Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Research on Semantic Segmentation Method of Urban Street Scene Based on DeepLabv3+

by

Huazhi Lin

Supervisor: Mike

Project submitted as part of the requirements for the award of MSc Software Engineering and Artificial Intelligence

September 2023

APPENDIX 1: DECLARATION AND FORM OF CONSENT

DECLARATION

I, Huazhi Lin declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.

Signed : Huazhi Lin

Date : September 2023

FORM OF CONSENT

I, Hazhi Lin here by consent that my Project, submitted in candidature for the *M.Sc. Applied Computing*, if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.

Signed : Huazhi Lin

Date : September 2023

Where necessary replace *M.Sc. Applied Computing* with your named award.

Abstract

With the advancement of science and technology, image semantic segmentation techniques based on deep learning have found increasingly widespread applications in people's daily lives. In urban road scenarios, high-precision semantic segmentation technology plays a vital role in providing essential information for intelligent vehicle planning and decision-making, serving as a crucial safeguard for vehicle safety. Therefore, this paper conducts research on semantic segmentation methods in road scenes based on the DeepLabV3+ model. Starting from the model itself, it analyzes the segmentation challenges in urban road scene images and tests and improves the encoder-decoder structure of the DeepLabV3+ model, effectively enhancing its segmentation performance.

The specific work and innovations of this paper are as follows:

(1) Addressing the issue of the enormous parameter count in the main feature extraction network of the DeepLabV3+ model, which hinders practical applications, this paper explores the segmentation performance of the DeepLabV3+ model using commonly used convolutional neural networks as the feature extraction network. It investigates the performance of DeepLabV3+ with ResNet50, ResNeXt50, and MobileNetV2 as feature extraction networks. By adjusting the output stride to reduce feature loss during downsampling, the segmentation performance of the model is improved. When the output stride is set to 8x, the DeepLabV3+ model achieves the best segmentation performance. Considering practical model applications, experimental comparative analysis indicates that when the output stride is set to 16x, the DeepLabV3+ model achieves a balance between parameter count, complexity, and segmentation accuracy.

(2) To address the problem of the DeepLabV3+ model's poor performance on small-scale objects and fine details at object boundaries, this paper proposes improvements to the original decoder structure. The main contributions are as follows: firstly, by using global convolution modules and boundary refinement modules to extract local features from shallower network layers, the GCN-DeepLabV3+ and

GCNG-DeepLabV3+ models fuse more detailed information, resulting in better representation capability for image details in segmentation. Secondly, by applying convolutional block attention modules to map the local information extracted from shallower network layers in both channel and spatial dimensions, the CBAM-DeepLabV3+ model captures more critical local features through attention mechanisms, enhancing segmentation performance on small-scale target objects and boundary details. Experimental results demonstrate that the DeepLabV3+ model with the improved decoding structure in this paper is better suited for semantic segmentation tasks in urban road scenes.

CHAPTER 1: INTRODUCTION

1.1 Research Background and Significance

Urban street scene imagery captures detailed views of natural and man-made landscapes at the street level from a pedestrian's perspective. It offers extensive coverage, large volumes of data, efficient collection, and relatively low costs, making it an indispensable data source for urban environmental assessment [1]. Understanding urban road scenes is fundamental to various research applications, such as urban environmental perception [2-3], urban safety [4-5], urban land use [6], and urban walkability [7]. Semantic segmentation of road scenes, as a component of environmental perception technology [8], is a crucial element in autonomous driving systems [9]. High-precision semantic segmentation technology provides essential environmental, vehicle condition, and road condition information, ensuring the safe operation of autonomous vehicles [10]. Consequently, research on semantic segmentation techniques for road scenes holds significant practical importance and promising applications, making it a current focal point in various interdisciplinary fields.

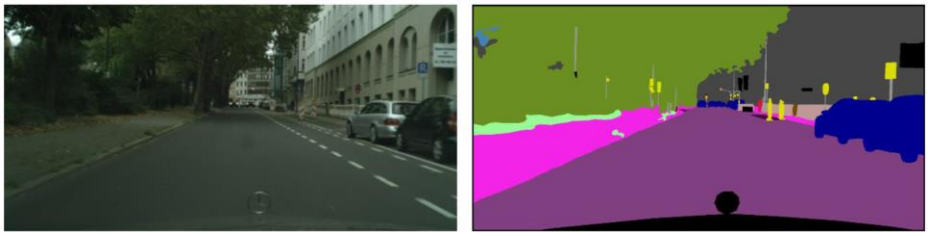


Fig 1.1 Semantic segmentation effect (cityscapes dataset image)

Semantic Segmentation of Road Scenes refers to the process of categorizing each pixel in captured street scene images based on its semantic information, essentially achieving pixel-level classification of the image [10]. The desired segmentation outcome is depicted in the right image in Figure 1.1, where the original image is interpreted to include multiple categories such as buildings, roads, grasslands, cars, etc., achieving pixel-wise classification of street scene images. Semantic segmentation of street scene images falls within the realm of computer vision research. Traditional semantic segmentation techniques rely on physical image features such as color, texture, grayscale, and others to perform image segmentation. Typical algorithms include

Normalized Cut [11], GrabCut [12], etc. However, due to their reliance on superficial image features, these algorithms often exhibit low segmentation accuracy, poor results, and limited real-world applications.

With the continuous development of computer hardware and deep learning technology, modern semantic segmentation techniques [13-17] based on deep learning have achieved breakthroughs and significant progress. Deep convolutional neural networks can extract abundant semantic information and local details from image annotations. Through end-to-end training, these models can learn richer global and local features, resulting in better generalization and more efficient and accurate image segmentation. However, the real-world application of semantic segmentation in street scenes is complex and dynamic. It requires consideration of both the similarities and differences between target objects, as well as the practical segmentation effects of objects of varying sizes. Factors such as occlusion, lighting conditions, temperature, and camera parameters in real scenes also affect the imaging of street scene images, increasing the difficulty of semantic segmentation in street scenes. These challenges merit further research.

Additionally, apart from focusing on algorithmic innovation, there is substantial research value and practical significance in flexibly applying existing deep learning models. Therefore, this paper focuses on the DeepLabv3+ semantic segmentation model and explores how to effectively apply it to street scene image data. It also addresses the limitations of the DeepLabv3+ model and proposes improvements to achieve better segmentation results in urban street scene contexts.

1.2 Project Aim and Objectives

With the continuous advancement of computer technology, there have been groundbreaking developments in semantic segmentation methods based on deep learning. Among them, the DeepLabV3+ model stands out as a semantic segmentation model with excellent segmentation performance. However, applying it to real-world urban street scenes requires consideration of several key factors:

(1) While the DeepLabV3+ model demonstrates strong segmentation performance,

its core feature extraction network has a substantial number of parameters, which can be impractical for model training and application. DeepLabV3+ adopts a typical encoder-decoder architecture, which is conducive to model refinement and replacement. Investigating the differences in models under different feature extraction networks holds significant insights for model selection, adjustment, and optimization in practical tasks.

(2) Despite the enhanced detail representation capabilities achieved through skip connections and concatenation fusion in the DeepLabV3+ model, it still exhibits limitations in capturing fine details of small-scale objects and object boundaries. High-precision semantic segmentation models are crucial for ensuring the safe operation of intelligent vehicles. Therefore, improving the DeepLabV3+ model's architecture to achieve superior segmentation results is of paramount research significance.

1.3 Report Structure

Chapter 1: INTRODUCTION

In this chapter, we provide an overview of the research background and project objectives. We conclude by briefly outlining the main tasks undertaken in this paper and the structure of the article.

Chapter 2: LITERATURE REVIEW

Chapter 2 categorizes deep learning-based semantic segmentation methods into three categories based on annotation and training approaches. It provides a detailed overview of the current state of research and developments both domestically and internationally.

Chapter 3: RESEARCH AND/OR DEVELOPMENT METHODOLOGY

Chapter 3 presents the design concepts and overall architecture of our research model, the DeepLabV3+ network. It introduces the datasets and evaluation metrics used to assess the improved algorithms and describes the experimental environment.

Chapter 4: DESIGN AND IMPLEMENTATION

Chapter 4 details various methods employed to enhance the DeepLabV3+ network and outlines the corresponding experiments conducted for implementation.

Chapter 5: RESULTS AND DISCUSSION / EVALUATION

Chapter 5 showcases the results obtained from the experiments conducted in Chapter 4. It includes an analysis of these results.

Chapter 6: CONCLUSIONS AND RECOMMENDATIONS

Chapter 6 summarizes the work presented in the preceding chapters, analyzes the limitations of the study, and provides insights into future research directions.

Chapter 7: REFLECTION

Chapter 7 offers a reflection on what has been learned during the course of this project and identifies areas for improvement in future work.

CHAPTER 2: LITERATURE REVIEW

2.1 Fully Supervised Semantic Segmentation Methods

Manually annotated image sample data contain rich surface information and local features, which greatly assist in training semantic segmentation models and improving segmentation accuracy. Therefore, fully supervised semantic segmentation methods are currently the most widely applied and effective segmentation techniques. The Fully Convolutional Network (FCN) [18], as a pioneering deep learning method in the field of semantic segmentation, replaced fully connected layers with convolutional layers after image feature extraction. This modification allowed the model to fully leverage the two-dimensional spatial information in images, achieving end-to-end, pixel-level dense predictions, marking a significant milestone in the entire semantic segmentation research domain.

Since then, researchers have continuously optimized, improved, and innovated semantic segmentation models and methods in various aspects, including optimizing convolutional structures. During the process of feature extraction from images, high-resolution images are advantageous for semantic segmentation. However, continuous downsampling and pooling operations can lead to a substantial reduction in image resolution, resulting in the loss of a significant amount of image information. Furthermore, this lost information is challenging to recover during the image upsampling process. Chen et al. [19] introduced the concept of dilated convolution in the DeepLabv1 model, innovatively enlarging the receptive field of convolutional kernels through dilation factors without increasing the number of parameters. This enabled the rapid and effective acquisition of deep image features while preserving the resolution of feature maps, thus improving segmentation accuracy. However, high dilation rates in dilated convolutions can lead to issues such as the loss of image details, feature discontinuities, and spatial gaps when extracting deep features. Wang et al. [20] addressed the "checkerboard effect" problem caused by dilated convolutions by mixing dilated convolutions and dense upsampling convolutions during the upsampling process. This further restored image details and enhanced the model's segmentation

accuracy.

To reduce model complexity, parameter count, and make deep learning models suitable for mobile devices, Paszke et al. [21] serially applied dilated convolutions to bottleneck modules, proposing the ENet model, which had fewer parameters while maintaining feature map resolution, thus advancing real-time semantic segmentation. Mehta et al. [22] constructed the ESPNet model by combining dilated convolution pyramids with point-wise convolutions, reducing model parameters while ensuring the receptive field and effectively improving the segmentation accuracy of real-time semantic segmentation models. In comparison to traditional convolutions mentioned above, Dai et al. [23] introduced the concept of deformable convolutions. These convolutions add a learnable offset during image computation, giving convolutional kernels the ability to adapt to spatial geometric deformations. While increasing the receptive field, deformable convolutions effectively accommodate geometric transformations in images, demonstrating their effectiveness in complex computer vision tasks.

Methods Based on Encoder-Decoder Structures. In semantic segmentation, the problem of reduced resolution and loss of spatial positional information due to convolution and pooling operations during image feature extraction is inevitable. The encoder-decoder structure is a commonly employed and effective approach to address this issue. Badrinarayanan et al. [24] drew inspiration from the feature extraction structure of the VGG16 network and designed a 5-layer symmetric encoder-decoder network called SegNet. In the encoder, local and global information of the image is extracted using convolutional layers, while pooling layers reduce model parameters through dimensionality reduction. In the decoder, the network incorporates position indices from max-pooling layers and utilizes reverse pooling operations to recover image details and spatial dimensions, mitigating information loss during downsampling and thereby enhancing segmentation accuracy. Noh et al. [25] proposed DeconvNet, a structure closely resembling SegNet but employing deconvolution operations in the decoder for pixel-wise classification predictions. UNet [26], named for its U-shaped architecture, is another classic encoder-decoder structure. UNet progressively restores

local information and resolution by concatenating and fusing feature maps corresponding to each encoder layer, achieving excellent segmentation results. Variants of UNet have significantly advanced the application of semantic segmentation in biomedical image segmentation.

The RefineNet network introduced by Lin et al. [27] adopts ideas from the residual identity mapping concept of backbone residual networks. It features residual convolution modules and multi-resolution fusion modules to effectively combine feature maps extracted at different layers. Additionally, it utilizes chained residual pooling modules to capture contextual information, resulting in high-resolution segmentation predictions. Peng et al. [28] introduced asymmetric convolution groups to achieve a large receptive field while saving parameters, innovatively enhancing segmentation accuracy. Their model design opened new avenues for semantic segmentation.Apart from these, real-time semantic segmentation approaches, such as ENet, LEDNet [29], and DFANet [30], also employ the typical encoder-decoder structure.

Methods Based on Feature Fusion Feature fusion involves the addition or concatenation of feature maps extracted from different layers during the downsampling process. Feature fusion not only captures hidden contextual information but also improves the utilization of image features, thereby enhancing segmentation effectiveness and accuracy. Lin et al. [31] introduced the concept of feature pyramids, which provide rich semantic information at various levels, benefiting computer vision tasks. Chen et al. [32] developed the DeepLabv2 model, drawing inspiration from the spatial pyramid concept and incorporating atrous convolutions introduced in DeepLabv1 to create the Atrous Spatial Pyramid Pooling (ASPP) module. This module effectively fuses multi-scale image features, enabling the segmentation of objects of various scales. In DeepLabv3 [33], improvements were made to the ASPP module's structure, and a DCNN network with both serial and parallel structures was designed to enhance segmentation accuracy.Liu et al. [34] extracted global information from images through global pooling. Recognizing that direct fusion of features at different scales can decrease model accuracy, ParseNet applied L2 regularization before fusing

extracted features with local information. This fusion improved segmentation results significantly. Zhao et al. [35] proposed PSPNet, which introduced more contextual information through pyramid pooling modules, effectively fusing features of different scales. This greatly enhanced scene parsing capabilities and reduced inter-class similarity errors and context mismatch issues. He et al. [36] addressed the balance between scale variations and dilation coefficients by introducing the DMNet model. They designed the Dynamic Convolutional Module (DCM) to obtain contextual information and achieved better segmentation results. These methods based on feature fusion contribute to the advancement of semantic segmentation by efficiently utilizing multi-scale information to enhance segmentation performance and accuracy.

Methods Based on Attention Mechanisms. Attention mechanisms involve focusing on crucial information for the target task while disregarding less relevant factors from a wealth of information. In the field of computer vision, attention mechanisms allocate different weights to different parts of an image, aiding neural networks in extracting more important and critical information. The PSANet network proposed by Zhao et al. [37] aggregates contextual information from various positions by introducing attention mechanisms. Although it consumes significant GPU memory, it introduces new perspectives into semantic segmentation. Li et al. [38] introduced the Pyramid Attention Network (PAN), which employs the Feature Pyramid Attention (FPA) module to add attention to higher-level features. By fusing contextual information from different scales, it enhances pixel classification consistency. Additionally, the Global Attention Upsample (GAU) module is employed to obtain high-level semantic information to restore lower-level local details, resulting in excellent segmentation performance.

Huang et al. [39] introduced CCNet, which utilizes a cross-cross attention mechanism. This model significantly reduces computational complexity and memory usage while improving efficiency and segmentation results. Fu et al. [40] proposed the Dual Attention Network (DANet) based on channel-wise and spatial-wise attention mechanisms to capture interdependencies in visual semantic features, yielding more precise segmentation results. Integrating attention mechanisms into semantic segmentation reduces dependency on external information, making it easier to uncover

internal feature correlations and improving segmentation accuracy.

Methods Based on Probabilistic Graph Models. Probabilistic graph models involve establishing dependencies among pixels as nodes in the late stages of convolutional neural network computation. These models balance the utilization of local and global image features, optimizing boundary information for segmentation targets and enhancing segmentation accuracy. Conditional Random Fields (CRFs) and Markov Random Fields (MRFs) are commonly used probabilistic graph models. Lin et al. [41] employed CRFs and CNNs for structured prediction in post-segmentation optimization. This approach reduces computational redundancy, incorporates rich global information, but may yield rough segmentation results due to the absence of Higher Order Potential (HOP) considerations. Arnab et al. [42] improved segmentation performance by training two types of HOP parameters embedded in CRFs using an end-to-end approach. Vemulapalli et al. [43] enhanced target segmentation results by introducing a fixed iteration Gaussian Conditional Random Field (GCRF). Liu et al. [44] integrated label context information and high-order features into MRF using a deep parsing network. They combined this with mean-field algorithms to improve segmentation efficiency and accuracy effectively.

Methods Based on Recurrent Neural Networks. Recurrent Neural Networks (RNNs) possess short-term memory capabilities, allowing them to utilize contextual information in images effectively and perform well in neural network models. Visin et al. [45] introduced ReSeg, a network that integrates CNNs and RNNs. Its decoding part is inspired by the ReNet network structure and effectively exploits image contextual information, resulting in improved segmentation. Li et al. [46] proposed LSTM-CF (Long Short-Term Memorized Context Fusion), which simultaneously integrates RGB and depth image contextual information for RGB-D image semantic segmentation. While LSTM-CF enhances segmentation accuracy, it often requires image chunking during processing, which can be cumbersome and may lead to underutilized image features. To address this, Liang et al. [47] proposed Graph-LSTM, which utilizes superpixels and their spatial connectivity in an adaptive undirected graph topology. This approach reduces computational redundancy, optimizes object boundaries, and

improves segmentation accuracy.

Methods Based on Generative Adversarial Networks (GANs). The semantic segmentation methods based on Generative Adversarial Networks (GANs) aim to enhance segmentation accuracy through adversarial training without increasing model algorithm complexity. Luc et al. [48] were the first to apply GAN concepts to semantic segmentation, treating traditional segmentation networks as generators and introducing a discriminative network for adversarial training. Through adversarial learning, the high-level inconsistencies between segmentation images and label images are gradually reduced, leading to improved segmentation accuracy. Xue et al. [49] combined GAN principles to design a Multi-scale L1 Loss network for optimizing network segmentation results. This effectively resolved the issue of inconsistent segmentation categories in the Unet network. While GANs can to some extent replace probabilistic graph models and pyramid models to enhance segmentation results, they exhibit scalability limitations on large datasets and are prone to instability. Consequently, GAN-based models are more frequently applied in semi-supervised learning scenarios with small sample datasets.

2.2 Weakly Supervised Semantic Segmentation Methods

While methods based on fully supervised learning offer better model training and yield superior segmentation results, the extensive need for labeled data also implies substantial human, material, and computational resources. Hence, researchers, considering practical applications and resource constraints, have shifted their focus towards using lightweight label data to train semantic segmentation models, aiming to enhance segmentation performance through weakly supervised learning methods. Currently, common weakly supervised learning semantic segmentation methods include boundary-level annotations, scribble-level annotations, image-level annotations, and various hybrid labeling approaches, as illustrated in Figure 2.1.

a. Border level annotation b. Graffiti level annotation c. Image level annotation

Fig 2.1 Common supervised learning method data examples [15]

Methods Based on Boundary-Level Annotations. Boundary-level annotations involve encompassing the entire object within a bounding box during sample annotation for model training. Dai et al. [50] proposed the BoxSup model, which trains on samples annotated with bounding boxes. The model iteratively optimizes the initial candidate regions obtained from the Multiscale Combinatorial Grouping (MCG) algorithm using Fully Convolutional Networks (FCN) until convergence, achieving segmentation accuracy comparable to fully supervised learning methods. Rajchl et al. [51] improved and extended the Grab-Cut algorithm, combining Conditional Random Fields (CRF) and CNN in a model that, after iterative training, yielded superior segmentation results. Khoreva et al. [52] recursively trained a CNN model using bounding box annotations generated by the GrabCut algorithm, achieving segmentation accuracy up to 95% of fully supervised learning methods. This demonstrates the effectiveness of weakly supervised learning segmentation methods based on boundary-level annotations.

Methods Based on Scribble-Level Annotations. Scribble-level annotations involve marking object regions in samples by drawing arbitrary lines, reducing the manual annotation workload. Lin et al. [53] proposed the ScribbleSup interactive semantic segmentation model, where annotators need to label only a small portion of the sample images. GraphCut algorithms automatically annotate the remaining samples, which are then input to the FCN model. Iterative training yields predicted segmentation results. Inspired by semi-supervised learning methods, Tang et al. [54] introduced the use of NormalizedCut regularization to compute the loss function, effectively improving model training quality and achieving better segmentation results. Obukhov et al. [55]

proposed a Gated Conditional Random Field Loss function suitable for weakly supervised learning semantic segmentation. This method effectively reduces the impact of irrelevant pixels in images, enhancing the training quality of scribble-level annotation data and further improving segmentation accuracy.

Methods Based on Image-Level Annotations. Image-level annotation methods construct connections between image category labels and pixel semantics to achieve semantic segmentation. These methods are straightforward to implement and widely applicable due to their simplicity. Pinheiro et al. [56] optimized segmentation models by allocating more weight to correctly classified pixel regions through the construction of aggregation layers. Durand et al. [57] introduced the WILDCAT model, capable of identifying multiple local regions, using a Multi-map Transfer Layer to obtain discriminative and prominent image features. Kolesnikov et al. [58] proposed the SEC (Seed, Expand, and Constrain) model composed of multiple loss functions. This model treats class activation map regions as target localization points, reasonably expands regions, and utilizes boundary loss functions to effectively enhance segmentation results. Wei et al. [59] introduced the STC (Simple To Complex) model, progressively enhancing network performance from simple to complex. The model initially identifies prominent regions in background images using saliency detection algorithms. These regions are then used as supervision information for the Enhanced-DCNN to obtain initial segmentation results. Finally, the results are input into the Powerful-DCNN, iteratively improving dense prediction results and enhancing segmentation accuracy.

Methods Based on Various Hybrid Annotations. In practice, weakly supervised learning methods often combine various types of weakly labeled data to train models due to the limitations of individual labeling techniques. Xu et al. [60] proposed a unified framework for handling weakly labeled data, using different constraint equations for different weakly labeled data to achieve pixel-level annotation effects, effectively improving segmentation results. Papandreou et al. [61] combined a small number of pixel-level labels with a large amount of weakly labeled data, using probabilistic graphical models and the Expectation-Maximization algorithm to predict the class of unknown pixels, achieving segmentation accuracy comparable to fully supervised

learning methods. Hong et al. [62] introduced the DecoupledNet model, which leverages mixed labeling data. The model independently performs image classification and segmentation tasks and enhances segmentation results through class activation maps.

CHAPTER 3: RESEARCH AND/OR DEVELOPMENT METHODOLOGY

3.1 DeepLabv3+ Network Model

DeepLabV3+ stands out as a leading semantic segmentation model known for its ability to effectively integrate multiscale information, as illustrated in Figure 3.1. Within the encoder-decoder architecture, it introduces the capability to flexibly control the encoder's feature extraction resolution, striking a balance between accuracy and computational efficiency through Atrous Convolution [68]. The DeepLabV3+ model is composed of two key components: the Encoder and the Decoder.

In the Encoder, the primary DCNN (Deep Convolutional Neural Network) employs a sequential Atrous Convolution approach to perform deep convolutional feature extraction on input images. The results are then divided into two segments. One segment serves as the low-level feature layer, directly feeding into the Decoder. The other segment undergoes parallel Atrous Convolutions, each utilizing Atrous Convolution with a different rate for feature extraction. This process captures context information at different scales. Subsequently, the features from different Atrous Convolution paths are merged, followed by 1×1 convolution to compress the features. The resulting high-level feature layer is then forwarded to the Decoder.

Within the Decoder component, the high-level feature layers from the Encoder undergo bilinear interpolation upsampling with a magnification factor of 4. These upsampled features are concatenated with the low-level feature layers, which have undergone channel reduction through 1×1 convolution. Further feature extraction is performed through two consecutive depth-wise separable convolution blocks. Ultimately, a 4× upsampling is applied to obtain the prediction map with the same dimensions as the original image.

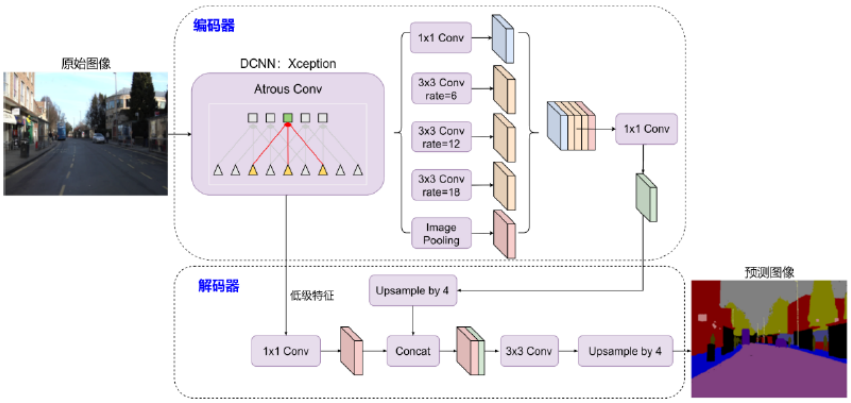


Fig 3.1 DeepLabv3+Model Structure

3.2 Improvements Implemented

This paper centers around the problem of semantic segmentation in urban road scenes and conducts experiments based on the DeepLabV3+ model. The specific research contributions are outlined as follows:

(1) Addressing the Issue of Excessive Parameters in the DeepLabV3+ Model for Feature Extraction: Recognizing the impracticality of the large parameter size in the DeepLabV3+ model's feature extraction network, this study explores the use of commonly employed convolutional neural networks (CNNs) as feature extractors. It investigates the segmentation performance of the DeepLabV3+ model with ResNet50, ResNeXt50, and MobileNetV2 as feature extraction networks. By adjusting the output stride to reduce feature loss during downsampling, the segmentation performance is improved. When the output stride is set to 8x, the DeepLabV3+ model exhibits the best segmentation results. However, considering practical model applications, experimental analysis reveals that when the output stride is adjusted to 16x, a balance is achieved between model parameters, complexity, and segmentation accuracy.

(2) Enhancements to the DeepLabV3+ Model Decoder for Small-Scale Objects and Object Boundary Details: Addressing the issue of the DeepLabV3+ model's subpar performance in handling small-scale objects and object boundary details, this paper introduces improvements to the original decoder structure. The key enhancements are as follows: a) Utilizing the Global Convolution Module and Boundary Refinement Module to extract local features from shallow network

layers. Models like GCN-DeepLabV3+ and GCNG-DeepLabV3+ with these features exhibit improved representational capabilities in capturing finer details for image segmentation. b) Employing Convolutional Block Attention Modules (CBAM) to map features from shallow network layers in both channel and spatial dimensions. The CBAM-DeepLabV3+ model, enhanced with attention mechanisms, extracts crucial local features, thereby enhancing segmentation performance for small-scale target objects and object boundary details. Experimental results demonstrate that the DeepLabV3+ model with these decoder structure improvements performs more effectively in semantic segmentation tasks involving urban road scenes.

3.3 Research Methodology

Following the algorithm improvements, the evaluation of the algorithm is conducted using the CamVid dataset. The CamVid dataset is publicly available from the University of Cambridge and is designed for road scene semantic segmentation using supervised learning. The dataset consists of over 700 manually annotated road scene images captured from the perspective of a vehicle. It is divided into training (367 images), validation (101 images), and test (233 images) sets, with image and label resolutions set at 480×360 pixels. The dataset comprises a total of 12 object classes, including sky, buildings, utility poles, roads, sidewalks, trees, traffic signs, fences, cars, pedestrians, bicycles, and background. This high-quality dataset serves as a fundamental experimental resource for studies related to object recognition, pedestrian detection, and other road scene research tasks.

Fig 3.2 Example of CamVid dataset images

In the experimental research, various performance metrics were employed to assess the effectiveness of the proposed methods. These performance metrics include Intersection over Union (IoU), Mean Intersection over Union (MioU), Precision, Recall, F1-Score, Pixel Accuracy (PA), Mean Absolute Error (MAE), among others. The mathematical expressions for these performance metrics are provided below:

$$IoU = \frac{TP}{(TP+FP+FN)} \tag{3-1}$$

$$MIoU = \frac{(IoU1+...+IoUn)}{n} \tag{3-2}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{3-3}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{3-4}$$

$$F1 - Score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{3-5}$$

$$PA = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{3-6}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3-7}$$

Among the terms used, TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives) represent the following: TP indicates cases where the model correctly predicts positives when they are indeed positive, TN denotes situations where the model correctly predicts negatives when they are indeed negative, FP signifies instances where the model erroneously predicts positives when they are actually negative, and FN represents occurrences where the model incorrectly predicts negatives when they are indeed positive. N represents the total number of pixels, y_i stands for the true class of the i-th pixel, and $\hat{y}_i$ represents the predicted class of the i-th pixel.

Among the various performance metrics, this paper selects MIoU (Mean Intersection over Union) and MAE (Mean Absolute Error) as the ultimate evaluation metrics for algorithm performance assessment. MIoU measures the relationship between the intersection and union of the predicted and actual regions, followed by averaging these values. MIoU comprehensively reflects the model's performance, taking into account the similarity and overlap between ground truth and predicted

values. MAE is a metric used to assess the accuracy of model predictions and represents the average difference between predicted values and ground truth values. In binary semantic segmentation tasks, MAE can be employed to evaluate the model's pixel-level classification accuracy, providing a more direct measure of the disparity between model predictions and ground truth. Since MIoU and MAE can simultaneously evaluate different aspects of binary semantic segmentation algorithm performance, this paper chooses them as the final evaluation metrics.

3.4 Experimental Design

The hardware environment used for experiments in this paper consists of an xeno(R)W-2145 CPU, an NVIDIA GeForce RTX 2080Ti GPU with 11GB of video memory. The software environment is based on Ubuntu 18.04, Python 3.6, PyTorch 1.10, and relevant Python libraries for neural networks.

CHAPTER 4: DESIGN AND IMPLEMENTATION

4.1 Problem Analysis

4.1.1 DeepLabV3+ Semantic Segmentation Model Based on Different Feature Extraction Networks

The DeepLabV3+ model has consistently been recognized as an outstanding performer in semantic segmentation tasks. However, it cannot be denied that despite its excellent segmentation performance, the model's overall parameter count remains substantial. This can pose challenges in practical applications. Specifically, this large model demands more computational resources for training and inference. Consequently, devices or scenarios with limited computing capabilities may face constraints when employing this model. Furthermore, large models require additional storage space, potentially increasing the complexity of model deployment and maintenance, especially in resource-constrained environments. This could also lead to longer model download and update times, which may not be ideal for applications requiring rapid deployment and updates.

To address this issue, one feasible approach is to consider replacing the backbone network with a smaller yet still performance-acceptable alternative. By doing so, it's possible to reduce the overall model's parameter count to some extent, thereby decreasing its computational and storage requirements. This makes it more suitable for deployment and operation in various application scenarios. With the continuous development of deep learning technologies, various convolutional neural networks have emerged, offering a variety of model choices while maintaining acceptable performance. DeepLabV3+ follows an encoder-decoder architecture, which is conducive to module replacement and structural improvement in practical applications.

4.1.2 Improved Decoding Structure of DeepLabV3+ Semantic Segmentation Model

In the field of autonomous driving, a high-precision semantic segmentation model is crucial because sufficient representational capacity is one of the prerequisites for a

model to be successfully applied in real-world scenarios. Urban road scenes are inherently complex and diverse, featuring objects of varying scales and being subject to multiple external factors such as lighting, temperature, and occlusions. Although the DeepLabV3+ model performs admirably in handling background-dominant, large-scale objects, its performance in small-scale objects and object boundaries is somewhat lacking.

To further enhance the DeepLabV3+ model's performance in segmenting complex urban road scenes, it is imperative to improve and optimize its decoder structure to achieve more refined segmentation results. The optimization goal is to enable the model to better capture information related to small-scale objects and object boundaries, thereby improving segmentation effectiveness. This entails enhancing the model's perceptual capabilities to adapt to scenes of varying scales and complexities.

Through this series of improvements, the DeepLabV3+ model achieves higher segmentation performance in urban road scenes, providing more reliable solutions for applications in autonomous driving and other related domains.

4.2 Residual Networks

4.2.1 ResNet50 Network

Residual Network (ResNet) [69], proposed by Microsoft Research, is a convolutional neural network architecture. Studies have shown that deeper network models can extract richer semantic features from images, which is advantageous for feature utilization in visual tasks. However, as networks become deeper, there is a challenge of gradient vanishing or exploding when stacking simple network layers. ResNet effectively addresses the issues of gradient vanishing and exploding by utilizing batch normalization layers, resolving problems related to slow convergence or non-convergence during network training.

In the practical process of model training, another issue known as the "degradation problem," distinct from overfitting, arises as network depth increases. To better train deep neural network models, ResNet introduces residual blocks through identity mapping, constructing deep convolutional neural networks. This design leads to

improved classification performance in image classification tasks. The overall structure of ResNet is similar, with the core part composed of stacked residual blocks.

Figure 4.1 illustrates the fundamental components of the ResNet50 network: (a) the bottleneck residual block that forms ResNet50, (b) detailed structural parameters of the feature extraction part of ResNet50, and (c) the change in feature map output stride after feature extraction in ResNet50, corresponding to the reduction factor after downsampling.

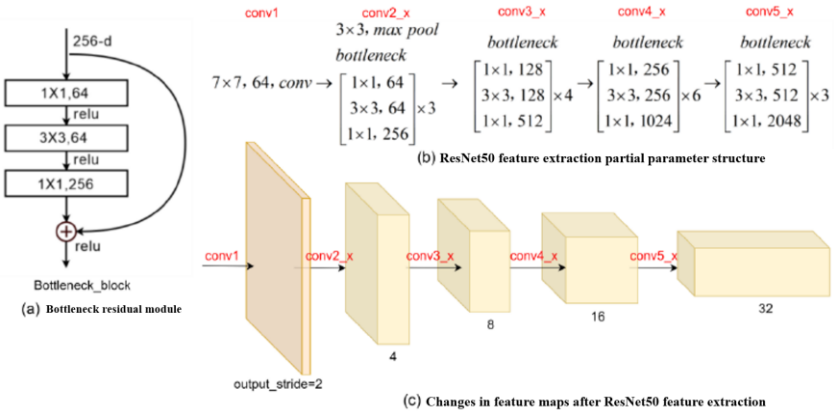


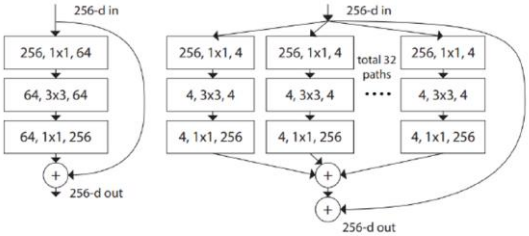
Fig 4.1 ResNet50 Detailed Structure

4.2.2 ResNeXt50 Network

ResNeXt [70] is a modified network that combines group convolution based on the ResNet architecture. It belongs to a variant of residual networks. Unlike Inception networks, which require manual design of intricate model structures, ResNeXt simplifies the design by using the same topology for each branch. This simplification aligns more with GPU design principles, making it more suitable for iterative model training.

ResNeXt structures the network by setting the number of groups through the "Cardinality" parameter. Under the same computational complexity, its building block units are fully equivalent to ResNet's residual modules. However, ResNeXt achieves superior classification performance in image classification tasks. Taking ResNeXt50 (32x4d) as an example, its basic residual block composition is illustrated in Figure 4.2(a). The overall structure of its feature extraction part is similar to ResNet50, but it employs 32 group convolutions to replace the bottleneck residual structures in layers

conv2_x to conv5_x, as detailed in Figure 4.2(b).



(a) **ResNeXt basic residual module**

conv1	$7\times7,\ 64,\ conv$	$\rightarrow$	$7\times7,\ 64,\ conv$
	$3\times3,\ max\ pool$	$\rightarrow$	$3\times3,\ max\ pool$
conv2_x	$\begin{bmatrix}1\times1,\ 64\\3\times3,\ 64\\1\times1,\ 256\end{bmatrix}\times3$	$\rightarrow$	$\begin{bmatrix}1\times1,\ 128\\3\times3,\ 128,\ C=32\\1\times1,\ 256\end{bmatrix}\times3$
conv3_x	$\begin{bmatrix}1\times1,\ 128\\3\times3,\ 128\\1\times1,\ 512\end{bmatrix}\times4$	$\rightarrow$	$\begin{bmatrix}1\times1,\ 256\\3\times3,\ 256,\ C=32\\1\times1,\ 512\end{bmatrix}\times4$
conv4_x	$\begin{bmatrix}1\times1,\ 256\\3\times3,\ 256\\1\times1,\ 1024\end{bmatrix}\times6$	$\rightarrow$	$\begin{bmatrix}1\times1,\ 512\\3\times3,\ 512,\ C=32\\1\times1,\ 1024\end{bmatrix}\times6$
conv5_x	$\begin{bmatrix}1\times1,\ 512\\3\times3,\ 512\\1\times1,\ 2048\end{bmatrix}\times3$	$\rightarrow$	$\begin{bmatrix}1\times1,\ 1024\\3\times3,\ 1024,\ C=32\\1\times1,\ 2048\end{bmatrix}\times3$

(b) **ResNeXt-50 (32x4d) feature extraction partial structural parameters**

Fig 4.2 Basic Structure of ResNeXt50

4.2.3 MobileNetV2 Network

MobileNetV2 [71] is a lightweight convolutional neural network developed by the Google team. This model strikes a fine balance between accuracy and real-time performance, showcasing innovations in two key structural aspects: Inverted Residuals and Linear Bottlenecks.

The Inverted Residuals structure, as illustrated in Figure 4.3, comes in two forms. The left diagram represents the specific configuration of the module when the stride is 1, and the input and output feature map tensors have the same shape. The right diagram illustrates the structure when the stride is 2. In contrast to the residual modules in ResNet, which consist of dimension reduction (compression), convolutional feature extraction, and dimension expansion (dilation), Inverted Residuals adopt an approach of dimension expansion (dilation), depth-wise separable convolutional feature extraction, and dimension reduction (compression). Additionally, the activation function in the residual structure is replaced with ReLU6. Inverted Residuals strike a balance between feature expansion and compression, effectively extracting valuable

data features.

The design of the Linear Bottlenecks structure aims to address the issue of low-dimensional feature information loss caused by ReLU functions. The specific architectural parameters of the MobileNetV2 network are outlined in Table 4.1, where "t" represents the expansion factor, "c" represents the output channel count, "n" represents the repetition count, and "s" represents the stride. Compared to the MobileNetV1 network, MobileNetV2 achieves higher accuracy, has a smaller model size, and finds broader applications in practice.

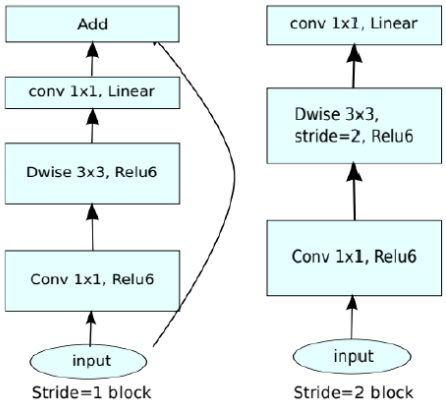


Fig 4.3 Inverted Residuals structure

Tab 4.1 MobileNetV2 Structure Parameter Table

Input	Operator	t	c	n	s
$224^2\times3$	conv2d	-	32	1	2
$112^2\times32$	bottleneck	1	16	2	1
$112^2\times32$	bottleneck	6	24	3	2
$56^2\times24$	bottleneck	6	32	4	2
$28^2\times32$	bottleneck	6	64	3	2
$14^2\times64$	bottleneck	6	96	3	1
$14^2\times64$	bottleneck	6	160	1	2
$7^2\times160$	bottleneck	6	320	1	1
$7^2\times160$	conv2d 1x1	-	1280	1	1
$7^2\times160$	avgpool 7x7	-	-		-
$1\times1\times1280$	conv2d 1x1	-	k		

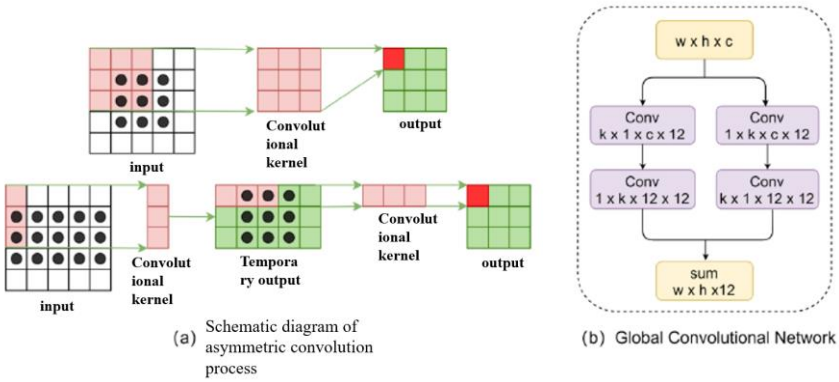
4.3 Improving decoding structure

4.3.1 Global Convolutional Network Module

The Global Convolutional Network module (GCN) [28], proposed by Tsinghua University and Megvii Technology, is a module designed to enhance model performance in semantic segmentation tasks. Its core principle involves the use of asymmetric convolutions to achieve a large receptive field, addressing pixel classification and localization challenges in semantic segmentation, ultimately leading to improved segmentation results. Compared to regular convolutions, asymmetric convolutions can produce feature mappings equivalent to those of large convolutional kernels while utilizing fewer parameters. As an example, consider a 3×3 regular convolution, whose feature mapping is equivalent to the convolution results of two asymmetric convolutions, 3×1 and 1×3, as illustrated in Figure 4.4(a).

The Global Convolutional Network module harnesses the power of asymmetric convolutions to extract richer image features with fewer parameters. Its specific structure, depicted in Figure 4.4(b), consists of two pairs of asymmetric convolutions: 1×k and k×1, and k×1 and 1×k. These pairs generate dense feature mappings, resulting in a k×k feature map, which is then fused and added to produce the final prediction image. Here, "k" represents the convolution kernel size, and "12" corresponds to the number of categories in the CamVid dataset's prediction images.

In semantic segmentation tasks, the Global Convolutional Network module effectively enhances model segmentation performance by leveraging a large receptive field. Its novel module design has introduced new perspectives into semantic segmentation research.



(a) Schematic diagram of asymmetric convolution process

(b) Global Convolutional Network

Fig 4.4 Schematic diagram of global convolution module

4.3.2 Boundary Refinement Module

The Boundary Refinement module (BR) [28] is an edge optimization module introduced within the Global Convolutional Network. Its structural design bears resemblance to residual structures and serves the purpose of further refining prediction results to enhance object localization and optimize boundaries. The specific architecture of the BR module is illustrated in Figure 4.5.

The BR module consists of two 3×3 convolutions connected via a residual short-circuit. It operates by adding the results of the residual structure to the original features, ultimately producing refined predictions with enhanced boundaries.

This module plays a crucial role in improving the localization and boundary accuracy of target objects. Its integration into the network contributes to more precise and optimized segmentation results.

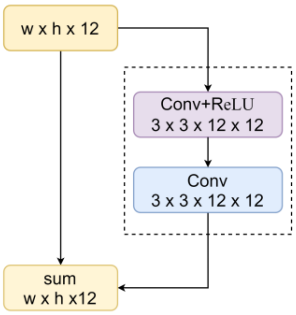


Fig 4.5 Boundary Refinement Module

4.3.3 Convolutional Block Attention Module

The Convolutional Block Attention Module (CBAM) [72], featured at ECCV 2018, stands as one of the notable representations of attention mechanisms. CBAM is a lightweight hybrid attention mechanism, requiring negligible parameters, and is easily integrated into CNN networks. It leverages attention mechanisms to direct the model's focus towards more crucial and essential information across channel and spatial dimensions, thereby enhancing model representational capacity and further improving performance.

CBAM's overall architecture, depicted in Figure 4.6, comprises two independent sub-modules: the Channel Attention Module (CAM) and the Spatial Attention Module

(SAM). Assuming an intermediate feature map, input features pass through the CAM and SAM modules sequentially to achieve channel and spatial attention mapping before producing the output feature map. The CAM structure, specifically illustrated in Figure 4.7(a), aggregates spatial information from feature maps using max-pooling and average-pooling operations. It then operates on the channel dimension, passing the feature map through a shared MLP network for channel information compression. Subsequently, it performs element-wise summation followed by sigmoid activation to yield the channel attention map. Similarly, the SAM module aggregates channel information using max-pooling and average-pooling and subsequently operates on the spatial dimension to generate the spatial attention map. Finally, it outputs the feature map after activation.

CBAM, with its dual attention mechanisms, plays a pivotal role in improving model performance by emphasizing essential information in both channel and spatial dimensions. Furthermore, it is a lightweight and parameter-efficient solution that can be seamlessly incorporated into CNN architectures.

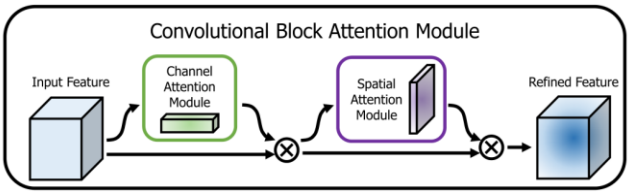


Fig 4.6 Convolutional Block Attention Module Structure [72]

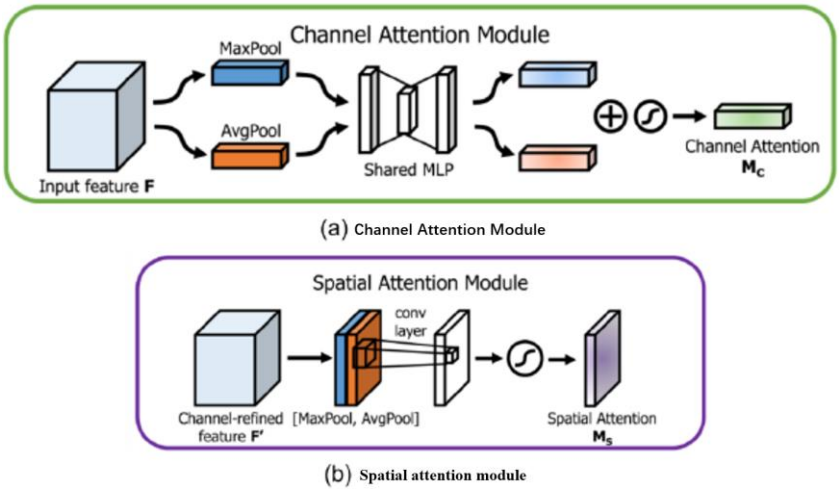


Fig 4.7 Convolutional Block Attention Module Submodule [80]

4.4 Semantic Segmentation Experiments Using DeepLabV3+ Models with Different Feature Extraction Networks

ResNet and MobileNet stand as quintessential representatives of traditional convolutional neural networks (CNNs) and lightweight CNNs in deep learning, widely applied in practical tasks. Building upon the experiments from Chapter 3, this subsection employs model transfer and logarithmic weighted loss functions to train models. The CamVid dataset is resized to 300×300, batched with a size of 4, and input to the DeepLabV3+ network for training. The initial learning rate for training is set to 0.01, decreasing exponentially to 0.001 using a poly decay strategy over 100 epochs. The experimental comparisons are made between Xception, ResNet50, ResNeXt50, MobileNetV2 networks, each serving as the feature extraction network for the DeepLabV3+ model. Importantly, ResNet50, ResNeXt50, and MobileNetV2 networks retain their original network structures, performing 32× downsampling to extract image features, and collaborating with the decoding portion of the original DeepLabV3+ model for segmentation prediction.

4.5 Exploration Based on the GCN-DeepLabV3+ Semantic Segmentation Model

Given practical computational costs and parameter considerations, ResNeXt50 is selected as the main feature extraction network with an output stride set at 16×. This choice is made as the baseline network for optimizing the decoding structure of the DeepLabV3+ model.

To further enhance the network's representational capabilities and address its minor shortcomings in capturing fine details of small-scale objects and object boundaries, two approaches are explored in this subsection. Firstly, the decoding structure of DeepLabV3+ is improved by incorporating the Global Convolutional Network (GCN) module (K=15). This module further extracts local features from shallow networks. It sequentially combines prediction segmentation images of 1/2 and 1/4 sizes from the shallow network with those obtained from the original network. These results are fused and upsampled to produce the segmentation prediction. This enhanced network is referred to as the GCN-DeepLabV3+ model, with its specific architecture depicted in

Figure 4.8. Secondly, a group module for local feature extraction, denoted as GCNG-DeepLabV3+, is constructed by parallelly using GCN modules of varying sizes (K=3, 7, 11, 15). It outputs segmentation prediction images for each module size and concatenates these local features with those from the original network. Finally, a 4× output stride is used to produce the prediction image. Both the GCN module and the final prediction image in this configuration are optimized using the Boundary Refinement (BR) module. The architecture of the enhanced network is illustrated in Figure 4.9.

The experimental settings for the training process are identical to those described in Section 4.4, involving resizing the CamVid dataset to 300×300, using a batch size of 4, and training the improved models. The initial learning rate is set to 0.01, decreasing exponentially to 0.001 using a poly decay strategy over 100 epochs. The experimental results and prediction images are summarized and presented in Section 5.

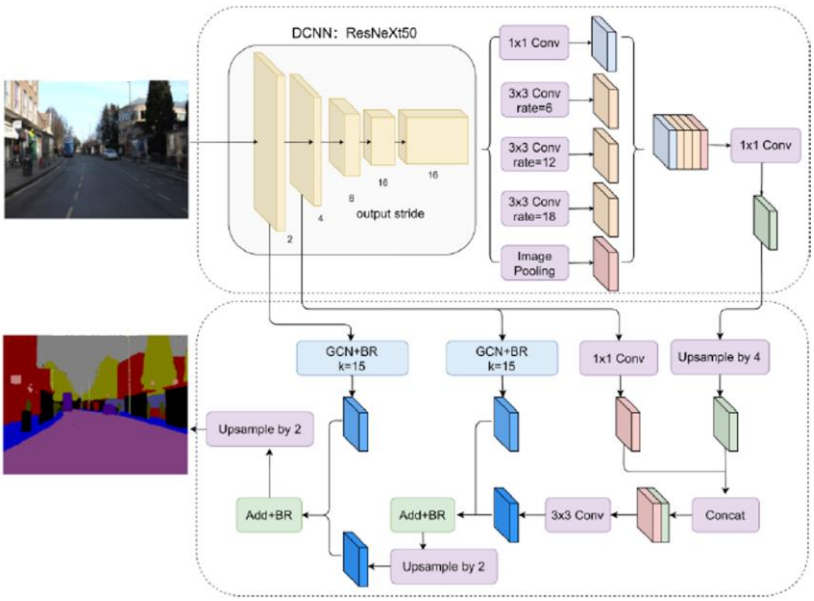


Fig 4.8 Structure diagram of GCN DeepLabV3+model

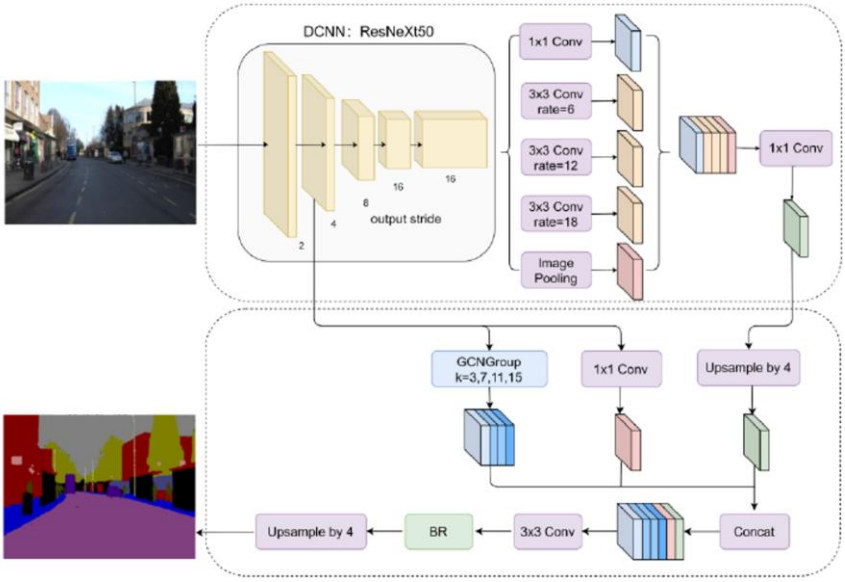


Fig 4.9 Structure diagram of GCNG-DeepLabV3+model

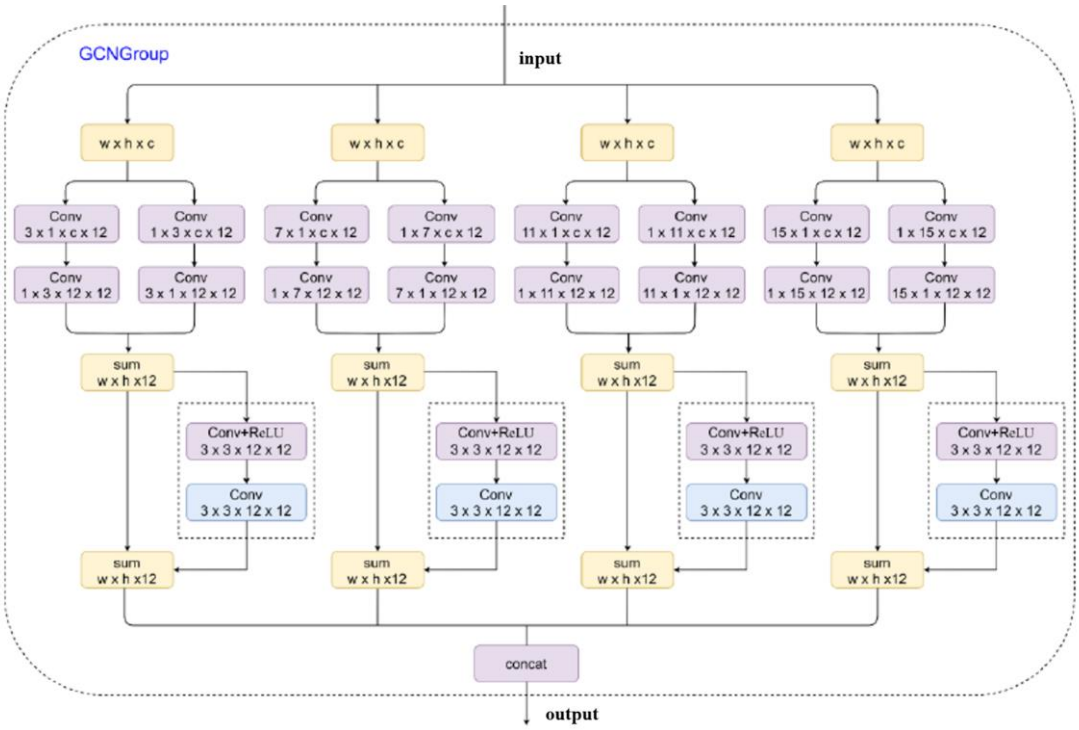


Fig 4.9 GCNGroup Structure Diagram

4.6 Exploration Based on the CBAM-DeepLabV3+ Semantic Segmentation Model

In order to further enhance the generalization capability of the DeepLabV3+ network and address its minor shortcomings in capturing fine details of small-scale objects and object boundaries, this subsection conducts experiments to improve the decoding structure of the DeepLabV3+ model by incorporating the Convolutional Block Attention Module (CBAM). This improvement leverages attention mechanisms

to extract more crucial and relevant local features, optimizing the model's segmentation performance.

Specifically, attention mechanisms are sequentially applied to the 1/4-sized feature vectors extracted by the decoder. Attention is applied separately to both channel and spatial dimensions. The feature maps obtained through attention mechanisms are then concatenated and fused with the prediction maps generated by the original network. Finally, segmentation prediction images are produced with a 4× output stride. This enhanced network is referred to as the CBAM-DeepLabV3+ model, and its specific architecture is illustrated in Figure 4.10.

The experimental settings for the training process are consistent with those described in Section 4.4, which involves resizing the CamVid dataset to 300×300, using a batch size of 4, and training the improved models. The initial learning rate is set to 0.01, decreasing exponentially to 0.001 using a poly decay strategy over 100 epochs. The experimental results and prediction images are summarized and presented in Section 5.

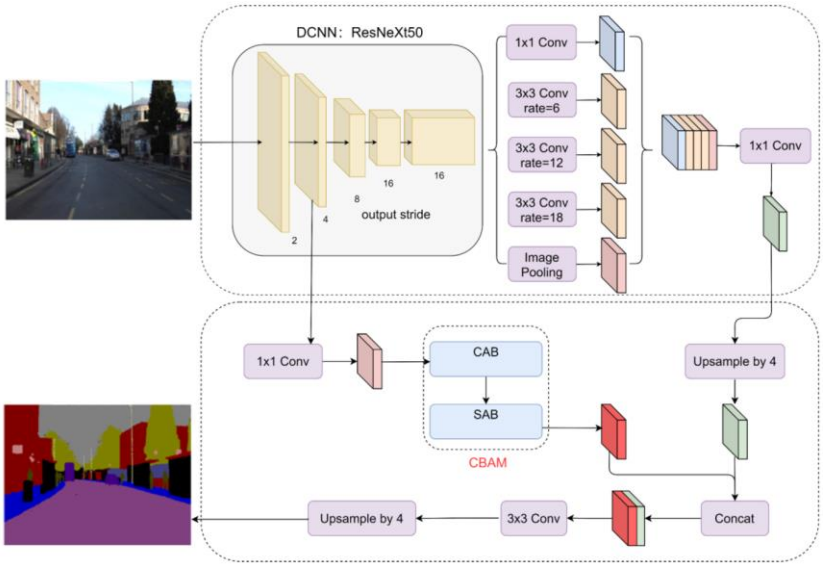


Fig 4.10 Structure diagram of CBAM-DeepLabV3+model

CHAPTER 5: RESULTS AND DISCUSSION / EVALUATION

5.1 Comparative Analysis of DeepLabV3+ Semantic Segmentation Model Results Based on Different Feature Extraction Networks

The loss function changes during the training process of different feature extraction networks are shown in Figure 5.1, and the accuracy changes on the validation set are shown in Figure 5.2.

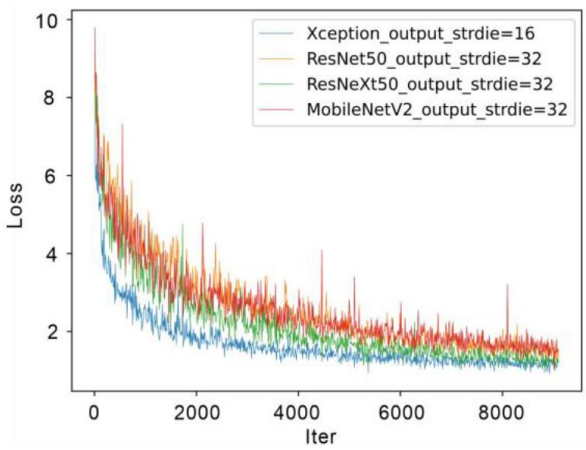
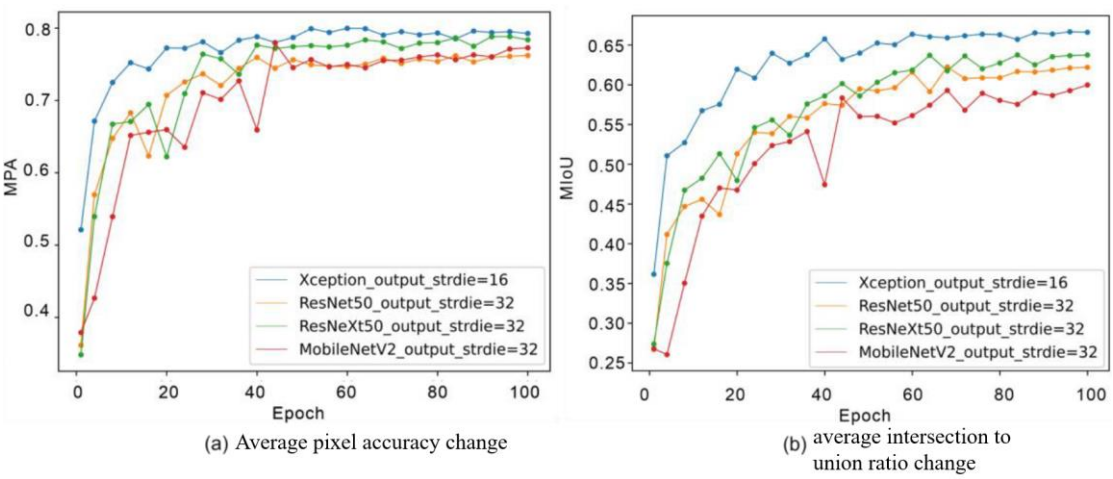


Figure 5.1 Model Training Loss Change Curve



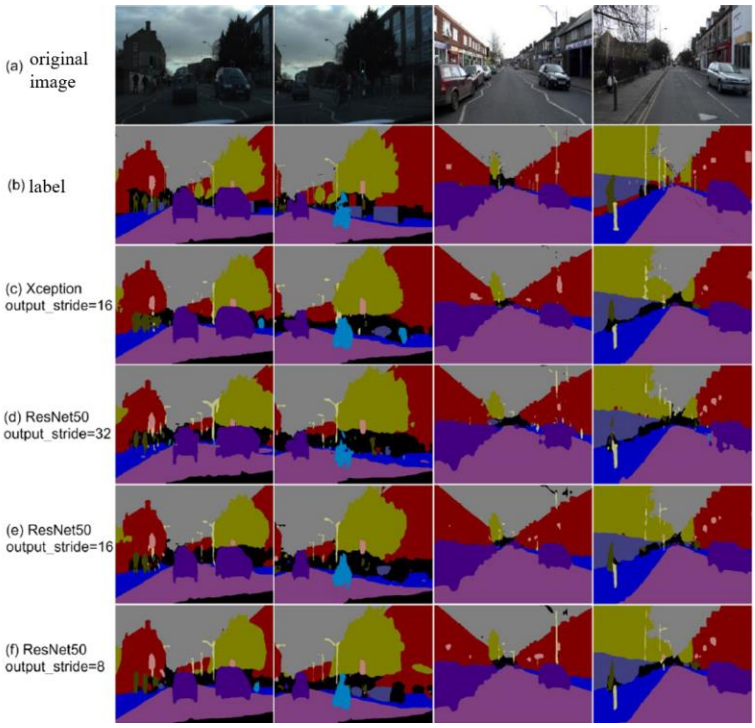
(a) Average pixel accuracy change

(b) average intersection to union ratio change

Figure 5.2 Precision variation curve of model training validation set

Tab 5.1 Overall Precision Statistics of DeepLabv3+under Different Feature Extraction Networks

Backbone	Output_stride	GFLOPS	Parameters/M	MPA/%	MIoU/%
Xception	16	40.99	80.21	**79.25**	**66.6**
	32	22.47	67.63	76.75	62.66
ResNet50	16	36.99	67.63	81.28	69.06
	8	105.34	67.63	**82.78**	**70.51**
	32	22.74	67.1	78.23	63.97
ResNeXt50	16	37.01	67.1	81.58	68.78
	8	104.69	67.1	**83.62**	**70.47**
	32	8.37	8.44	77.5	60.08
MobileNetV2	16	10.06	8.44	79.04	63.82
	8	17.65	8.44	**80.73**	**65.95**



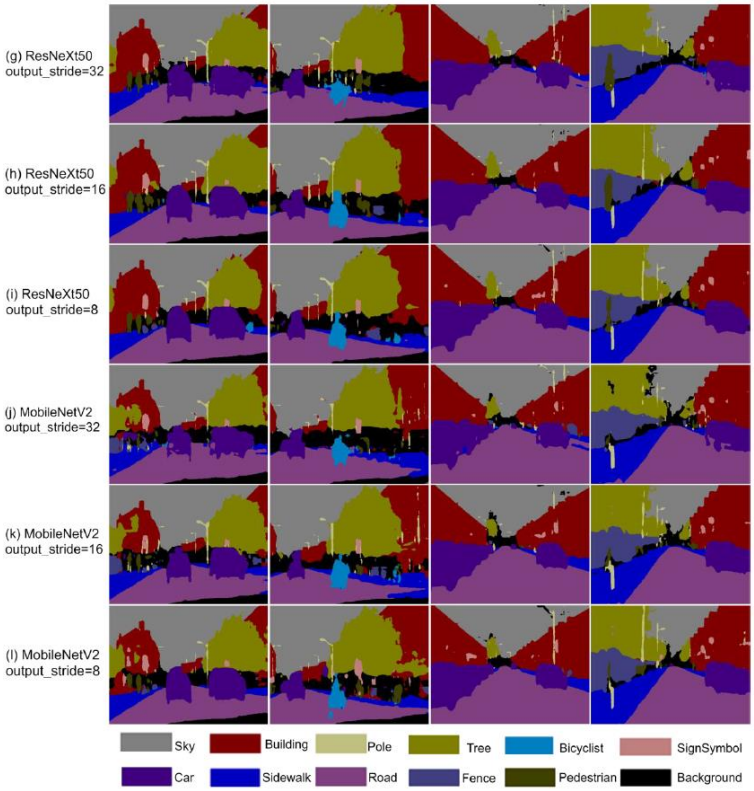


Fig 5.3 Comparison of Model Prediction Images

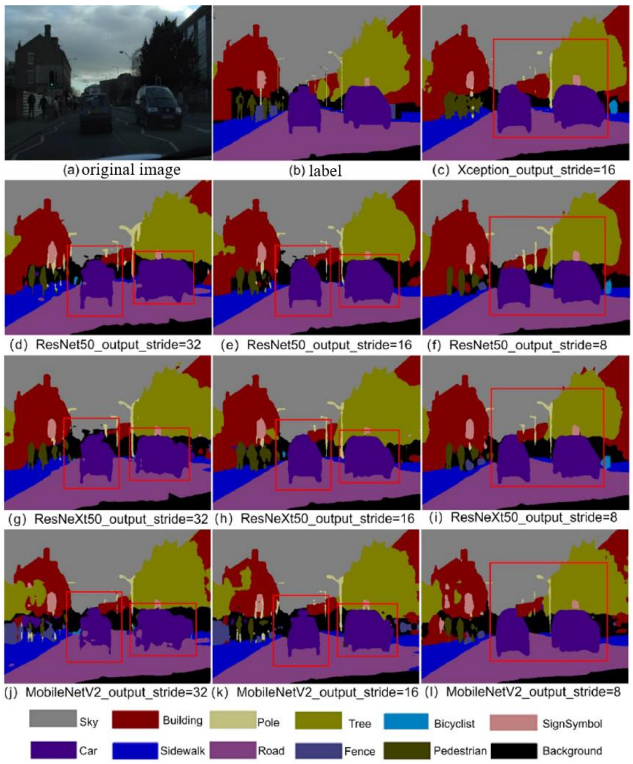


Fig 5.4 Comparison of Model Prediction Image Details

(1) Without altering the initial structure of the feature extraction networks and with identical model training settings, the behavior of different feature extraction networks

on the CamVid validation dataset can be observed from the loss values during training (Figure 5.1) and the changes in accuracy (Figure 5.2). Xception performs the best among the feature extraction networks, exhibiting faster and lower loss convergence, as well as higher Mean Pixel Accuracy (MPA) and Mean Intersection over Union (MIoU) values. ResNeXt50 generally outperforms ResNet50, while MobileNetV2 displays loss convergence patterns similar to ResNet50 but with more outliers and greater oscillations in accuracy, along with slower model convergence.

(2) Combining the overall accuracy statistics on the CamVid validation dataset for different output strides of common feature extraction networks (Table 5.1), it is evident that reducing the downsampling layer stride in the feature extraction network can effectively enhance the segmentation performance of the DeepLabV3+ model. When the model's predicted images have an output stride of 16x, ResNet50, ResNeXt50, and MobileNetV2 achieve the best MIoU values on the CamVid validation dataset, namely 69.06%, 68.78%, and 63.87%, respectively, indicating noticeable improvements in accuracy. However, it should be noted that adjusting the output stride, although not increasing the model's training parameters, consumes more computational resources. When the model's predicted images have an output stride of 8x, computational resource usage increases significantly, training time lengthens, and complexity rises, but the gain in segmentation accuracy is relatively minimal. Overall, an output stride of 16x appears to strike a balance between parameters, complexity, and segmentation accuracy.

(3) Examining the predicted images with different output strides from various feature extraction networks (Figure 5.3) and the detailed comparison images (Figure 5.4), it is apparent that when the output stride is set to 32x, ResNet50 and ResNeXt50 networks yield relatively poor prediction results, exhibiting inconsistencies between classes and blurred object boundaries. However, when the output stride is reduced to 16x, ResNet50 and ResNeXt50 networks significantly improve their prediction quality, resulting in better segmentation performance. Furthermore, when the output stride is 8x, the predicted images show the best segmentation performance for foreground classes, with clearer object boundaries and fewer misclassifications or omissions.

(4) These experiments highlight that the original DeepLabV3+ model is

characterized by a large number of parameters and complex architecture. Due to dataset limitations, Xception as the backbone network struggles to train a well-generalized model on small sample datasets. These experiments demonstrate that choosing different convolutional neural networks for feature extraction based on practical requirements is a viable approach. While MobileNetV2 as a backbone network may not match the segmentation performance of traditional CNNs, it offers the advantages of fewer model parameters and lower complexity, making it a favorable choice for practical tasks.

5.2 Comparative Analysis of Results for GCN/CBAM-DeepLabV3+ Semantic Segmentation Model Exploration

This section compares the segmentation results between the improved CBAM-DeepLabV3+ model and the baseline network. The segmentation accuracy for various categories in the improved DeepLabV3+ model is presented in Tables 5.2 and 5.3, while Table 5.4 provides overall segmentation accuracy statistics. The second and third rows in the tables present the complexity and parameter count results for testing the improved models under the same criteria, with testing performed on images of size 300x300 pixels. For a more intuitive comparison of segmentation results for the improved decoding structure of the DeepLabV3+ model, partial predicted images on the CamVid dataset are displayed in Figure 5.4. Specifically, (a) shows the original test images, (b) displays the ground truth labels, (c) represents the predictions from the baseline network, and (d)-(f) illustrate the predictions from the GCN-DeepLabV3+, GCNG-DeepLabV3+, and CBAM-DeepLabV3+ models, respectively. Figure 5.5 provides a detailed comparison of the second column of predicted images from Figure 5.4.

Tab 5.2 DeepLabV3+MPA values for each category of improved decoder

Category	Baseline	GCN-DeepLabV3+	GCNG-DeepLabV3+	CBAM-DeepLabV3+
Sky	95.61	**96.3**	95.08	95.54
Building	86.38	86.57	87.49	**88.66**
Pole	16.35	18.13	**25.15**	15.14
Road	97.77	97.59	**98.15**	98.08
Sidewalk	96.12	96.93	**97.12**	96.04
Tree	96.25	**96.9**	96.65	96.83
SignSymbol	**76.85**	76.3	76.26	73.9
Fence	93.76	91.5	**94.03**	92.34
Car	89.07	89.8	90.1	**90.35**
Pedestrian	81.85	84.57	**86.14**	81.1
Bicyclist	89.79	**95.06**	92.16	93.21
Background	58.61	58.77	**58.97**	57.56

Tab 5.3 DeepLabV3+MIoU values for each category of improved decoder

Category	Baseline	GCN-DeepLabV3+	GCNG-DeepLabV3+	CBAM-DeepLabV3+
Sky	93.51	**93.67**	93.09	93.29
Building	84.62	84.96	85.52	**86.55**
Pole	12.65	15.53	**18.91**	12.95
Road	96.57	96.75	**97.15**	96.8
Sidewalk	85.08	86.63	**87.4**	87.22
Tree	90.89	90.53	**91.36**	90.97
SignSymbol	58.97	**61.97**	60.31	61.42
Fence	**65.17**	69.4	68.21	69.25
Car	81.85	83.18	**84.67**	82.36
Pedestrian	49.94	51.24	**55.58**	55.47
Bicyclist	77.04	77.95	**80.06**	79.91
Background	29.11	28.55	**30.48**	29.93

Tab 5.3 DeepLabV3+MIoU values for each category of improved decoder

Category	Baseline	GCN-DeepLabV3+	GCNG-DeepLabV3+	CBAM-DeepLabV3+
Sky	93.51	**93.67**	93.09	93.29
Building	84.62	84.96	85.52	**86.55**
Pole	12.65	15.53	**18.91**	12.95
Road	96.57	96.75	**97.15**	96.8
Sidewalk	85.08	86.63	**87.4**	87.22
Tree	90.89	90.53	**91.36**	90.97
SignSymbol	58.97	**61.97**	60.31	61.42
Fence	**65.17**	69.4	68.21	69.25
Car	81.85	83.18	**84.67**	82.36
Pedestrian	49.94	51.24	**55.58**	55.47
Bicyclist	77.04	77.95	**80.06**	79.91
Background	29.11	28.55	**30.48**	29.93

Tab 5.4 Overall segmentation accuracy statistics of DeepLabV3 with improved decoder

Category	Baseline	GCN-DeepLabV3+	GCNG-DeepLabV3+	CBAM-DeepLabV3+
GFLOPS	37.01	38.31	39.01	37.01
Parameters/M	67.1	69.78	69.78	67.12
MPA	81.58	82.37	**83.11**	81.56
MIoU	68.78	70.03	**71.06**	70.51

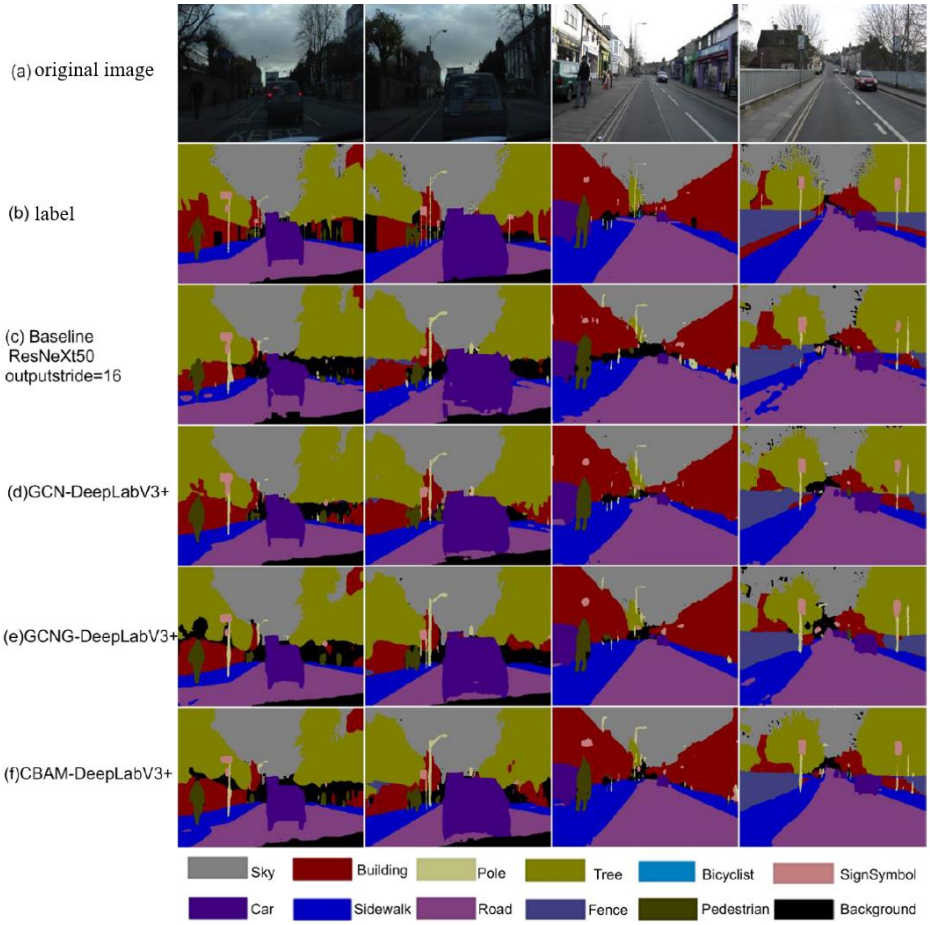


Fig 5.4 Comparison of Model Prediction Images

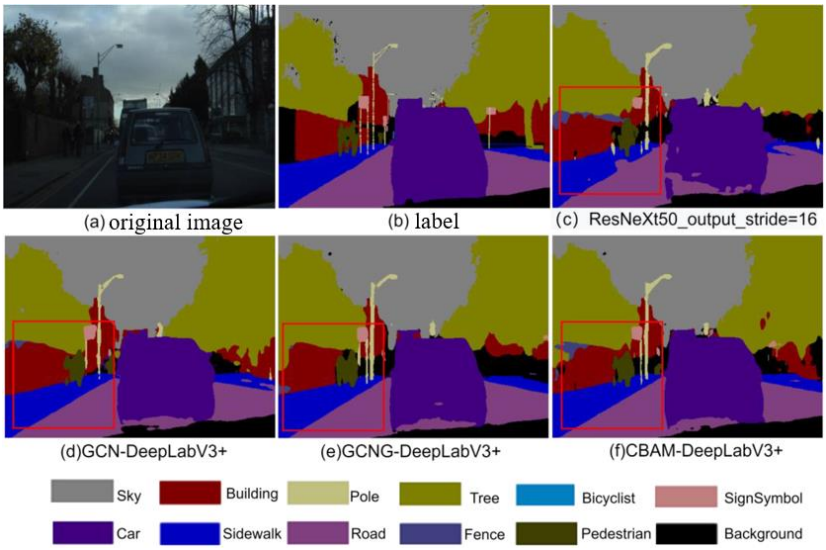


Fig 5.5 Comparison of Model Prediction Image Details

Tab 5.5 Comparison of Precision Statistics of Different Models in the CamVid Dataset

Model	MIoU
SegNet[24]	60.1
Bayesian SegNet[73]	63.1
DFANet[30]	64.7
ICNet[74]	67.1
BiSeNet[75]	68.7
Baseline	68.78
GCN-DeepLabV3+	70.03
CBAM-DeepLabV3+	70.51
GCNG-DeepLabV3+	**71.06**

(1) From the overall accuracy statistics table (Table 5.4), it can be observed that the improved GCN-DeepLabV3+, GCNG-DeepLabV3+, and CBAM-DeepLabV3+ models, with minimal changes in parameters and complexity, achieved an increase of 1.25, 2.28, and 1.73 in MIoU values, respectively, compared to the Baseline network. Specifically, the results of the multi-scale fusion GCNG-DeepLabV3+ model are superior to the stepwise summation fusion GCN-DeepLabV3+ model, but the former has higher model complexity and longer training iterations. In contrast to using a global convolution module to further extract local features for model optimization, adding a lightweight attention mechanism to shallow layers also effectively enhances the segmentation performance of the DeepLabV3+ model, with a simpler model construction.

(2) The segmentation accuracy statistics for different categories (Tables 5.2 and 5.3) show that the GCNG-DeepLabV3+ model outperforms the GCN-DeepLabV3+ model in various segmentation categories. Specifically, it exhibits significant improvements, especially for smaller-scale objects with fewer pixels, such as street lamps and pedestrians, with gains of 6.26 and 5.64 MIoU, respectively. This demonstrates that the local features extracted by the global convolution module are effectively integrated into the final segmentation predictions, improving the model's segmentation performance. For the CBAM-DeepLabV3+ model, after attention mapping in both channel and spatial dimensions, there is an overall improvement across various categories, with the pedestrian category showing the most significant improvement, with a 5.53 MIoU increase. This confirms that adding an attention mechanism to local features in shallow

layers helps the model capture more crucial fine-grained details, thereby enhancing segmentation performance.

(3) The improved model's segmentation prediction images (Figure 5.4) and detail comparison images (Figure 5.5) reveal that the GCN-DeepLabV3+, GCNG-DeepLabV3+, and CBAM-DeepLabV3+ models perform well in capturing fine details, resulting in clearer boundaries of target objects. GCN-DeepLabV3+ and GCNG-DeepLabV3+ models further reduce inter-class errors and perform better on background categories through feature fusion. Meanwhile, CBAM-DeepLabV3+ excels in capturing internal details of segmented objects, showing improved internal consistency of target objects.

(4) The experiments demonstrate that using global convolution modules in shallow layers further enriches the locally extracted features, benefiting feature utilization in segmentation tasks. Combining boundary refinement modules results in smoother and more detailed segmentation prediction images. Additionally, adding attention mechanisms to shallow features helps the model capture more critical and important feature information, leading to better segmentation results. The experimental results confirm the effectiveness of the decoding structure improvements, providing valuable insights for the enhancement and application of the DeepLabV3+ model.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusion

(1) The DeepLabV3+ model is a powerful semantic segmentation model with excellent segmentation performance. However, its complex backbone feature extraction network structure and large number of parameters can be challenging for training and practical applications. Depending on specific requirements, different convolutional neural networks can be chosen for feature extraction. During model training, adjusting the downsampling layer stride to reduce feature map resolution loss can effectively improve segmentation accuracy. Experimental comparisons using ResNet50, ResNeXt50, and MobileNetV2 as the DeepLabV3+ model's backbone networks indicate that when the output stride of the model is set to 8x, the best segmentation results are achieved. When the backbone network's output stride is set to 16x, a balance is achieved between model parameters, complexity, and segmentation performance.

(2) Sufficient representational capacity is crucial for model applications, particularly in the context of high-precision semantic segmentation for autonomous driving. In urban road scenes, the original DeepLabV3+ model exhibited shortcomings in segmenting small-scale objects and capturing object boundary details. To further enhance its segmentation performance, this study proposed improvements to the model's decoding structure. Firstly, by employing global convolution modules and boundary refinement modules to extract local features from shallow layers, the GCN-DeepLabV3+ and GCNG-DeepLabV3+ models, which incorporate more detailed information, demonstrated improved representation capabilities in capturing image details. Secondly, by utilizing convolutional block attention modules to map local features in both channel and spatial dimensions, the CBAM-DeepLabV3+ model, equipped with an attention mechanism, captured critical local features and enhanced segmentation performance for small-scale objects and boundary details. Experimental results suggest that the improved decoding structure of the DeepLabV3+ model is better suited for semantic segmentation tasks in urban road scenes.

6.2 Outlook

(1) For the semantic segmentation of small-sample urban street image data, unsupervised learning methods can be explored for model training. Additionally, domain adaptation methods that leverage well-trained models on high-computational devices can be employed to adapt models for practical use.

(2) While the improved models achieved favorable segmentation results on the CamVid dataset, real-world urban street scenes often feature more distractions and complexity. Further validation of model segmentation performance in actual urban environments is necessary. Additionally, the applicability of the improved DeepLabV3+ model to other scenarios and datasets requires further research and verification.

(3) In pixel-wise dense prediction tasks like semantic segmentation, large model parameter sizes can hinder practical application due to memory and computational constraints. Balancing segmentation accuracy and memory efficiency remains a key challenge in the deployment of real-time semantic segmentation models.

(4) Currently, supervised semantic segmentation techniques are well-developed, but weakly supervised and unsupervised methods are more aligned with real-world requirements. Research on leveraging weakly supervised and unsupervised semantic segmentation methods to improve segmentation accuracy on small-sample datasets is a future research focus.

CHAPTER 7: REFLECTION

During the course of completing this paper, I gained insights into the historical development of semantic segmentation networks and learned about the optimization directions in neural networks. While optimizing the DeepLabv3+ network structure, I developed a deeper understanding of the architecture of each layer, which will be beneficial for future network improvements.

REFERENCES

[1] L. Zhang, T. Pei, Y. Chen, et al., "A Review of Urban Environmental Assessment Based on Street View Images," Journal of Earth Science Information, vol. 21, no. 1, pp. 46-58, 2019.

[2] Z. Fan, B. Z. Liu, et al., "Measuring Human Perceptions of a Large-Scale Urban Region Using Machine Learning," Landscape and Urban Planning, vol. 180, pp. 148-160, 2018.

[3] Y. Zhang, X. Fu, "Quantitative Interpretation of Visual Environment Based on Deep Learning Street Image Interpretation and Aesthetics Ecology," Acta Ecologica Sinica, vol. 40, no. 22, pp. 8191-8198, 2020.

[4] L. Porzi, S. R. Bulò, B. Lepri, et al., "Predicting and Understanding Urban Perception with Convolutional Neural Networks," in Proceedings of the 23rd ACM International Conference on Multimedia, ACM, 2015, pp. 139-148.

[5] X. Li, C. Zhang, W. Li, "Does the Visibility of Greenery Increase Perceived Safety in Urban Areas? Evidence from the Place Pulse 1.0 Dataset," ISPRS International Journal of Geoinformation, vol. 4, no. 3, pp. 1166-1183, 2015.

[6] X. Li, C. Zhang, W. Li, "Building Block Level Urban Land-Use Information Retrieval Based on Google Street View Images," Giscience & Remote Sensing, vol. 54, no. 6, pp. 819-835, 2017.

[7] K. Hara, V. Le, J. Froehlich, "Combining Crowdsourcing and Google Street View to Identify Street-Level Accessibility Problems," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2013, pp. 631-640.

[8] Z. Chen, A. Li, X. Qiu, et al., "Research Status of Intelligent Vehicle Environmental Visual Perception and Its Key Technologies," Journal of Hebei University of Science and Technology, vol. 40, no. 1, pp. 15-23, 2019.

[9] F. Pan, "Research Progress on Reinforcement Learning in Autonomous Driving Control Technology," Journal of Image and Graphics, vol. 26, no. 1, pp. 28-35, 2021.

[10] L. Wang, C. Yan, "A Review of Road Scene Semantic Segmentation," Progress in Laser and Optoelectronics, vol. 58, no. 12, pp. 44-66, 2021.

[11] Y. Yang, W. Zhao, Z. Wang, et al., "Research on the Normalized Cut Image Segmentation Method Based on Graph Theory," Computer and Modernization, 2010, (1): 113-116.

[12] Q. H. Zheng, W. Q. Li, W. H. Hu, et al., "An Interactive Image Segmentation Algorithm Based on Graph Cut," Procedia Engineering, 2012, 29: 1420-1424.

[13] Z. W. Jing, H. Y. Guan, D. F. Peng, et al., "A Comprehensive Review of Image Semantic Segmentation Based on Deep Neural Networks," Computer Engineering, 2020, 46(10): 117.

[14] X. Tian, L. Wang, Q. Ding, "A Comprehensive Review of Image Semantic Segmentation Methods Based on Deep Learning," Journal of Software, 2019, 30(2): 440-468.

[15] C. Qing, J. Yu, C. B. Xiao, J. Duan, "Research Progress in Deep Convolutional Neural Networks for Image Semantic Segmentation," Journal of Image and Graphics, 2020, 25(06): 1069-1090.

[16] P. Huang, Q. Zheng, C. Liang, "A Comprehensive Review of Image Segmentation Methods," Journal of Wuhan University (Natural Science Edition), 2020, 66(06): 519-531.

[17] H. Xu, Y. Zhu, T. Zhen, Z. H. Li, "A Comprehensive Review of Deep Neural Network Image Semantic Segmentation Methods," Computer Science and Exploration, 2021, 15(01): 47-59.

[18] E. Shelhamer, J. Long, T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[19] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, "Semantic Image Segmentation

with Deep Convolutional Nets and Fully Connected CRFs," arXiv preprint arXiv:1412.7062, 2014.
[20] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, "Understanding Convolution for Semantic Segmentation," in Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 1451-1460.
[21] A. Paszke, A. Chaurasia, S. Kim, et al., "ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation," arXiv:1606.02147v1, 2016.
[22] S. Mehta, M. Rastegari, A. Caspi, et al., "ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation," in Berlin, Germany: Springer, 2018, pp. 561-580.
[23] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, "Deformable Convolutional Networks," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 764-773.
[24] V. Badrinarayanan, A. Kendall, R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.
[25] H. Noh, S. Hong, B. Han, "Learning Deconvolution Network for Semantic Segmentation," in Proceedings of 2015 IEEE International Conference on Computer Vision, Santiago, Chile: IEEE Computer Society, 2015, pp. 1520-1528.
[26] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Oct 5-9, 2015. Berlin, Heidelberg: Springer, 2015, pp. 234-241.
[27] G. S. Lin, A. Milan, C. H. Shen, et al., "RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York: IEEE Press, 2017, pp. 5168-5177.
[28] C. Peng, X. Zhang, G. Yu, G. Lulo, J. Sun, "Large Kernel Matters - Improve Semantic Segmentation by Global Convolutional Network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1743-1751.
[29] Y. Wang, Q. Zhou, J. Liu, et al., "LEDNet: A Lightweight Encoder-Decoder Network for Real-Time Semantic Segmentation," in Proceedings of 2019 IEEE International Conference on Image Processing, Washington D. C., USA: IEEE Press, 2019, pp. 1860-4864.
[30] H. C. Li, P. F. Xiong, H. Q. Fan, et al., "DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation," in Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Washington D. C., USA: IEEE Press, 2019, pp. 9522-9531.
[31] T. Y. Lin, P. Dollár, R. Girshick, et al., "Feature Pyramid Networks for Object Detection," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017, pp. 936-944.
[32] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," in arXiv:1606.00915, 2016.
[33] L. C. Chen, G. Papandreou, F. Schroff, H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," in arXiv:1706.05587, 2017.
[34] W. Liu, A. Rabinovich, A. Berg, "ParseNet: Looking Wider to See Better," in Computer Science, 2015.
[35] H. S. Zhao, J. P. Shi, X. J. Qi, et al., "Pyramid Scene Parsing Network," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017, pp. 6230-6239.

[36] J. J. He, Z. Y. Deng, Y. Qiao, "Dynamic Multi-Scale Filters for Semantic Segmentation," in Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019, pp. 3561-3571.
[37] H. S. Zhao, Y. Zhang, S. Liu, et al., "PSANet: Point-Wise Spatial Attention Network for Scene Parsing," in Computer Vision - ECCV 2018, V. Ferrari, M. Hebert, and C. Sminchisescu (Eds.), Lecture Notes in Computer Science, Cham: Springer, 2018, pp. 270-286.
[38] H. C. Li, P. F. Xiong, J. An, et al., "Pyramid Attention Network for Semantic Segmentation," in arXiv:1805.10180, 2018.
[39] Z. L. Huang, X. G. Wang, L. C. Huang, et al., "CCNet: Criss-Cross Attention for Semantic Segmentation," in Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Washington D. C., USA: IEEE Press, 2019, pp. 603-612.

[40] J. Fu, J. Liu, H. J. Tian, et al., "Dual Attention Network for Scene Segmentation," in Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Washington D. C., USA: IEEE Press, 2019, pp. 3146-3154.
[41] G. Lin, C. Shen, I. Reid, et al., "Deeply Learning the Messages in Message Passing Inference," in Computer Science, M. Hebert, B. Leibe, J. Matas, N. Sebe, and C. Sminchisescu (Eds.), Lecture Notes in Computer Science, Cambridge, USA: MIT Press, 2015.
[42] A. Arnab, S. Jayasumana, S. Zheng, et al., "Higher Order Conditional Random Fields in Deep Neural Networks," in Computer Vision - ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling (Eds.), Lecture Notes in Computer Science, Cham: Springer, 2016, pp. 524-540.
[43] R. Vemulapalli, O. Tuzel, M. Y. Liu, R. Chellapa, "Gaussian Conditional Random Field Network for Semantic Segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3224-3233.
[44] Z. W. Liu, X. X. Li, P. Luo, et al., "Semantic Image Segmentation via Deep Parsing Network," in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Washington D. C., USA: IEEE Press, 2015, pp. 1377-1385.
[45] F. Visin, A. Romero, K. Cho, et al., "ReSeg: A Recurrent Neural Network-Based Model for Semantic Segmentation," in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Washington D. C., USA: IEEE Press, 2016, pp. 41-48.
[46] Z. Li, Y. K. Gan, X. D. Liang, et al., "LSTM-CF: Unifying Context Modeling and Fusion with LSTMs for RGB-D Scene Labeling," in Computer Vision - ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling (Eds.), Lecture Notes in Computer Science, Berlin, Germany: Springer, 2016, pp. 541-557.
[47] X. D. Liang, X. H. Shen, J. S. Feng, et al., "Semantic Object Parsing with Graph LSTM," in Computer Vision - ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling (Eds.), Lecture Notes in Computer Science, Berlin, Germany: Springer, 2016, pp. 125-143.
[48] P. Luc, C. Couprie, S. Chintala, et al., "Semantic Segmentation Using Adversarial Networks," in arXiv:1611.08408v1, 2016.
[49] Y. Xue, T. Xu, H. Zhang, et al., "SegAN: Adversarial Network with Multi-Scale Loss for Medical Image Segmentation," in Neuroinformatics, 2018, 16(3/4): 383-392.
[50] J. F. Dai, K. M. He, J. Sun, "BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation," in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015, pp.

1635-1643.
[51] M. Rajchl, M. C. H. Lee, O. Oktay, et al., "DeepCut: Object Segmentation from Bounding Box Annotations Using Convolutional Neural Networks," in IEEE Transactions on Medical Imaging, vol. 36, no. 2, pp. 674-683, 2017.
[52] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple Does It: Weakly Supervised Instance and Semantic Segmentation," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA: IEEE, 2017, pp. 1665-1674.
[53] D. Lin, J. F. Dai, and J. Y. Jia, "ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation," in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Washington D. C., USA: IEEE Press, 2016, pp. 3159-3167.
[54] M. Tang, A. Djelouah, F. Perazzi, et al., "Normalized Cut Loss for Weakly-Supervised CNN Segmentation," in Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Washington D. C., USA: IEEE Press, 2018, pp. 1818-1827.
[55] A. Obukhov, S. Georgoulis, D. X. Dai, et al., "Gated CRF Loss for Weakly Supervised Semantic Image Segmentation," in arXiv:2003.13067, 2020.
[56] P. O. Pinheiro and R. Collobert, "From Image-Level to Pixel-Level Labeling with Convolutional Networks," in Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA: IEEE, 2015, pp. 1713-1721.
[57] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Point-Wise Localization and Segmentation," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA: IEEE Computer Society, 2017, pp. 5957-5966.
[58] A. Kolesnikov and C. H. Lampert, "Seed, Expand and Constrain: Three Principles for Weakly Supervised Image Segmentation," in Proceedings of the European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, Cham: Springer, 2016, vol. 9908, pp. 695-711.
[59] Y. C. Wei, X. D. Liang, Y. P. Chen, et al., "STC: A Simple to Complex Framework for Weakly Supervised Semantic Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2314-2320, 2017.
[60] J. Xu, A. G. Schwing, and R. Urtasun, "Learning to Segment Under Various Forms of Weak Supervision," in Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Washington D. C., USA: IEEE Press, 2015, pp. 3781-3790.
[61] G. Papandreou, L. C. Chen, K. P. Murphy, et al., "Weakly-and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation," in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Washington D. C., USA: IEEE Press, 2015, pp. 16-23.
[62] S. Hong, D. Yeo, S. Kwak, et al., "Weakly Supervised Semantic Segmentation Using Web-Crawled Videos," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Washington D. C., USA: IEEE Press, 2017, pp. 7322-7330.
[63] Y. H. Zhang, Z. F. Qiu, T. Yao, et al., "Fully Convolutional Adaptation Networks for Semantic Segmentation," in Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA: IEEE Press, 2018, pp. 6810-6818.
[64] Y. Cheng, F. Wei, J. Bao, et al., "Dual Path Learning for Domain Adaptation of Semantic Segmentation," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9082-9091.

[65] J. H. Cho, U. Mall, K. Bala, et al., "PiCIE: Unsupervised Semantic Segmentation using Invariance and Equivariance in Clustering," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16794-16804.
[66] S. H. Gao, Z. Y. Li, M. H. Yang, et al., "Large-Scale Unsupervised Semantic Segmentation," arXiv preprint arXiv:2106.03149, 2021.
[67] R. Harb, P. Knöbelreiter, "InfoSeg: Unsupervised Semantic Image Segmentation with Mutual Information Maximization," arXiv preprint arXiv:2110.03477, 2021.
[68] L. C. Chen, Y. Zhu, G. Papandreou, et al., "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801-818.
[69] K. He, X. Zhang, S. Ren, et al., "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.
[70] S. Xie, R. Girshick, P. Dollár, et al., "Aggregated Residual Transformations for Deep Neural Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1492-1500.
[71] M. Sandler, A. Howard, M. Zhu, et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520.
[72] S. Woo, J. Park, J. Y. Lee, et al., "CBAM: Convolutional Block Attention Module," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3-19.
[73] A. Kendall, V. Badrinarayanan, R. Cipolla, "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding," in Proceedings of the British Machine Vision Conference (BMVC), London, UK: British Machine Vision Association, 2017, pp. 1-2.
[74] H. Zhao, X. Qi, X. Shen, et al., "ICNet for Real-Time Semantic Segmentation on High-Resolution Images," in "ICNet for real-time semantic segmentation on high-resolution images," in Berlin, Germany: Springer, 2017.
[75] C. Yu, J. Wang, C. Peng, et al., "BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation," in Berlin, Germany: Springer, 2018, pp. 334-349.