc129605843)

**[Chapter 2 Background Review](#_Toc129605844)** [2](#_Toc129605844)

**[Chapter 3 Methodology](#_Toc129605845)** [3](#_Toc129605845)

**[3.1](#_Toc129605846)****[Approach](#_Toc129605846)** [3](#_Toc129605846)

**[3.2](#_Toc129605847)****[Technology](#_Toc129605847)** [3](#_Toc129605847)

**[3.3](#_Toc129605848)****[Project Version Management](#_Toc129605848)** [3](#_Toc129605848)

**[Chapter 4 Results](#_Toc129605849)** [4](#_Toc129605849)

**[Chapter 5 Professional Issues](#_Toc129605850)** [5](#_Toc129605850)

**[5.1](#_Toc129605851)****[Project Management](#_Toc129605851)** [5](#_Toc129605851)

**[5.1.1](#_Toc129605852)****[Activities](#_Toc129605852)** [5](#_Toc129605852)

**[5.1.2](#_Toc129605853)****[Schedule](#_Toc129605853)** [5](#_Toc129605853)

**[5.1.3](#_Toc129605854)****[Project Data Management](#_Toc129605854)** [5](#_Toc129605854)

**[5.1.4](#_Toc129605855)****[Project Deliverables](#_Toc129605855)** [5](#_Toc129605855)

**[5.2](#_Toc129605856)****[Risk Analysis](#_Toc129605856)** [5](#_Toc129605856)

**[5.3](#_Toc129605857)****[Professional Issues](#_Toc129605857)** [5](#_Toc129605857)

**[Chapter 6 Conclusion](#_Toc129605858)** [6](#_Toc129605858)

**[References](#_Toc129605859)** [7](#_Toc129605859)

**[6.1](#_Toc129605860)****[Formatting Requirements](#_Toc129605860)** [7](#_Toc129605860)

**[6.2](#_Toc129605861)****[Written Presentation](#_Toc129605861)** [7](#_Toc129605861)

**[Appendices](#_Toc129605862)** [8](#_Toc129605862)

# **Abstract**

Semantic segmentation is an important research direction in the field of computer vision , which plays a key role in many daily life applications, such as self-driving, medical image analysis and video monitoring. Nevertheless, today's semantic segmentation techniques still face challenges such as complex scene structure, multi-scale object recognition and changing environmental conditions, which bring very huge disturbances to semantic segmentation. To deal with these challenges, this study designs a high-performance semantic segmentation framework. The model does this by fusing multiple advanced deep learning techniques, such as using the ASPP module with deeply separable convolution to capture multi-scale contextual information, adopting the Transformer module to enhance model's ability to capture global dependency, and applying multi-scale pooling to optimize the model's ability to process features at different scales. This module fusion method significantly improves the accuracy and robustness of semantic segmentation, and achieves excellent segmentation results on diverse datasets while increasing the segmentation speed. By evaluating the model on test datasets, the model demonstrates its excellent performance in handling complex image segmentation tasks. The core contribution of this research is to propose an efficient and accurate semantic segmentation framework integrating deep separable convolution, ASPP, Transformer and multi-scale pooling, which brings new research directions and application potentials to the field of semantic segmentation.

Keywords: Semantic Segmentation, Computer Vision, Deep Learning, Technological integration, ASPP, Transformer Module, Multi-scale Pooling

# **Abbreviations**

CNN: Convolutional Neural Network

ASPP: Atrous Spatial Pyramid Pooling

ResNet: Residual Network

MIOU: Mean Intersection Over Union

GAP: Global Average Pooling

DWC: Depth-wise Convolution

SGD: Stochastic Gradient Descent

PA: Pixel Accuracy

CPA: Class Pixel Accuracy

FPS: Frames Per Second

ReLU: Rectified Linear Unit

BN: Batch Normalization

TN: True Negative

TP: True Positive

FP: False Positive

FN: False Negative

GUI: Graphical User Interface

GPU: Graphics Processing Unit

# **Glossary**

**Semantic Segmentation :** This is a technique that segments an image into multiple parts and each part is labeled with different categories for understanding scenes and objects in the image.

**Deep Learning :** Learning from large amounts of data by using networks with multiple layers of processing units, and is widely used for tasks such as image recognition, speech recognition, and semantic segmentation.

**Convolutional Neural Network (CNN) :** A convolutional neural network is a deep learning model specifically designed to process data that has a clear grid structure, such as images. By stacking convolutional, pooling, and fully connected layers, CNN are able to extract features from images efficiently and use them for tasks such as classification, detection, and segmentation.

**Image Augmentation :** This is a technique where a series of transformations (e.g., rotation, scaling, cropping, etc.) are applied to the training image to increase the variety of the dataset artificially. This helps the model learn more generalized features, which improves performance on unseen data.

**Feature Extraction :**In deep learning, feature extraction usually means using high-level features learned from previous layers of the model that are helpful for the task at hand.

**Muti-scale Feature Fusion :** This is a technique that can combine features from different resolutions of an image to capture information ranging from coarse to detailed.

**Attention Mechanism :** Attention Mechanism allows deep learning models to dynamically focus on important parts of the input data. It plays a key role in improving the explainable and performance of the model.

**Stochastic Gradient Descent (SGD) :** It is an optimization algorithm used to minimize the loss function of the model during the training process.The SGD optimizer can effectively reduce the consumption of computational resources and speed up the training process.

**Atrous Spatial Pyramid Pooling (ASPP) :** ASPP is a technique to capture multi-scale information by using null convolution with different sampling rates.ASPP can effectively improve the feature resolution and model performance in semantic segmentation tasks.

**Cross-entropy Loss :** It is a loss function commonly used in classification tasks to quantify the difference between the probability distribution predicted by the model and the true label. It is essential for training high-performance classification models.

**Dice Loss :** It is a loss function used in image segmentation tasks. It is particularly suitable for dealing with the problem of category imbalance, where the accuracy of segmentation is improved by optimizing the model to the overlap of the segmented regions of the Jung family.

# **Introduction**

## **Background**

With urban traffic congestion and frequent accidents becoming an increasing problem, self-driving vehicles are widely recognized as a potential solution to reduce accident rates and improve traffic flow **[1]**. However, the application of self-driving in urban environments faces a number of challenges, which include dealing with complex backgrounds, variable weather conditions, and diverse traffic scenarios **[2]**. These factors make it difficult for traditional computer vision techniques to be adapted, increasing the difficulty of realizing self-driving technology **[2]**.

In this case, semantic segmentation is especially crucial as a machine vision technique that allows a fine understanding of the surrounding environment **[3]**. It can relate each pixel in an image to a semantic category of roads, buildings, vehicles, and pedestrians **[4]**. This can help self-driving systems to recognize and understand their surroundings more accurately, providing a powerful environment perception tool for vehicles **[5]**.

Although semantic segmentation technology shows great potential in self-driving, it still faces many challenges in practical applications, including how to effectively deal with complex scene structures, adapt multi-scale object recognition, and deal with changing environmental conditions **[6]**. These challenges require semantic segmentation models not only have better accuracy and robustness, but also need to realize real-time and fast image processing to adapt the real-time decision-making requirements for self-driving vehicles **[7]**.

Therefore, it has become an urgent problem for how to make the semantic segmentation technology adapt the application requirements of self-driving vehicles in complex urban environments. This study is devoted to in-depth research on semantic segmentation technology, aiming to develop a high-accuracy real-time semantic segmentation model to provide safer and more efficient self-driving technology for urban transportation systems to promote sustainable urban development.

* + 1. **Convolutional Layers**

The convolutional layer is the core building block of a convolutional neural network and is responsible for executing most of the computations. It requires several components, including input data, filters, and feature maps. The input is assumed to be a color image, consisting of a three-dimensional matrix of pixels. This means that the input has three dimensions: height, width, and depth, corresponding to the RGB in the image. There is also a feature detector, also known as a kernel or filter, which moves through the various sense fields of the image, checking for the presence of features. This process is called convolution **[8]**.

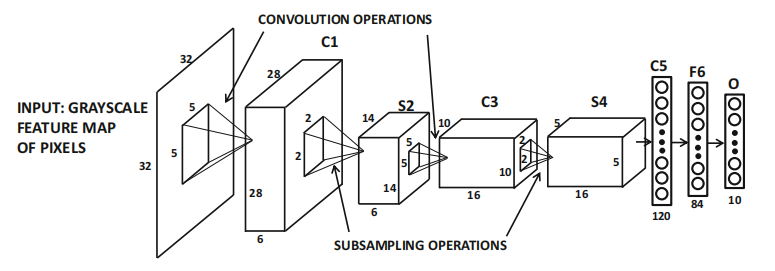


Figure 1: convolutional architecture[29]

* + 1. **Activation Functions**

Activation functions are a type of function added to an artificial neural network designed to help the network learn complex patterns in the data. The Sigmoid function compresses the input to the (0,1) interval, which is suitable for binary classification problems, but it is easy to cause gradient vanishing. Tanh function outputs in the range of (-1,1), and its output centering improves the stability of the model training, but there is the problem of gradient vanishing.The ReLU function is the most widely used. Input is greater than 0 to remain unchanged, the output is 0 when the input is less than 0. It solves the gradient disappearance problem, the training speed is fast **[8]**.

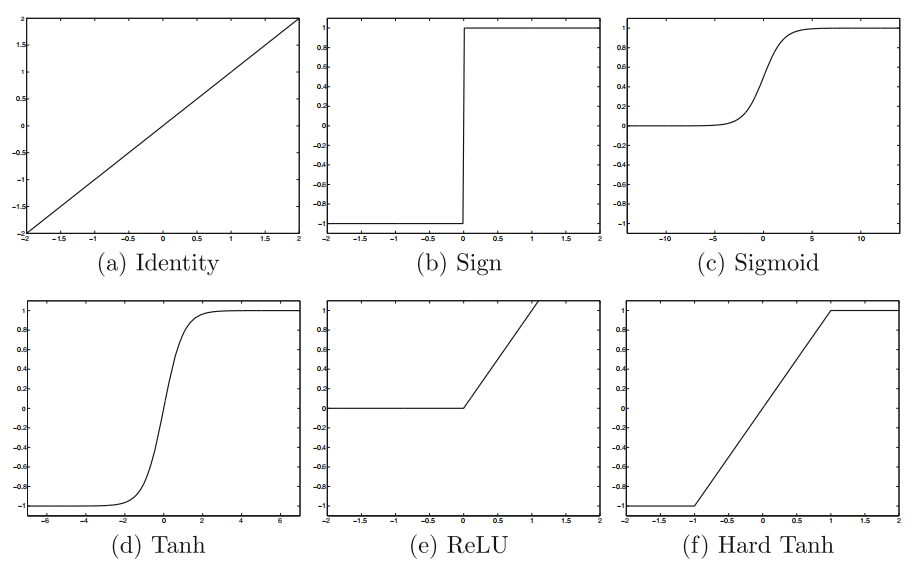


Figure 2: Activation function image[29]

* + 1. **Pooling Layers**

The pooling layer, also known as the downsampling layer, performs a dimensionality reduction operation designed to reduce the number of parameters in the input. Similar to the convolutional layer, the pooling operation lets the filter scan the entire input, but the difference is that this filter has no weights. The kernel applies an aggregation function to the values in the sensory field to populate the output array. There are two main types of pooling: 1. Maximum pooling: as the filter moves through the input, it selects the pixel with the largest value and sends it to the output array. 2. Average pooling: as the filter moves through the input, it calculates the average value in the receptive field and sends it to the output array. Although a lot of information is lost in the pooling layer, it still brings many benefits to the convolutional neural network. The layer helps reduce complexity, increase efficiency, and limit the risk of overfitting **[8]**.

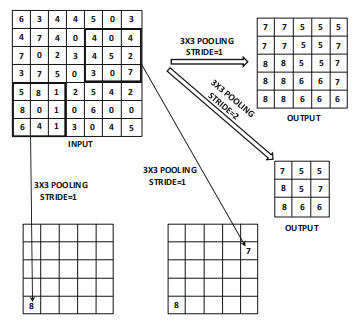


Figure 3: Pooling schematic[29]

* + 1. **Up-Sampling**

Up-sampling increases the size of the data by inserting new pixels or feature points, using methods such as transposed convolution to recover detail and support accurate prediction.

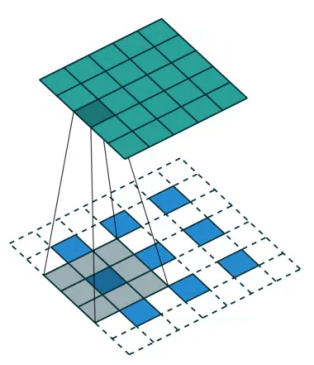


Figure 4: Up-Sample process

* + 1. **Batch Normalization**

Batch normalization is a technique used to improve the efficiency and stability of deep neural network training by normalizing the distribution of the inputs at each layer so that the mean is 0 and the standard deviation is 1. This helps to address the problem of internal covariate bias in training, allowing the network to use higher learning rates and be less sensitive to the choice of initialization weights. Batch normalization also indirectly provides a regularization effect that accelerates the convergence of the network and improves the generalization ability of the model **[8]**.

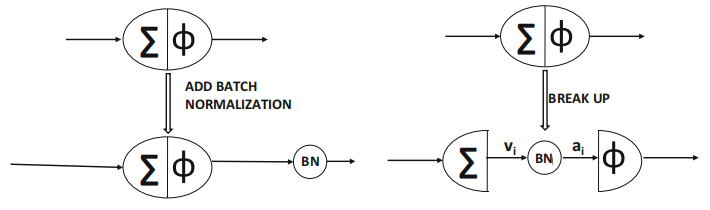


Figure 4: Batch Normalization

* + 1. **ASPP**

ASPP module improves the accuracy of semantic segmentation by processing images in parallel using null convolution with different sampling rates, effectively capturing information at different scales. This method expands the sensory field of the model, allowing it to simultaneously understand both details and broader contextual information in the image, and is particularly suitable for dealing with size variations in images, thus improving segmentation performance. In short, ASPP allows the model to better adapt to targets of different sizes and improve semantic segmentation **[9]**.

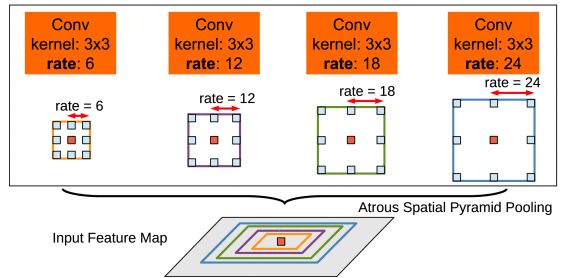


Figure 5: ASPP Module**[19]**

* + 1. **Loss Functions**

Cross-entropy Loss **[21]**: Cross-entropy loss measures the discrepancy between model predictions and actual labels, and is a key tool for optimizing model accuracy in tasks dealing with classification and semantic segmentation.

Dice Loss **[22]**: Dice Loss is a loss function based on Dice coefficients for evaluating the similarity of two samples, which is particularly suitable for the case of category imbalance in image segmentation, and significantly improves the accuracy and performance of semantic segmentation by optimizing the overlap rate between predicted and real labels.

* + 1. **Optimization Algorithms**

SGD Optimizer:SGD is an optimization algorithm commonly used for deep learning model training, which progressively reduces the value of the loss function by randomly selecting a sample of data to compute the gradient and updating the model parameters in each iteration.SGD aims to efficiently find the global optimal solution, and although the update at each step may not be as accurate as gradient descent based on the entire data, it is more efficient and practical when dealing with large datasets **[20]**.

* + 1. **LinkNet**

As shown in Figure 6, LinkNet is a light weight,efficient neural network structure for semantic segmentation that uses an encoder-decoder architecture. Special jump connections solve the gradient loss problem. The encoder catches image features and decoder recover image details. The encoder gradually reduces the image resolution and extracts the features and the decoder restores the resolution while fusing the features and assigning semantic categories to each pixel **[24]**.

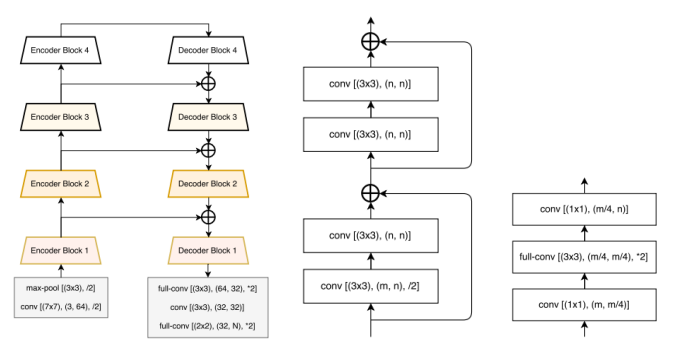


Figure 6. LinkNet structure diagram and encoder-decoder**[24]**

* + 1. **U-Net**

U-Net ass depicted in Figure 7 is a deep learning model for image segmentation. It uses an encoder-decoder architecture and jump connections to reduce images, extract features and recover to original size, which solves the gradient loss problem, improves accuracy and preserves features at all layers [25].U-Net is able to catch image features accurately, meanwhile its simple architecture is easy to modify.

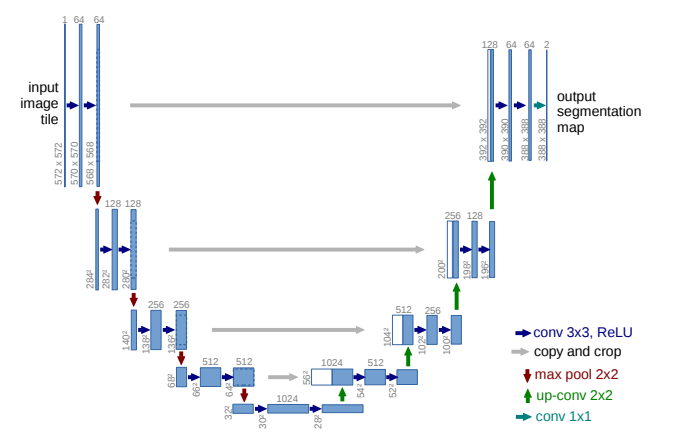


Figure 7. U-Net model **[25]**

* + 1. **Dilated Convolution**

In the fields of deep learning and computer vision, dilation convolution can be used to improve the performance of convolutional neural networks. It is used to significantly increase the size of the network's receptive field without adding additional computational burden by introducing predefined spacing "holes" between neighboring elements of the convolutional kernel. This means that the network is able to observe a wider region of the input data, rather than being limited to immediately neighboring pixels.**[27]**

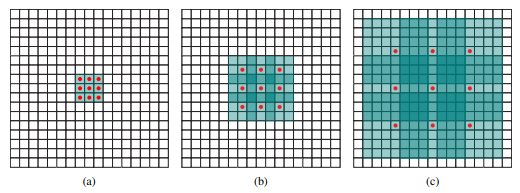


Figure 8. Dilated Convolution **[27]**

* + 1. **Transformer Encoder**

The Transformer module, originally designed to solve sequence-to-sequence tasks in natural language processing, has been successfully extended to the field of computer vision as a powerful tool complementary to traditional convolutional neural networks. Integrating the Transformer module into a CNN can greatly enhance the model's understanding of the global contextual information of an image. This combination exploits the efficiency of CNN in extracting local visual features and the power of Transformer in modeling long-range dependencies. In many visual tasks, such as image classification, object detection, and semantic segmentation, this combination provides a novel way to capture the details and overall structure of an image, thereby improving the performance of the model.**[28]**

## **Aim**

The primary aim of this project is to create a high-performance semantic segmentation model using deep learning techniques. The model integrates a variety of advanced deep learning techniques, including the ASPP module based on DWC, the Transformer module, and multi-scale pooling. It aims to significantly improve the accuracy and robustness of the semantic segmentation model in complex scene structures, multi-scale object recognition and diverse environmental conditions. In addition, to enhance the applicability and interactivity of the project, I plan to use Flask build a website to implement GUI, which will display the segmentation results of the semantic segmentation model, aiming able to intuitively see the results of the model processing various images. In conclusion, this project aims to provide a novel semantic segmentation framework and promote the development and popularization of semantic segmentation technology in practical applications by developing a user-friendly GUI.

## **Objectives**

In order to achieve the main goal of this project, the following specific objectives have been set:

* In-depth research and experimentation of advanced deep learning techniques: research and implement multiple deep learning modules, such as DWC-based ASPP module, Transformer module, edge detection module, and multi-scale pooling module, and try to integrate them. Aims to improve the performance of semantic segmentation models through these advanced techniques.
* Improve the accuracy and robustness of the model in complex environments: through the above techniques, the accuracy and stability of the model in complex scene structures, multi-scale object recognition, and diverse environmental conditions are improved to meet the common challenges in semantic segmentation.
* Evaluating model performance: evaluate the accuracy and robustness of the model using evaluation metrics (e.g., accuracy, recall, MIOU, PA, Dice coefficients, Kappa scores, FPS) by running the model on a test dataset. This will provide an objective benchmark for comparing different model architectures and configurations.
* Developing a Graphical User Interface (GUI): Build a website using the Flask framework to implement a GUI for presenting semantic segmentation results so that users can easily upload images, start the segmentation process and view the model outputs intuitively, enhancing the applicability and interactivity of the project.
* Contribute to the development and popularization of semantic segmentation technology: Through the development of an easy-to-use GUI and a high-performance semantic segmentation model, we provide new ideas and tools for the research, development, and application of semantic segmentation technology, and promote the development and popularization of this technology in practical applications.

## **Project Overview**

### **Scope**

The core goal of this project is to develop a high-performance semantic segmentation model using deep learning techniques, aiming to significantly improve accuracy and robustness in complex scenarios, multi-scale object recognition, and variable environmental conditions, which will contribute to the realization of self-driving technologies. The scope of the project includes the design and implementation of a semantic segmentation model that integrates multiple deep learning techniques, in particular the fusion of the DWC-based ASPP module, the Transformer module, and multi-scale pooling techniques. This integration strategy works on extracting richer image features as well as reducing the training parameters of the model to improve its efficiency and accuracy. In addition to this, the project creates a website to implement a GUI through the Flask framework to visualize the results of semantic segmentation to enhance the user's application experience and interactivity. Through the GUI, users can upload images for segmentation and instantly view the processing results of the model. Finally, the project will evaluate the performance of the model on a test dataset and compare it with other existing models to show the advantages of my model. Through the above efforts, this project is not only devoted to promote the innovation of semantic segmentation technology, but also to promote the popularization and application of this technology in real-world applications through the development of an easy-to-use GUI.

### **Audience**

For government and city planners, this project enhances traffic management efficiency. Realtime road condition sensing in autonomous driving can optimize traffic flow and reduce congestion **[2]**. Autonomous driving also mitigates accidents caused by human factors, enhancing road safety. For drivers, it ensures safe, fatigue-free driving, and selects optimal routes based on real-time conditions, saving time, reducing stress, and improving travel efficiency **[8]**. For the urban environment, improved traffic management and reduced congestion cut emissions, enhancing city air quality.

# **Background Review**

**2.1 Summary of Related Literature**

The research background of semantic segmentation techniques centers around a central task in the field of computer vision - understanding the category to which each pixel in an image belongs **[4]**. This technique enables computers not only to recognize the objects present in an image, but also to accurately classify the boundary of each object, which is an important foundation in the field of image processing and analysis **[3]**.

Nowadays, the introduce of deep learning, especially the application of CNN, greatly improves the accuracy and efficiency of semantic segmentation, allowing machines to process more complex image data, recognize and segment multiple objects in an image **[23]**. The progress of this technology not only promotes the development of the computer vision field, but also brings innovative possibilities for a number of application fields, such as self-driving, medical image analysis, and environmental monitoring **[23]**.

By combing through the related literature, I will compare the MIOU of some existing semantic segmentation models, and below is Table 1 which summarizes the performance comparison of different semantic segmentation models by different researchers.

|  |  |  |  |
| --- | --- | --- | --- |
| **Author** | **Model** | **MIou** | **Dataset** |
| Badrinarayanan et.al. **[9]** | SegNet | 56.1% | Cityscapes |
| Abdigapporov et.al.**[10]** | BiFPN | 56.4% | Cityscapes |
| Paszke et.al. **[11]** | ENet | 58.3% | Cityscapes |
| Poudel et.al. **[12]** | Fast-SCNN | 68% | Cityscapes |
| Yu et al **[13]** | BiSeNet | 69% | Cityscapes |
| Fourure et al **[14]** | GridNet | 69.5% | Cityscapes |
| Chen et al **[15]** | Deep-Lab CRF | 70.4% | Cityscapes |
| Lin et al **[16]** | RefineNet | 73.6% | Cityscapes |
| Li et.al.**[17]** | BiAttnNet | 74.7% | Cityscapes |
| My Model | DeepSegASPP+Transformer | 75% | Cityscapes |

**Table 1 Comparison of results**

First, the SegNet model proposed by Badrinarayanan et al **[9]**. is a convolutional network using a nonlinear upsampling technique that achieves a MIOU of 56.1%, which is more suitable for simple scene understanding. Later, the BiFPN network developed by Abdigapporov et al.**[10]** achieved a MIOU of 56.4% by enhancing the multi-scale fusion of features. ENet designed by Paszke et al. **[11]** is a lightweight network designed to satisfy the requirements of real-time applications with 58.3% MIOU, while Fast-SCNN proposed by Poudel et al. **[12]** achieves 68% MIOU by optimizing the computational efficiency, and both networks are suitable for semantic segmentation applications on mobile.

BiSeNet developed by Yu et al. **[13]** uses dual network structure to achieve a model that balances speed and performance with 69% MIOU. Fourure et al.**[14]**'s GridNet model enhances feature fusion through its grid structure and achieves 69.5% MIOU. Chen et al. **[15]'**s DeepLab-CRF, which is constructed using ASPP with CRF technology, achieves 69.5% MIOU on the edge processing was refined and achieved 70.4% MIOU. RefineNet proposed by Lin et al. **[16]** enhanced the information fusion by multipath refinement technique and achieved 73.6% MIOU. BiAttnNet model by Li et al. **[17]** with the introduction of bi-directional attention mechanism drastically improved the segmentation accuracy and achieved 74.7% MIOU.

My model obtained a higher MIOU on the Cityscapes dataset, which indicates that my model has a higher performance in recognizing complex urban scenes. This achievement not only outperforms other models in terms of accuracy, but also provides an important advancement in computer vision systems for complex urban scenes. With the ASPP module based on depth-separable convolution, my model effectively captures contextual information on multiple scales; while the Transformer part enhances the model's understanding between different regions of an image. Thus, compared to the work of other researchers, my model not only sets a new performance benchmark for semantic segmentation tasks, but also introduces an innovative approach to effectively utilize the cutting-edge techniques of deep learning to achieve a more accurate and efficient understanding of images.

# **Methodology**

## **Approach**

These following aspects will be followed in this project:

* The Cityscapes dataset is used, containing 2975 training images, 500 validation images and 500 test images.
* Data preprocessing includes data balancing and enhancement to improve model generalization. Specifically, evaluating pixel classes and ignoring classes with fewer pixels. Different weather is simulated to adjust the saturation, hue, contrast and random cropping of the images.
* The model will use ASPP module, Transformer encoder, Edge Detection and Contextual Enhancement modules and select Resnet50 as the backbone network.

## **Dataset**

The Cityscapes dataset is a large dataset focusing on urban street scenes, widely used in computer vision and autonomous driving research. It contains high-resolution images from 50 different cities and provides accurate pixel-level annotations for about 5,000 images covering 30 different categories, such as roads, pedestrians, etc. The dataset also includes a number of other images that have been annotated to provide a better understanding of the city's streetscape. In addition, the dataset includes about 20,000 roughly annotated images, enriching the training data. In this project 2975 images will be used as training dataset, 500 as validation and 500 as testing.

## **Pre-processing**

### **Data Balancing**

When dealing with the semantic segmentation dataset CityScapes, a crucial step is to evaluate the balance of the different categories in the dataset, which is usually achieved by counting the number of pixels in each category. This process involves traversing the annotations of all images in the dataset to count the number of pixels in each category. Given the large volume of the CityScapes training dataset, a representative validation dataset was selected for this analysis for pixel counting. By generating graphs, we can visualize the quantitative relationships between different categories in the dataset. The analysis results reveal that in the CityScapes dataset, the number of samples for certain categories such as roads and buildings is relatively high, while the number of samples for specific vehicle or pedestrian categories is low.In order to make the segmentation result more accurate, I will ignore some categories with small amount of pixels, so as to avoid the situation that the small category segmentation affects the large category segmentation.

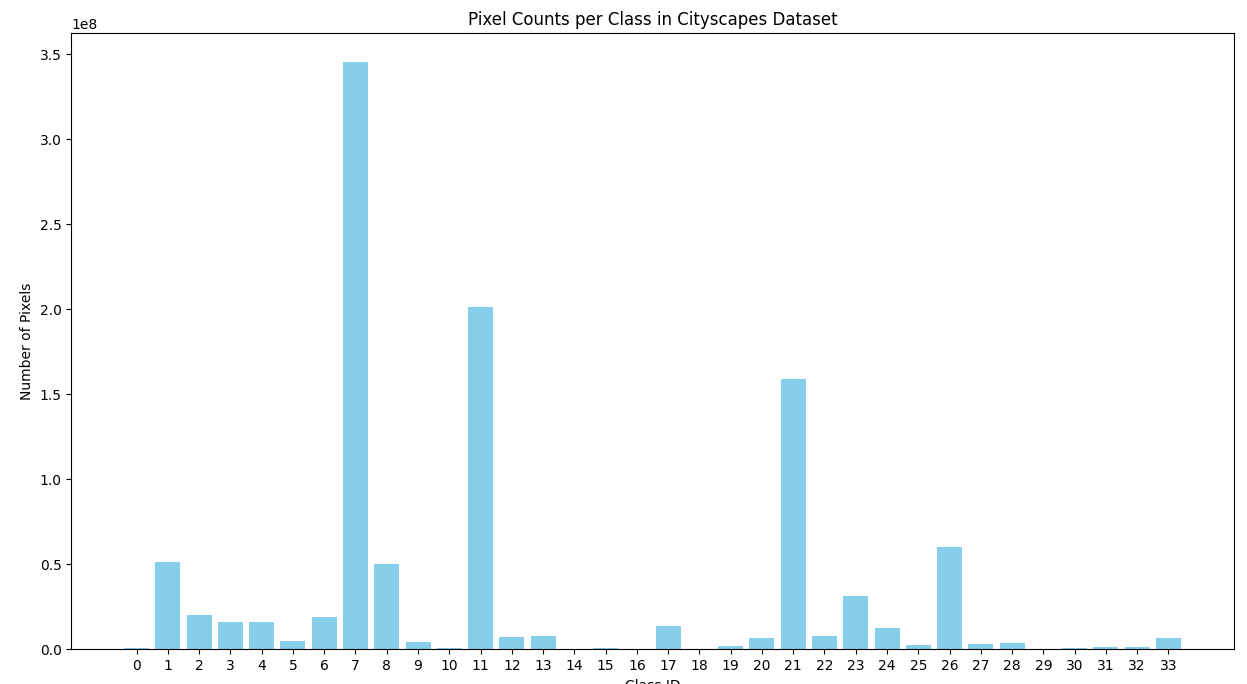
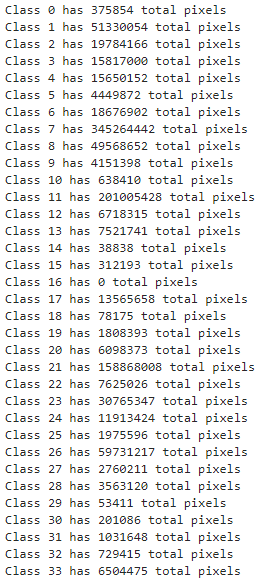


Figure 8. Pixel class profile of the original cityscapes dataset

### **Data Enhancement**

Data enhancement techniques play a key role in developing image segmentation models for urban scenes. Due to the variety of urban weather conditions, I first added a weather changing data enhancement operation to the original cityscapes dataset to simulate the visual effects under different weather conditions such as foggy, rainy and snowy days. This is shown in the figure below:



Figure 9.Weather Change Effect

In addition, to further improve the model's adaptability and robustness to features such as urban image lighting, I use several stochastic transformation techniques. A series of varied training samples are generated by adjusting the saturation, hue, and contrast of the images as well as implementing random cropping. Specifically, the saturation of the images was randomly varied between 0.5 and 1.5, the hue was varied up to a maximum range of 0.2, and the contrast was adjusted between 0.5 and 1.5 as a way of ensuring that the model was able to deal with images under different lighting conditions. In addition by randomly cropping images with 1024\*2048 resolution to 384\*384 size, this reduces the model training parameters while avoiding the interference of extreme values.

**original after**

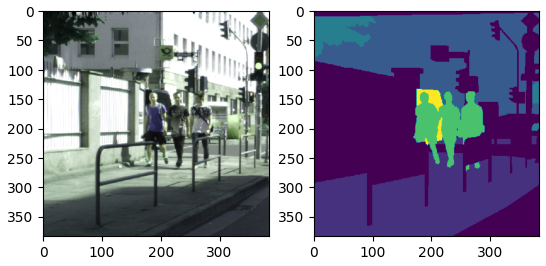
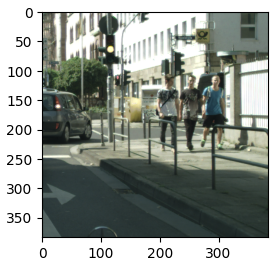


Figure 10. Image Enhancement Processing

In addition to this, I also tried to add an edge enhancement effect to the original image, a step designed to improve the model's ability to capture image edge information and further enhance the model's accuracy in recognizing the boundaries of objects in complex urban scenes. However, as far as the results of the training are concerned, the results of this are very unsatisfactory and this data enhancement approach is abandoned. The enhancement results are shown below:

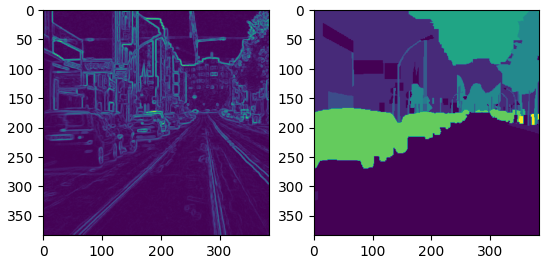


Figure 11. Edge Enhancement Processing

## **Component Modules and Model Architecture**

With the description of the aforementioned concepts, it becomes more straightforward to grasp the network model utilized in this project. The following section will provide a detailed introduction to the modules and architecture of the model employed within this project.

### **Convolution-based ASPP Module:**

The objective of the ASPP (Atrous Spatial Pyramid Pooling) module is to enhance the capture of multi-scale information from images, thereby increasing the precision of semantic segmentation. As depicted in Figure 6, this module represents a modified version of the ASPP module, where the original four dilated convolutions of varying dilation rates have been reduced to three layers, with dilation rates of 6, 12, and 18, respectively. This significant reduction in dilated convolution layers considerably decreases redundant computations and training parameters. Such a streamlined architecture not only diminishes the model's training duration but also aids in preventing over-fitting, markedly improving the model's generalization capability during segmentation. Moreover, the conventional convolution layers have been substituted with depthwise separable convolutions, which substantially reduce the model's parameter count and computational load. Compared to standard dilated convolutions, depthwise separable convolutions achieve better performance while significantly reducing computational complexity and memory requirements.

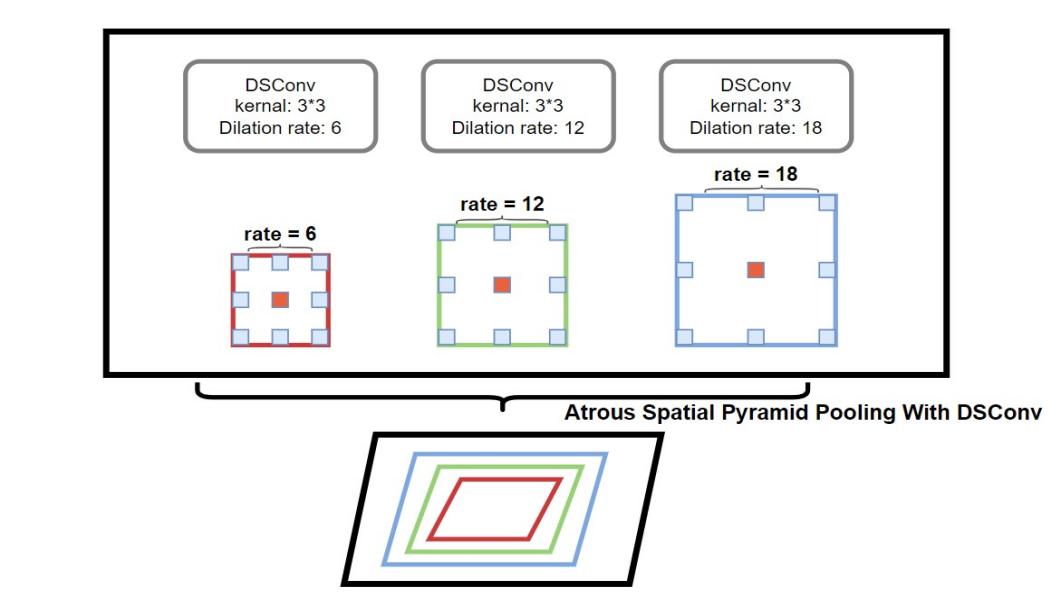


Figure 12. Modified ASPP module

Figure 7 illustrates the flowchart of the improved ASPP (Atrous Spatial Pyramid Pooling) module. Initially, the input X is fed into both an upsampled average pooling layer and four depthwise separable convolutions with varying dilation rates (DR=1, DR=6, DR=12, DR=18). Subsequently, the upsampled global context feature map (y\_pool) is fused with the outcomes of the four depthwise separable convolutions with different dilation rates (y1, y2, y3, y4). Finally, a rich set of multi-scale features is obtained as the output of this module. This module enhances the model's capability to perceive features of varying sizes, thereby improving its performance in processing objects across different scales.

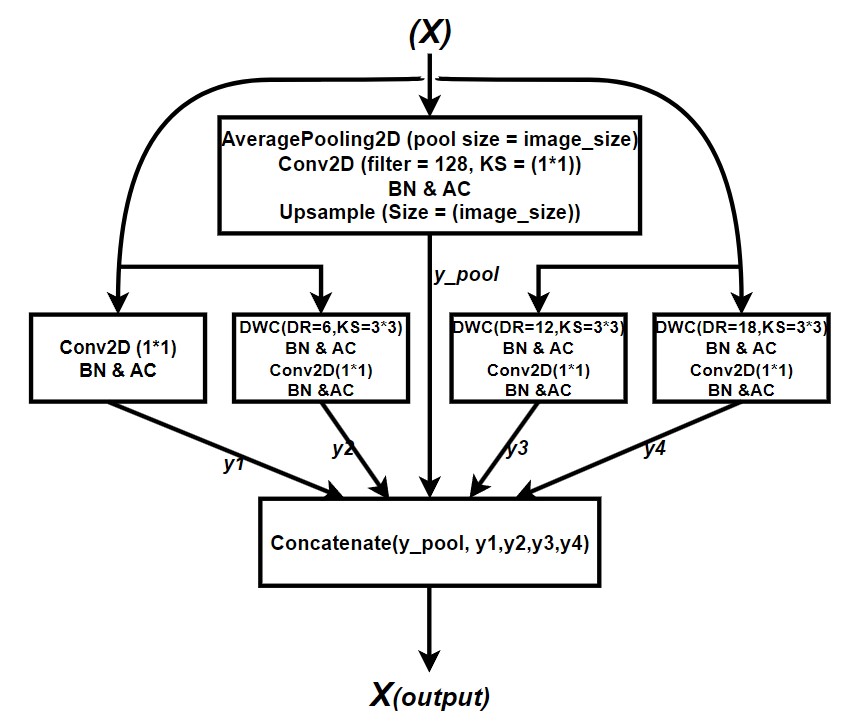


Figure 13. Modified ASPP Module Flow Diagram

### **Transformer Encoder Module:**

Figure 8 presents the flowchart of the Transformer Encoder module, whose input originates from the output of the ASPP module that has been reshaped into a sequence. Initially, the input features undergo layer normalization and a multi-head attention mechanism. This process aids in stabilizing the training process and allows the model to focus on different parts of the input features simultaneously. This capability enhances the model's ability to capture the global dependencies between pixels, significantly improving semantic segmentation performance. The epsilon value is set to 1e-6 to ensure computational stability, the key dimension is set to 128 to enhance the model's capacity to capture diverse features, and the number of heads is set to 4 for parallel computation of attention.

This module is appended to the end of the ASPP module because, while the ASPP module aids the model in **observing** the image at different scales, it enables the model to understand the connections between different parts of the image. It is akin to equipping the model with both a **telescope** and a **microscope**, allowing not only the observation of the overall shape of objects within the image but also the details of these objects. Thus, the model can perceive both the local information of each pixel and a broader range of global information.

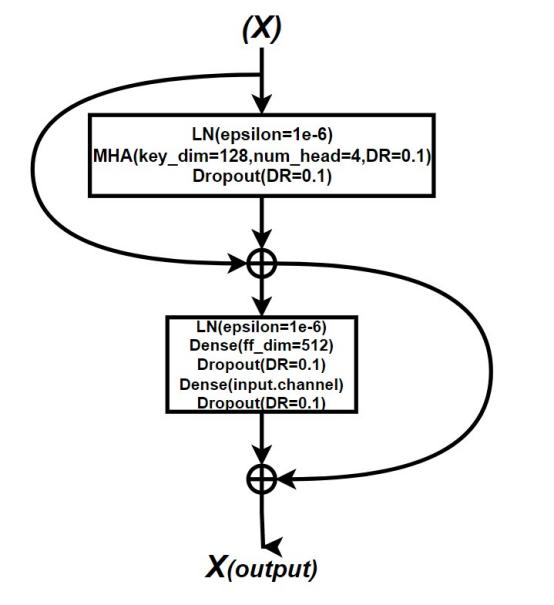


Figure 14. Transformer Module Flow Diagram

### **Edge Detection Module:**

Figure 9 depicts the edge detection module designed for this project. In this module, depthwise separable convolutions are integrated with the SE (Squeeze-and-Excitation) attention mechanism to achieve efficient edge detection. Initially, the module extracts features using depthwise convolutions (DWC). Subsequently, the results are fed into the SE module to perform an **excitation** operation on the input features, making the model focus more on the feature channels that are crucial for the edge detection task. Finally, the module employs a convolution with a kernel size of 1x1 to obtain the final output results. This approach ensures that the module efficiently emphasizes features relevant to edge detection, enhancing the overall performance of the task.

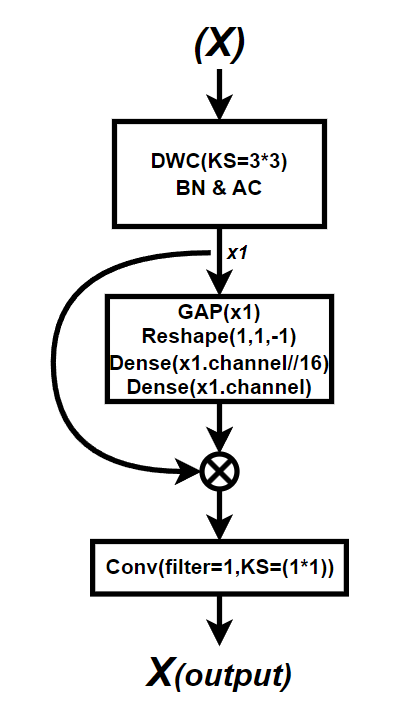


Figure 15. Edge Detection Module Flow Diagram

### **Context Enhancement Module:**

The following describes a context enhancement module flowchart, designed to enable the model to focus more on both local details and global context information within a scene. Initially, the input to this module is processed through three convolutional layers with varying dilation rates to capture features at different scales. These features are then combined through concatenation and weighted operations (Add and Multiply) to produce a comprehensive feature set. The weighting aims to utilize convolutional layers activated by a sigmoid function to determine the fusion weights for each feature. Such feature fusion allows the module to leverage various contextual information, enhancing segmentation accuracy.

Subsequently, a global average pooling operation is employed to extract the global information of the image. This information is then processed through two connected layers, activated by ReLU and Sigmoid functions, respectively, to enhance the global features. The global information generated by this module aids the model in better recognizing the overall scene layout and class distribution, thereby enhancing its semantic understanding of complex scenes.

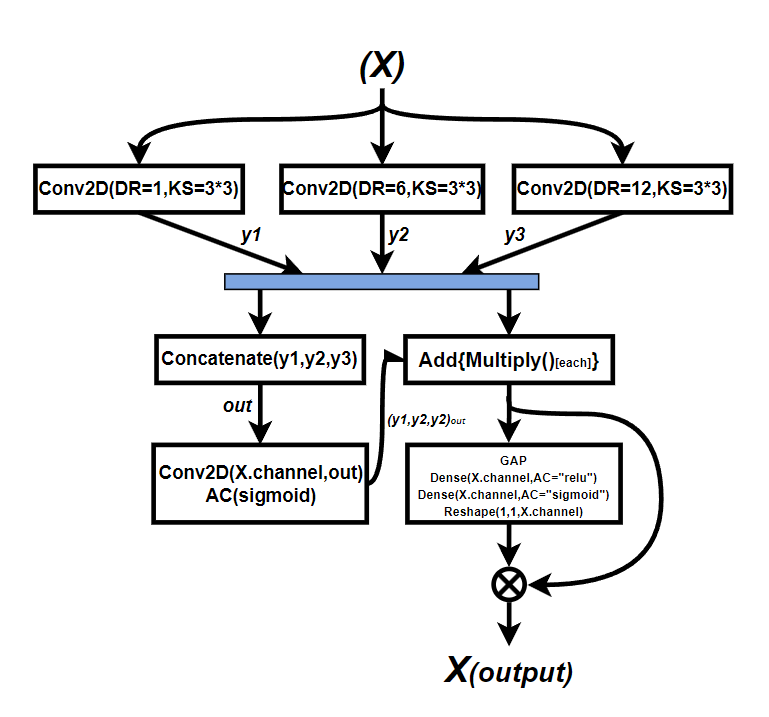


Figure 16. Context Enhancement Module Flow Diagram

### **ResNet50:**

ResNet50 is a residual network featuring 50 convolutional layers, designed to mitigate the training difficulties that arise with increasing network depth. The network includes residual modules that utilize concatenation of inputs and outputs to prevent the issue of vanishing gradients. The primary reason for selecting ResNet50 as the backbone network for the semantic segmentation model is to balance classification effectiveness with computational efficiency. This choice ensures that the model maintains high accuracy while also being relatively efficient to train and run, making it suitable for a wide range of semantic segmentation tasks.

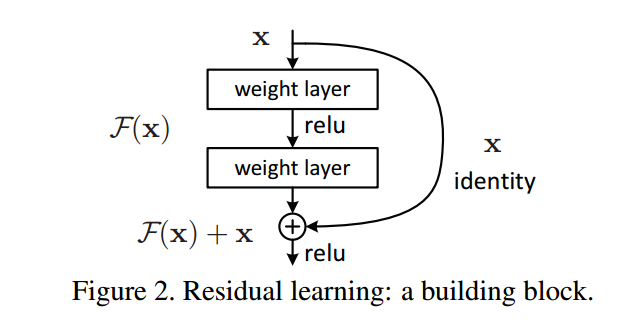


Figure 17. Residual module **[26]**

### **Overall Architecture of the Model:**

Figure 10 show the architecture of the model, with an input size of 384x384x3. Initially, the input data is directed into two separate paths: the edge detection module and ResNet50 equipped with pretrained weights. The former enhances the model's sensitivity to image edges, while the latter accelerates model training and improves generalization capabilities. Subsequently, weights from three layers of the pretrained ResNet50 model—namely, activation\_9, activation\_23, and activation\_39—are utilized.

The weights from the activation\_9 layer, representing low-level features, undergo convolutional layer processing and multi-scale pooling, followed by concatenation with the output of the convolutional layer. This step aids in capturing fundamental image information, beneficial for maintaining image details.

Activation\_23 represents a higher-level feature layer within ResNet50. Its weights are fed into the context enhancement module to produce output that assists the model in understanding more complex image content and contextual relationships, aiding in the recognition of complex objects.

Following this, the activation\_39 layer, which provides even higher-level abstract features compared to the activation\_23 layer, has its output directed into an ASPP module based on depthwise separable convolutions, and then the output is fed into a Transformer encoder. The ASPP module enhances the model's ability to process different parts of an image, while the Transformer encoder leverages attention mechanisms to improve segmentation accuracy.

Subsequently, the results from the edge detection module and the final outputs from activation\_9, activation\_23, and activation\_39 are combined. Finally, the combined result is passed through two layers of depthwise separable dilated convolutions based on a residual structure to obtain the final segmentation output. This significantly reduces the consumption of computational resources, ensuring that the model maintains high performance while keeping computational costs low.

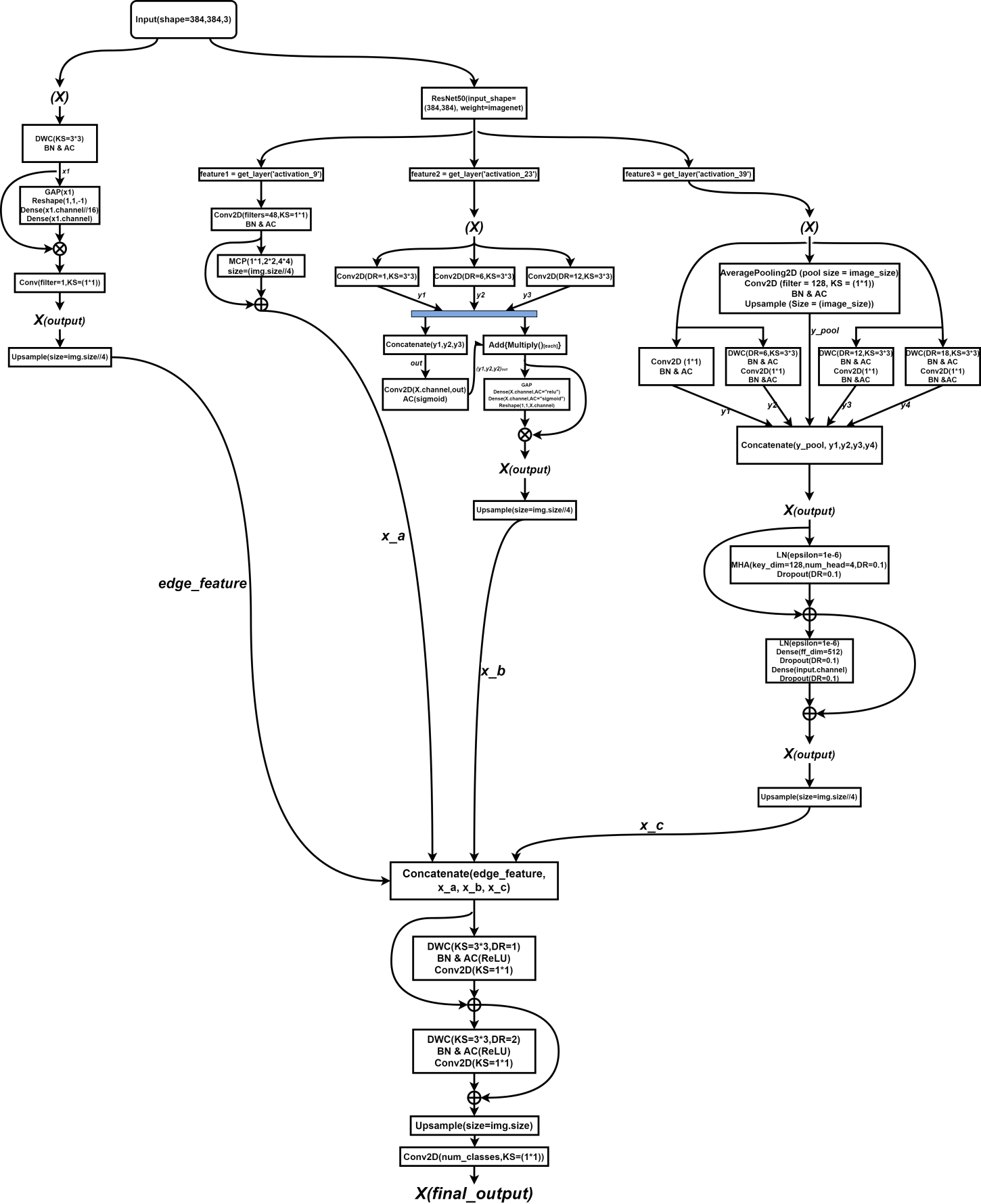


Figure 18. Model Architecture

## **Technology**

The part information is shown in Table 2.

|  |  |  |
| --- | --- | --- |
| Software | Framework | Tensorflow 2.10.0  Cudatoolkit 11.8.0  Cudnn 8.9.2.26 |
| Language | Python 2.9 |
| Libraries | Numpy  Matplotlib  Pandas  Keras 2.10.0  Glob3 |
| Version management plan | GitHub |
| Operation System | Windows 10 |
| Hardware | Central processing unit(CPU) | Intel Xeon Gold 6142 Processor 22M Cache 2.60 GHz (Cloud Server) |
| Graphic Processing Unit(GPU) | NVIDIA GeForce RTX 3090 (Cloud Server) |

Table 2 The technologies of the project

## **Testing and Evaluation plan**

### **Data testing**

* 500 test images from the Cityscapes dataset were selected to ensure coverage of all annotation categories.
* Verify the annotation completeness of each image in the test set to ensure that there is no missing data.
* Check that the test set contains images from different cities, different weather and different time periods to ensure comprehensiveness.
* Apply the same preprocessing steps to the test set as to the training set, including image resizing and normalization, to eliminate the effect of processing differences on the test results.

### **Model evaluation & testing**

**Primary indicators:**

* TP denotes the number of true cases (the number of samples that the model correctly predicts to be in the positive category)
* FN is the number of false negative cases (the number of samples that the model incorrectly predicts as negative)
* FP denotes the number of false positive cases (the number of samples that the model incorrectly predicts as positive)
* TN denotes the number of true negative cases (the number of samples correctly predicted by the model to be in the negative category).

**Secondary and tertiary indicators:**

### **Mean Intersection over Union(MIoU):**

MIoU [15] is mainly used to measure the degree of overlap between the segmentation results predicted by the model and the true results. In formula, 'k' represents the category number, 'Pii' represents number of overlapping pixels, and 'Pji' represents the number of misassigned pixels.' 1/(k+1)' is the average weight to ensure that each category contributes equally to the mIoU. Thus, mIoU is affected by the category number, the positive sample number, and the pixel overlap between different categories.

(equation 1)

1. **Accuracy:**

Accuracy is one of the most common metrics to measure the performance of a model it indicates the ratio of the number of samples correctly predicted by the model to the total number of samples. In semantic segmentation, Accuracy is equal to the total number of correctly classified pixels divided by the total number of pixels in the image and it gives a quick overview of how well the model performs on the entire dataset.

***Accuracy* =**  (equation 2)

1. **Loss Function:**

In this project, I used a combined loss function. This loss function is a combination of the Cross Entropy Loss Function, which takes into account the accuracy of the predicted probability distribution, and the Dice Loss, which focuses on the overlap of the shapes between the predicted and real labels. This combined loss function not only allows the model to learn accurate pixel categorization and overall region similarity but also alleviates the problem of category imbalance in semantic segmentation tasks.

Categorical Cross-Entropy Loss is used to measure the difference between the probability distribution predicted by the model and the probability distribution of the true label. Here, M is the number of categories,  is 1 if the true label of category c is observed and 0 otherwise, and  is the predicted probability that category c is observed.

(equation 3)

This following Dice Loss is applicable to multi classification task. Where Y is the binarized matrix of true labels, P is the predicted probability matrix, yi and pi are the true and predicted values respectively, and N is the total number of pixel points.

(equation 4)

Portfolio losses are realized through a weighted sum, where α and β are weighting parameters used to adjust the relative importance of the two loss terms.

(equation 5)

1. **Precision:**

Precision measures how many of all the samples classified as positive instances by the model are true instances. Precision is calculated using the following formula:

(equation 6)

Precision also ranges from 0 to 1, with closer to 1 indicating that the model is more accurate in the samples classified as true cases.

1. **Recall:**

Recall measures the ability of the model to correctly identify positive examples, also recall is known as True Positive Rate or Sensitivity.Recall is calculated as follows:

(equation 7)

The Recall Rate ranges from 0 to 1, the closer to 1 means the better performance of the model in identifying positive examples.

1. **Kappa**

Kappa is used to measure the consistency of two evaluators in a classification task beyond pure chance consistency. It is commonly used to evaluate the performance of machine learning models, especially in consistency tests for data labeling.

(equation 8)

Where is the observed consistency and is the chance consistency, it takes values between -1 and 1, with higher values indicating better consistency.

1. **PA**

Pixel accuracy is one of the most intuitive metrics for evaluating the performance of an image segmentation model, which calculates the percentage of all correctly categorized pixels out of the total pixels.

(equation 9)

In particular, TPi, TNi, FPi, and FNi represent the number of true cases, true negative cases, false positive cases, and false negative cases of the i th category, respectively, and n is the total number of categories.

1. **CPA**

The category accuracy calculates the proportion of pixels correctly categorized in each category out of the total pixels in that category, which is then averaged over all categories.

(equation 10)

For each category i, its CPA is given by the following equation, where TPi and FNi are defined as above.

1. **Dice Coefficient**

Dice coefficient is a statistical tool that measures the similarity of two samples and is commonly used in medical image segmentation. It calculates the ratio of the size of the intersection between twice the predicted and true labels to the sum of the respective sizes of the predicted and true labels.

(equation 11)

Here, Y is the set of real labels, P is the set of predicted labels, and yi and pi represent the value of each pixel point in the real and predicted labels, respectively.

1. **FPS**

FPS is a measure of the speed of image processing, especially important in video processing or real-time systems, and indicates how many frames per second the model can process.

(equation 12)

The average processing time is the average time it takes for the model to process a single frame of an image, and the unit is usually seconds.The higher the value of FPS, the faster the model can process.

# **Results**

## **Results of Model Training**

### **Final Result**

The following result is the final result of my model, which was run under the use of a combined loss function (Categorical Cross Entropy Loss + Dice Loss). Also, I have used SGD optimizer based on weight decayed as well as polynomial decayed learning rate scheduling to ensure the segmentation results. In the SGD optimizer, I used an initial learning rate of 0.01 and 1.2 as a power of polynomial decay, which ensures that the learning rate decreases slowly early in the training and accelerates later on, which is well suited for tasks that require long periods of time to explore the parameter space at a high learning rate. After that, I also set the weight decay value to 0.0001, which can be used for regularization to avoid overfitting. Finally, I set the momentum of the SGD optimizer to 0.9, which ensures that the model strikes a certain balance between speeding up training and avoiding excessive oscillations.

1. **Accuracy**

According to Figure, the model finally obtained an accuracy of 93.58. It can be clearly seen that the validation accuracy oscillates greatly in the early stage of training, and is relatively stable in the later stage. This is due to the fact that the learning rate is set too high in the early stage, and the weights of the model are too aggressive leading to too much change in each step, thus causing sharp fluctuations in the accuracy rate. In the later stage of training, the learning rate gradually stabilizes by decaying to a relatively small value. For the semantic segmentation of complex scenes and multi-category fine-grained segmentation, the accuracy rate reaches 93.58% indicating that the model can recognize and segment various objects in the image very well.

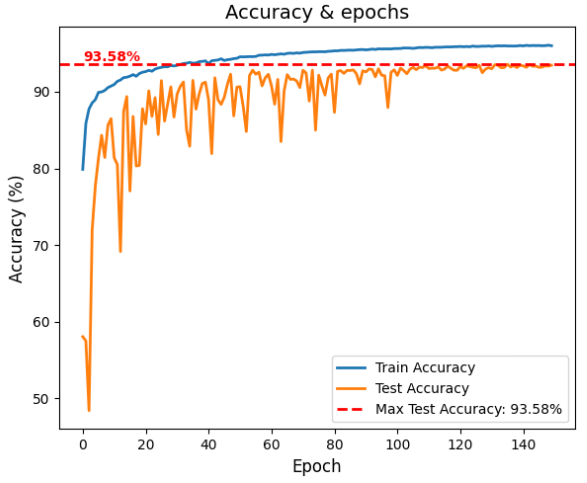


Figure 19. Accuracy

1. **Loss**

The change of loss is shown in Figure, which also has a strong vibration in the early stage and stabilizes in the later stage. This is also due to the high learning rate in the early stage of training, the update step of the model weights may be too large, resulting in the model "jumping" on the surface of the loss function, which causes sharp fluctuations in the loss. In addition, because I was in the 512 \* 1024 size of the image randomly cropped out of the 384 \* 384 size of the image, which makes it difficult for the model to learn all the features in the pre-training period, with the introduction of more data and the model's adaptation to the distribution of data, the size of the loss gradually stabilized. However, in the semantic segmentation task, the loss function is mainly used as a metric to optimize the model during training rather than an indicator to evaluate the model's performance, which can only help to determine whether the model is learning from the data and whether it is moving in the direction of reducing the prediction error.

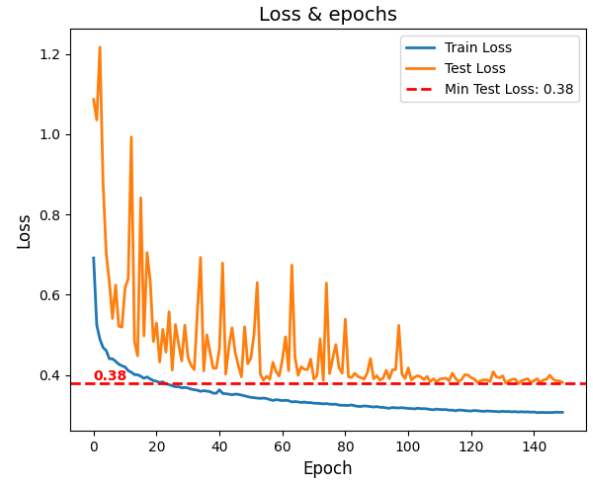


Figure 20. Loss

1. **MIOU**

The figure below shows the rising curve of MIOU during the training process. It can be clearly seen that the value of MIOU reaches 0.77, which is an okay result on the Cityscapes dataset, indicating that the model has a high generalization ability when dealing with complex urban street scenes. Based on the images, it can be seen that the training set consistently outperforms the test set, which is an expected situation since the model learns directly on the training set. However, the MIOU of the test set stabilizes at a high level, which means that the model has better convergence. In the pre-training period, the size of the MIOU is much smaller than that of the validation set, which is due to the fact that I randomly crop the original image, I crop a smaller region of the original image for training to get a larger field of view, which is good for the model to learn more details.

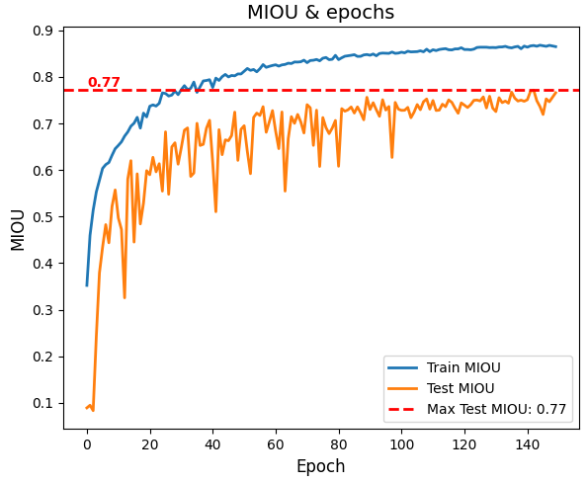
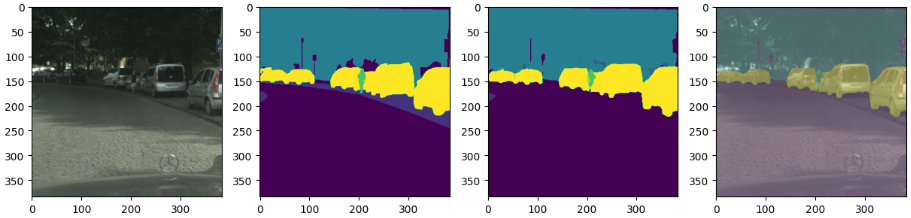
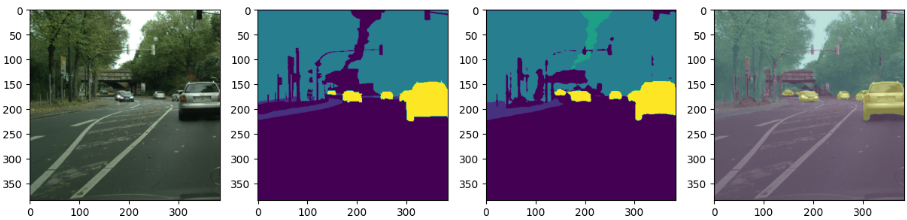


Figure 21. MIOU





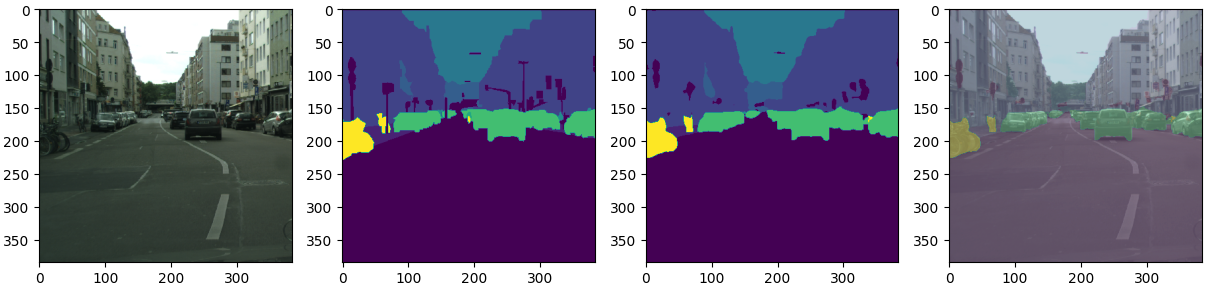


Figure. Segmentation Result

1. **Precision**

According to the figure below, the model achieved an accuracy of 0.95, which is a relatively high result, indicating that 95% of the samples predicted by the model to be in the positive category are indeed in the positive category, which shows that the model is very accurate and stable in determining pixel categories, indicating that the model has a high generalization ability. In addition, the training and testing accuracies are very close to each other, signifying that the model has a high generalization ability on both the training and testing sets. At about 80 epochs, the accuracy reaches a steady state and does not change much at subsequent times.

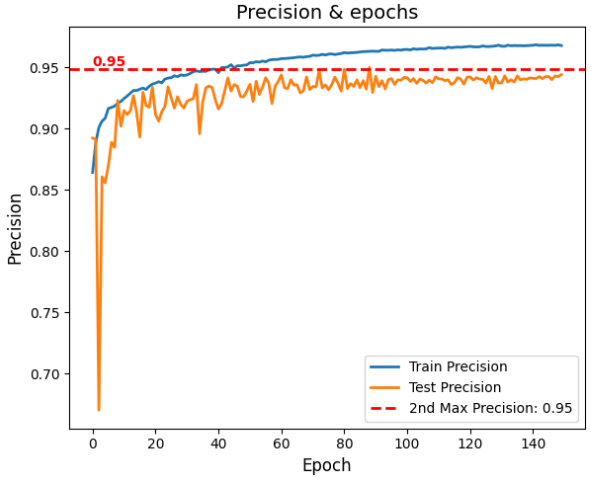


Figure 22. Precision

1. **Recall**

The recall of the model reached 0.97, which is a relatively high value, which means that the model was able to recognize the majority of positively classified samples in the dataset. In addition, the results of the training and validation sets are very similar and remain high, indicating that the model has good recognition ability on both datasets and there is no overfitting. Although the recall fluctuates more drastically in the early stages, it stabilizes at about 100 epochs, which is due to the fact that the model did not learn enough features in the early stages.

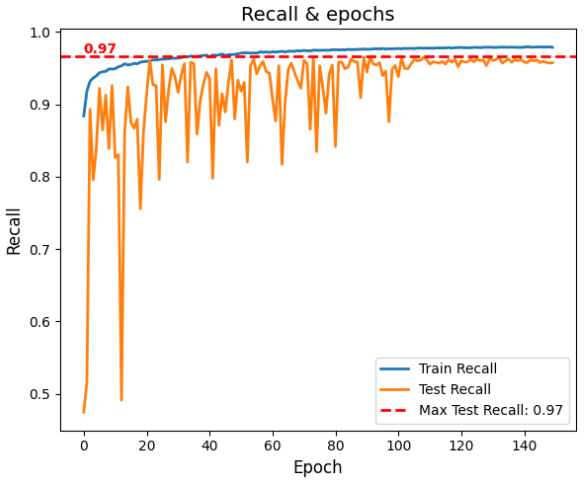


Figure 23. Recall

1. **Other Metrics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PA** | **CPA** | **Dice Coefficient** | **Kappa** | **FPS** |
| 0.9316 | 0.851 | 0.853 | 0.9357 | 65 |

**Table 3. Other Metrics**

According to the above table it can be seen that the pixel accuracy (PA) of the model reaches 0.9316, which indicates that the model correctly classifies 93.16% of the pixel points, which means that the model has a better performance on Cityscapes, which is a more complex dataset. Also, the average pixel accuracy reached 0.851, which indicates that the model has an average accuracy of 85.1% on each category and it shows the performance of the model on all categories.

The Dice coefficient reached 0.853, and although this metric is commonly used for medical image segmentation, I think it gives a good indication of the model's performance for similar cases like data imbalance. This metric is similar to IOU, which also measures the overlap between predicted segmentation and true segmentation, and 0.853 means that the model has good segmentation performance.

The kappa score reaches 0.9357 and this metric measures the classification accuracy while considering data imbalance. And the kappa score of 93.57% indicates that the model has a very reliable classification performance.

Finally, the model achieves an FPS of 65, indicating that the model can process 65 frames per second while processing the video stream. This is very important for real-time semantic segmentation, as it ensures that the model can respond very quickly to complex scene inputs while ensuring accuracy.

### **Training strategy adjustment and fine tuning**

1. **The training for fine-tuning has not been completed yet**

## **Comparison with Other Models**

### **Comparison of Common Models**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Loss** | **Precision** | **Recall** | **MIOU** | **PA** | **CPA** | **Dice Coefficient** | **Kappa** | **FPS** |
| LinkNet | 63.34 | 0.96 | 0.89 | 0.83 | 0.15 | 0.61 | 0.19 | 0.1978 | 0.41 | 53.1 |
| UNet | 63.75 | 0.95 | 0.9 | 0.86 | 0.15 | 0.59 | 0.20 | 0.2036 | 0.401 | 24.4 |
| My\_Model | 93.58 | 0.38 | 0.95 | 0.97 | 0.77 | 0.9316 | 0.851 | 0.853 | 0.9357 | 63 |

Table 3. Comparison of Common Models

In this study, I compare my model with the common semantic segmentation models LinkNet and UNet through an exhaustive comparative analysis. The results show that My\_Model significantly outperforms the comparison models in several key performance metrics. Specifically, my model achieves 93.58% in terms of accuracy, far exceeding LinkNet's 63.34% and UNet's 63.75%; in terms of the loss function value, my model is only 0.38, which indicates that the model's prediction is very close to the actual results, while the loss values of 0.96 and 0.95 for LinkNet and UNet, respectively, show a higher error rate. In addition, my model shows excellent performance in terms of precision, recall, MIOU, PA, CPA, Dice coefficient, and Kappa coefficient, especially in MIOU, My\_Model reaches 0.77, which is much higher than the 0.15 of LinkNet and UNet, and fully demonstrates its efficient ability in segmenting each pixel category of the image. In terms of processing speed, My\_Model runs at a rate of 65 frames per second (FPS), which not only ensures high accuracy, but also realizes efficient processing speed, which is especially important for application scenarios that require real-time processing. Overall, the performance of my model outperforms LinkNet and UNet in all comparative dimensions, validating the efficiency and sophistication of our model.

### **Comparison of Other backbones**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pre-trained** | **Accuracy** | **Loss** | **Precision** | **Recall** | **MIOU** | **PA** | **CPA** | **Dice Coefficient** | **Kappa** | **FPS** |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| My\_Model | 93.58 | 0.38 | 0.95 | 0.97 | 0.77 | 0.9316 | 0.851 | 0.853 | 0.9357 | 63 |

Table 4. Comparison of Other backbones models

### **Models in literature**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Author** | **Model** | **MIou** | **Accuracy** | **Recall** | **FPS** | **Para(M)** |
| Badrinarayanan et.al. **[20]** | SegNet | 56.1% | \* | \* | \* | 29.46 |
| Abdigapporov et.al.**[21]** | BiFPN | 56.4% | 89.6 | 79.8 | 65.7 | \* |
| Paszke et.al. **[22]** | ENet | 58.3% | \* | \* | 46.8 | 0.4 |
| Poudel et.al. **[23]** | Fast-SCNN | 68% | 83.5 | \* | 123.5 | 1.11 |
| Yu et al **[24]** | BiSeNet | 69% | 65.5 | \* | 65.5 | 14.1 |
| Fourure et al **[25]** | GridNet | 69.5% | \* | \* | \* | \* |
| Chen et al **[26]** | Deep-Lab CRF | 70.4% | \* | \* | \* | 15.2 |
| Lin et al **[27]** | RefineNet | 73.6% | 80.6 | \* | \* | \* |
| Li et.al.**[28]** | BiAttnNet | 74.7% | \* | \* | 89.2 | 2.2 |
| My Model | My\_Model | 75% | 93.58 | 0.97 | 65 | 10 |

Table 5.Comparison of Other models in literature

Through in-depth analysis and comparison, my model is compared with other well-known semantic segmentation models, demonstrating its outstanding performance and innovations in various key performance metrics.

First, in terms of MIoU, my model leads all compared models with a score of 75%, including the recent top performers BiAttnNet (74.7%) and RefineNet (73.6%). This result demonstrates the superior performance of my model in accurately segmenting the categories to which each pixel of an image belongs.

In terms of accuracy, my model achieves a high score of 93.58%, which is much better than BiFPN (89.6%) and RefineNet (80.6%), demonstrating its superior ability in correctly identifying image categories. Specifically, my model also achieves 0.97 in recall, showing that it is able to re-identify positive class samples almost perfectly, which is excellent among all the models listed.

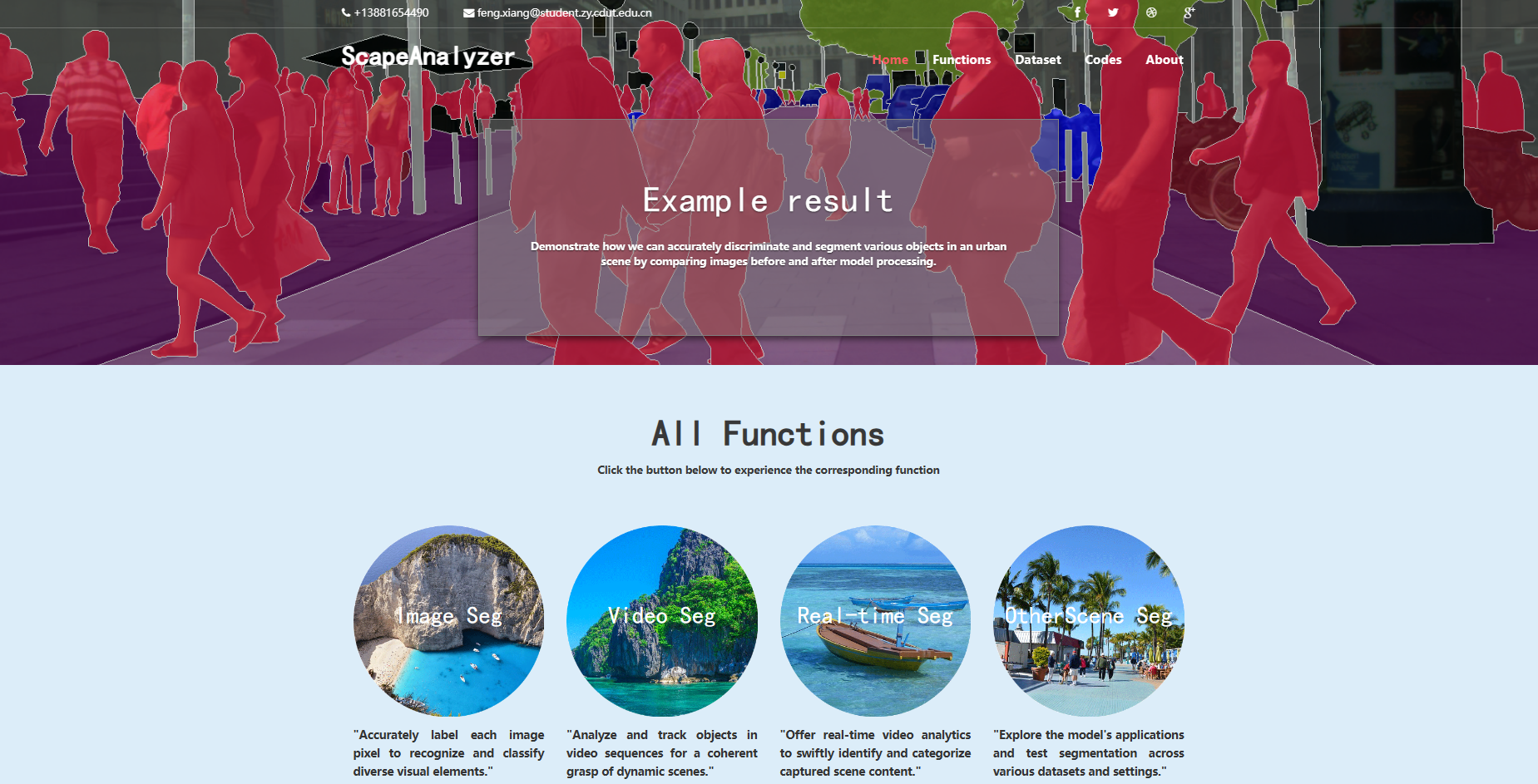
For FPS (frames per second), my model ensures good real-time processing with a performance of 65 FPS. Although slightly lower than Fast-SCNN's 123.5 FPS, this processing speed is a reasonable balance between accuracy and real-time performance, given my model's significant advantages in other performance metrics.

In terms of the number of model parameters (Para(M)), my model has a parameter count of 10M, which still achieves the best segmentation performance while maintaining a lower complexity compared to other models. This is in comparison to ENet (0.4M) and Fast-SCNN (1.11M), which have a much lower number of parameters, further demonstrating the results of my model in optimizing the model structure and improving efficiency.

In summary, my model performs well in the semantic segmentation task, not only leading in key metrics such as MIoU, accuracy and recall, but also achieving an excellent balance between processing speed and model efficiency. These results fully demonstrate the efficiency and sophistication of my model, providing new perspectives and solutions for future image processing and analysis tasks.

## **GUI Demonstration**

In this project, for the study of semantic segmentation of urban datasets, I designed and implemented an interactive website with the aim of improving user experience and demonstrating the practical application of semantic segmentation techniques. The website provides four main functions: image segmentation, video segmentation, real-time segmentation and other dataset segmentation. Users can upload static images of cityscape to the image segmentation module, and the system will automatically segment and label different city elements. In the video segmentation part, users can upload video data, and the website will analyze and segment the dynamic city scenes in the video frame by frame. The real-time segmentation function supports instant analysis of live video streams, which is suitable for real-time surveillance systems. Finally, the other dataset segmentation function provides users with the flexibility to perform semantic segmentation on non-urban datasets, which enhances the applicability and usefulness of the website. The entire website interface is designed to be intuitive and user-friendly, ensuring that even non-specialized users can easily get started and effectively utilize our semantic segmentation technology.



**Figure. Web GUI**

# **References**

1. Duarte, F. & Ratti, C., 2018. The Impact of Autonomous Vehicles on Cities: A Review. Journal of Urban Technology, 25(4), pp.3-18. Available at: https://doi.org/10.1080/10630732.2018.1493883
2. Janai, J. et al. (2020) ‘Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art’, Foundations and Trends® in Computer Graphics and Vision, 12(1–3), pp. 1–308. Available at: https://doi.org/10.1561/0600000079.
3. H. Hu, H. Cai, Z. Ma, and W. Wang, "Semantic segmentation based on semantic edge optimization," in Proc. 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS), Changchun, China, 2021, pp. 612-615, doi: 10.1109/EIECS53707.2021.9587939.
4. Hao, S., Zhou, Y. and Guo, Y. (2020) ‘A Brief Survey on Semantic Segmentation with Deep Learning’, Neurocomputing, 406, pp. 302–321. Available at: https://doi.org/10.1016/j.neucom.2019.11.118.
5. Li, J. et al. (2021) ‘Lane-DeepLab: Lane semantic segmentation in automatic driving scenarios for high-definition maps’, Neurocomputing, 465, pp. 15–25. Available at: https://doi.org/10.1016/j.neucom.2021.08.105.
6. Y. Guo and B. Yang, "A Survey of Semantic Segmentation Methods in Traffic Scenarios" in Proc. 2022 International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM), Xiamen, China, 2022, pp. 452-457, doi: 10.1109/MLCCIM55934.2022.00083.
7. K. Singh, V. Varshney, and M. Rajl, "Image Segmentation Role in Self Driving Car," in Proc. 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 861-864, doi: 10.1109/ICAC3N56670.2022.10074404.
8. Martínez-Díaz, M. and Soriguera, F. (2018) ‘Autonomous vehicles: theoretical and practical challenges’, XIII Conference on Transport Engineering, CIT2018, 33, pp. 275–282. Available at: https://doi.org/10.1016/j.trpro.2018.10.103.
9. V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481–2495, 2017, doi: 10.1109/TPAMI.2016.2644615.
10. S. Abdigapporov, S. Miraliev, V. Kakani, and H. Kim, “Joint Multiclass Object Detection and Semantic Segmentation for Autonomous Driving,” IEEE Access, vol. 11, pp. 37637–37649, 2023, doi: 10.1109/ACCESS.2023.3266284.
11. A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation,” CoRR, vol. abs/1606.02147, 2016. Available: http://arxiv.org/abs/1606.02147
12. R. P. K. Poudel, S. Liwicki, and R. Cipolla, “Fast-SCNN: Fast Semantic Segmentation Network,” CoRR, vol. abs/1902.04502, 2019. Available: http://arxiv.org/abs/1902.04502
13. C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation,” CoRR, vol. abs/1808.00897, 2018. Available: http://arxiv.org/abs/1808.00897
14. D. Fourure, R. Emonet, É. Fromont, D. Muselet, A. Trémeau, and C. Wolf, “Residual Conv-Deconv Grid Network for Semantic Segmentation,” CoRR, vol. abs/1707.07958, 2017, [Online]. Available: http://arxiv.org/abs/1707.07958
15. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834–848, 2018, doi: 10.1109/TPAMI.2017.2699184.
16. G. Li, L. Li, and J. Zhang, “BiAttnNet: Bilateral Attention for Improving Real-Time Semantic Segmentation,” IEEE Signal Processing Letters, vol. 29, pp. 46–50, 2022, doi: 10.1109/LSP.2021.3124186.
17. G. Lin, A. Milan, C. Shen, and I. D. Reid, “RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation,” CoRR, vol. abs/1611.06612, 2016, [Online]. Available: http://arxiv.org/abs/1611.06612
18. C. C. Aggarwal, Neural Networks and Deep Learning: A Textbook. Cham, Switzerland: Springer International Publishing AG, part of Springer Nature, 2018. doi: 10.1007/978-3-319-94463-0.
19. L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834-848, Apr. 1, 2018, doi: 10.1109/TPAMI.2017.2699184.
20. X. Wang, L. Yan, and Q. Zhang, "Research on the Application of Gradient Descent Algorithm in Machine Learning," in Proc. 2021 International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, 2021, pp. 11-15, doi: 10.1109/ICCNEA53019.2021.00014.
21. A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in Proc. ICML 2023. [Online]. Available: https://arxiv.org/pdf/2304.07288.pdf
22. S. Kato and K. Hotta, “Adaptive t-vMF dice loss: An effective expansion of dice loss for medical image segmentation,” Computers in Biology and Medicine, vol. 168, p. 107695, 2024, doi: https://doi.org/10.1016/j.compbiomed.2023.107695.
23. F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," Neurocomputing, vol. 338, pp. 321–348, 2019. Available: https://doi.org/10.1016/j.neucom.2019.02.003.
24. A. Chaurasia and E. Culurciello, "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation," CoRR, vol. abs/1707.03718, 2017. [Online]. Available: http://arxiv.org/abs/1707.03718
25. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," CoRR, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597
26. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," CoRR, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385
27. F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," in 2016 International Conference on Learning Representations (ICLR), 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1511.07122
28. C.-F. (Richard) Chen, Q. Fan, and R. Panda, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2021, pp. 357-366.