Development of an Attention Mechanism Based Multi-Layer Feature Fusion Module in Semantic Segmentation

by

Sen Shao

Supervisor：Nik Whitehead

List of Figures

List of Tables

I. INTRODUCTION

Semantic segmentation, as a basic and critical task in computer vision, had the core objective of accurately classifying each pixel in an image into semantic categories. In recent years, with the rapid advancement of deep learning techniques, a major breakthrough had been achieved in convolutional neural network-based semantic segmentation methods [1]. In this progress, the R-CNN family of models, including Fast R-CNN [2], Faster R-CNN [3], and Mask R-CNN [4], significantly improved the efficiency and accuracy of segmentation through the introduction of innovative mechanisms such as Region Proposal Network (RPN) and mask branching. Meanwhile, FCN (Fully Convolutional Network) provided a concise and efficient solution to the pixel-level classification problem through its end-to-end fully convolutional architecture [5]. In driving the development of semantic segmentation technology, different levels of feature layers carried different information and meanings. [6] and [7] had proposed that shallow features are usually rich in details and localisation information, while deeper features contained more semantic information about the object, and multi-layer feature fusion is gradually becoming a key factor in improving model performance. For example, techniques such as FPN (Feature Pyramid Network) [8] and ASPP (Adaptive Spatial Pyramid Pooling) [9] greatly enhanced the model's in-depth understanding of the image content by cleverly integrating the image features at different scales and levels. These strategies not only improve the accuracy of segmentation, but also extend the model's ability to handle multi-scale objects.

Although current semantic segmentation models had made significant progress in feature fusion techniques, existing strategies still faced challenges in achieving effective information integration. The design of HRNet [10] focused on fusing multiple layers of features to integrate different levels of information, but this process was extremely sensitive to the efficiency of information transfer. As the transfer of information through the network could lead to loss of detail and semantics, HRNet specifically retained high-resolution feature branches to ensure that more of the original information is retained in the final output. In addition, some traditional methods may have failed to fully consider the importance of contextual information during feature fusion, which may lead to the omission of critical global information during segmentation. To overcome this limitation, specialised context modules, such as OCNet [11] and OCRNet [12], had been introduced. The purpose of these modules was to capture and integrate global contextual information during the feature fusion process, thus enhancing the model's ability to fully understand the scene. In addition, as the size of the dataset grew and the complexity of the scene increased, the model still faced challenges in dealing with complex scenes containing multiple overlapping objects [13]. This requires

models to have been able to perform selective integration when fusing similar information to optimise segmentation performance. For example, AttaNet proposed an attention fusion module in the feature fusion module, which obtained the fusion coefficients between different layers through the interaction of different layers, enabling the model to identify and fuse key features more accurately [14].

A. Research Problem

However, these studies aimed to fuse the features but ignored the up-sampling process in multilayer feature fusion [9], [15], [16], [17]. Bilinear interpolation was used in the up-sampling process, but this introduced erroneous information and led to performance degradation. This study mainly explores how to fix the up-sampling error information during the multilayer feature fusion process.

B. Research Aim

The aim of this study is to develop a multi-layer feature fusion module based on the attention mechanism. By improving the feature fusion strategy, the prediction accuracy of the model and its ability to adapt to complex scenes are improved as much as possible.

C. Research Objectives

1) Develop an improved up-sampling technique to reduce the error caused by bilinear interpolation in the process of multilayer feature fusion, to improve the prediction accuracy of semantic segmentation models significantly.
2) Experimentally validate the robustness of the new up-sampling technique when dealing with complex scenes and targets of different scales, ensuring that the model maintains high performance on diverse datasets.
3) To evaluate and demonstrate the generalisation capability and integration simplicity of the proposed up-sampling techniques in different semantic segmentation model frameworks, to facilitate their adoption in a wider range of research and application scenarios.

II. Literature Review

This section will be divided into four core parts for in-depth discussion. The first part introduces the basic concepts and technological advances in semantic segmentation. The second part describes the theoretical foundations of multi-layer feature fusion, key techniques, and contributions to improving the performance of semantic segmentation. The third part introduces the application and extension of the attention mechanism to language segmentation. The fourth part introduces how Transformer is applied to the field of semantic segmentation.

A. Semantic Segmentation

The application of deep learning techniques in semantic segmentation is rapidly becoming the forefront of AI research, to accurately recognize each region in an image and assign it a corresponding semantic label, allowing computers to gain a deeper understanding of the image content. Advances in deep learning for image recognition have skyrocketed since AlexNet's historic breakthrough in the ImageNet competition in 2012, which not only validated the potential of deep convolutional neural networks for large-scale image recognition tasks, but also inspired subsequent research with its innovative architecture [18]. GoogleNet introduces the Inception module, which realizes the multi-scaling and multi-feature recognition. It also realizes the parallel processing of multi-scale features, greatly improves the efficiency and accuracy of feature extraction, and provides the possibility of fine regional division in semantic segmentation tasks [19]. ResNet solves the problem of gradient vanishing in the training of the deep network through the residual learning framework, which allows for the construction of a deeper network, and paves the way for the development of semantic segmentation technology [20].

The development of the R-CNN series of models is not only a revolutionary breakthrough of deep learning in computer vision, but also provides rich experience and profound insights for the development of semantic segmentation technology. The innovation of the R-CNN [1] model is that it is the first time that the deep learning technique is applied to the large-scale object detection task. It efficiently extracts regions that may contain target objects through a selective search algorithm, and then utilizes a pre-trained convolutional neural network for feature extraction. Compared with traditional manual feature extraction methods, R-CNN is able to capture richer and more abstract feature representations, which largely improves the accuracy and reliability of object detection. Fast R-CNN is further optimized on the basis of R-CNN, which significantly improves the computational efficiency. It saves computational resources by sharing the computation of convolutional layers and reducing the repetitive forward propagation process. In addition, Fast R-CNN [2] introduces Region Proposal Network (RPN), which is

an innovative structure

that is not only used to generate high-quality candidate regions, but also shares convolutional features with the detection network, further improving the detection accuracy while avoiding the addition of extra computational burdens. Faster R-CNN, on the other hand, is a further optimization of the Fast R-CNN is another important improvement of Fast R-CNN, which integrates RPN into the network structure and realizes an end-to-end training process. This design not only further improves the detection speed, but also makes the generation of candidate regions more closely integrated with the target detection process. Faster R-CNN [3] significantly improves the detection efficiency while maintaining high accuracy, providing strong support for real-time object detection tasks.

Based on the R-CNN family, Mask R-CNN further extends this framework to not only perform object detection, but also to be able to accurately segment the pixel-level mask of each instance. Mask R-CNN [4] achieves the object by adding a branch, the mask branch, to Faster R-CNN, and utilizing a convolutional neural network to predict the mask of each candidate region, enabling the instance's accurate contour segmentation. This improvement not only improves the segmentation accuracy, but also provides richer information for understanding the contour of each individual object in the image. Despite the remarkable achievements of the R-CNN family in the field of object detection, its application to semantic segmentation tasks still faces challenges. The multi-stage processing flow of the R-CNN family of models increases the computational cost and limits the feasibility of the models in real-time applications. For example, in the field of autonomous driving, the model's recognition results are crucial for vehicle navigation and decision-making. Incorrect prediction results not only lead to navigation errors of self-driving vehicles, but also may cause traffic accidents, resulting in damages to people and property. The proposal of FCN [5] has revolutionized the field of semantic segmentation. FCNs realize an end-to-end training approach by converting a traditional convolutional neural network into a fully convolutional architecture, which converts a semantic segmentation problem into a pixel-level classification task, as shown in Fig. 1. This not only simplifies the model design, but also significantly improves the speed and efficiency of segmentation. by learning the mapping from pixel to category, FCN is able to predict the category label of each pixel directly in a single network, achieving more fine-grained and accurate semantic segmentation.

With the continuous deepening of computer vision research, subsequent work has further optimized and improved on the basis of FCN. DeconvNet [21] firstly proposes an innovative network architecture to gradually recover the spatial resolution of the feature map by introducing an inverse convolutional layer. This innovation not only improves the segmentation accuracy, but also significantly enhances the network's ability to capture local features and boundary information, providing a new way of thinking

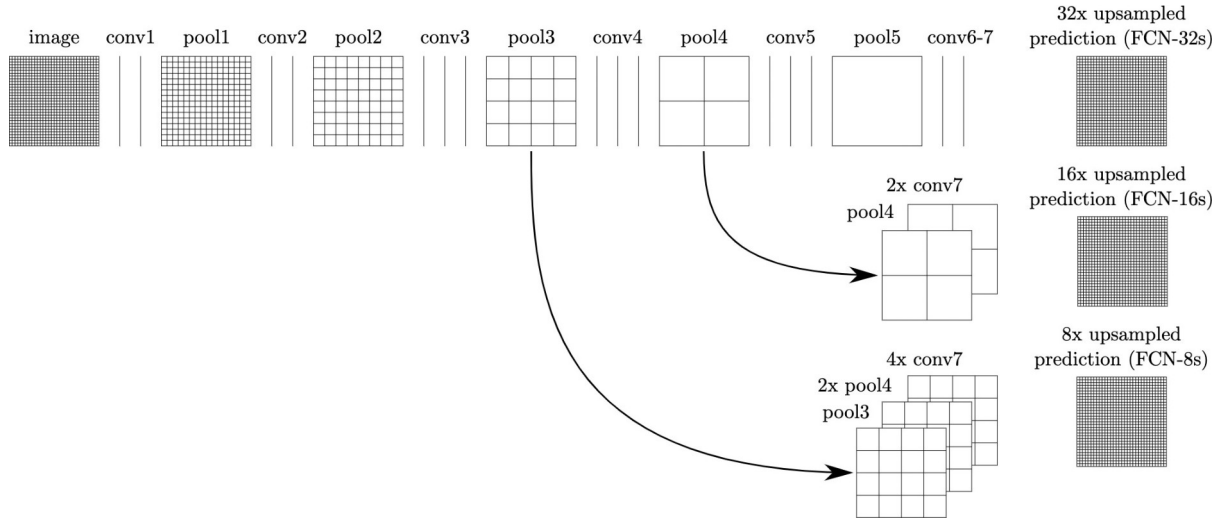

Fig. 1: The structure of the FCN (Fully Convolutional Network))[5]

about segmentation networks. SegNet [22] addresses some of the limitations of the traditional pixel-level semantic segmentation methods with its deep convolutional encoder-decoder architecture. SegNet's design is unique in the sense that, by encoder stacking followed by a corresponding decoder stacking, SegNet Effectively up-sampling the low-resolution feature maps to the same size as the input image, this design restores the spatial resolution while maintaining the classification accuracy, further advancing the development of semantic segmentation techniques. UNet [23], on the other hand, connects the high-resolution feature maps in the encoder directly to the corresponding layers in the decoder by establishing a jump connection between the encoder and decoder. This design not only preserves the detailed information in the image, but also significantly improves the accuracy of the network in making pixel-level predictions. The introduction of jump connections allows the network to fully utilize the rich details captured during the encoding process during decoding, which is particularly important for application scenarios such as medical image segmentation that require high accuracy. In addition, the exploration of techniques such as multi-scale feature fusion, attention mechanism, and adaptive pooling further improves the performance and robustness of semantic segmentation. The fusion of these techniques enables the network to simultaneously capture feature information at different scales, adaptively focus on key regions in the image, and dynamically adjust the pooling strategy based on contextual information to better handle objects of different shapes and sizes.

ParseNet [24] researchers have identified a key issue: although the fc7 layer of FCN theoretically has

a receptive field of up to 404×404 pixels, in practice, this receptive field is often significantly smaller, suggesting that there are limitations in FCN in capturing the global context of an image. This is an important obstacle for tasks that require understanding the overall semantic structure of an image, such as image segmentation and scene understanding. To overcome this challenge, the researchers introduced a global pooling strategy, which is an effective means of sensory wild expansion. Global pooling allows the network to integrate information from the entire feature map without changing the feature dimensions, thus significantly expanding the perceptual field of the model. This allows the network to capture the overall contextual information of the image more comprehensively, which is crucial for understanding the overall structure and semantics of the image. The global pooling strategy provides an efficient mechanism for the network to integrate and extract global features from images without adding excessive network parameters and computational overhead. Another approach to solve the sensory field problem is the Dilated Convolutional Network [25], which effectively increases the sensory field of the model by introducing dilation in the convolutional kernel, which expands the coverage of the convolutional kernel without increasing the number of parameters. This enables the model to capture a wider range of contextual information while maintaining high resolution. For tasks that require understanding the overall structure of an image, such as semantic segmentation and object detection, this ability to expand the perceptual field is particularly important, helping the model to more accurately recognize and segment different regions in an image. Overall, global pooling and null convolution are two effective techniques to expand the perceptual field, which can significantly enhance the model's ability to understand the overall structure and semantics of the image. These methods play a key role in computer vision tasks that require global contextual awareness, and provide efficient mechanisms for networks to integrate and extract global features from images, thereby improving model performance on such tasks.

In addition, RepLKNet [26] can significantly enhance the receptive field of the network by resizing the convolutional kernel to 31x31, thus allowing the model to integrate feature information over a wider range. This large-size convolutional kernel design may be more computationally intensive, but it shows great potential in capturing global features of an image. The large size of the convolutional kernel can cover a wider range of perceptual fields, allowing the model to integrate contextual information over a wider range to better understand the overall semantic structure of the image. This is crucial for computer vision tasks that require global perceptual capabilities, such as image segmentation and scene understanding. In addition to the two methods of pooling strategy and expanding convolutional range, Deformable Convolutional Networks [27], [28] provides a more advanced technology to enhance the adaptive ability of convolutional neural networks to the geometrical transformations of the object. By

introducing deformable convolutional and deformable pooling modules, DCN realizes adaptive sampling of the input feature maps, thus capturing the geometrical transformations more efficiently. The DCN can realize adaptive sampling of the input feature maps by introducing deformable convolution and deformable pooling module, so as to capture the shape and appearance changes of the objects more effectively. The core of deformable convolution lies in its ability to adaptively adjust the sampling position of the convolution kernel according to the actual shape and position of the target object. This flexibility is achieved by adding additional offsets to the fixed sampling points of traditional convolution operations, which are learned. The deformable pooling module improves the traditional RoI pooling operation by adding offsets to each sub-region of the RoI, allowing the pooling operation to capture key parts of the target object more accurately, rather than being constrained to a fixed grid structure. These innovations help the model to better adapt to deformations and changes in the target object, resulting in better performance in tasks that require sensitivity to object boundaries and appearance.

Multi-task learning is widely recognized as an important strategy to improve the network's ability to understand the real world. UperNet addresses some of the key challenges in multi-task learning through its innovative network design. In traditional multitask learning frameworks, different tasks usually share the same network backbone and segmentation header, which may lead to limited performance on specific tasks [29]. UperNet does this by customizing specialized prediction heads for each task, as shown in Fig. 2. The scene segmentation head helps to recognize environmental elements in the image, such as the sky and buildings. The object segmentation head accurately finds and describes objects such as pedestrians and vehicles. The material segmentation head distinguishes between different materials on the surface of an object, such as metal or fabric. The semantic segmentation head is further refined to recognize instances of all categories in the image, which is crucial for applications such as autonomous driving. These features enable UperNet to adapt to a variety of vision tasks and improve performance. MaskFormer represents another innovation in multi-task learning networks. It skillfully combines the powerful feature extraction capabilities of Convolutional Neural Networks (CNNs) with the self-attention mechanism of the Transformer architecture, especially in image segmentation tasks. The distinguishing feature of MaskFormer is its ability to perform semantic segmentation and instance segmentation at the same time without having to rely on additional network architectures or complex post-processing steps [30]. This achievement is made possible by its well-designed network architecture that efficiently extracts feature representations suitable for both segmentation tasks.

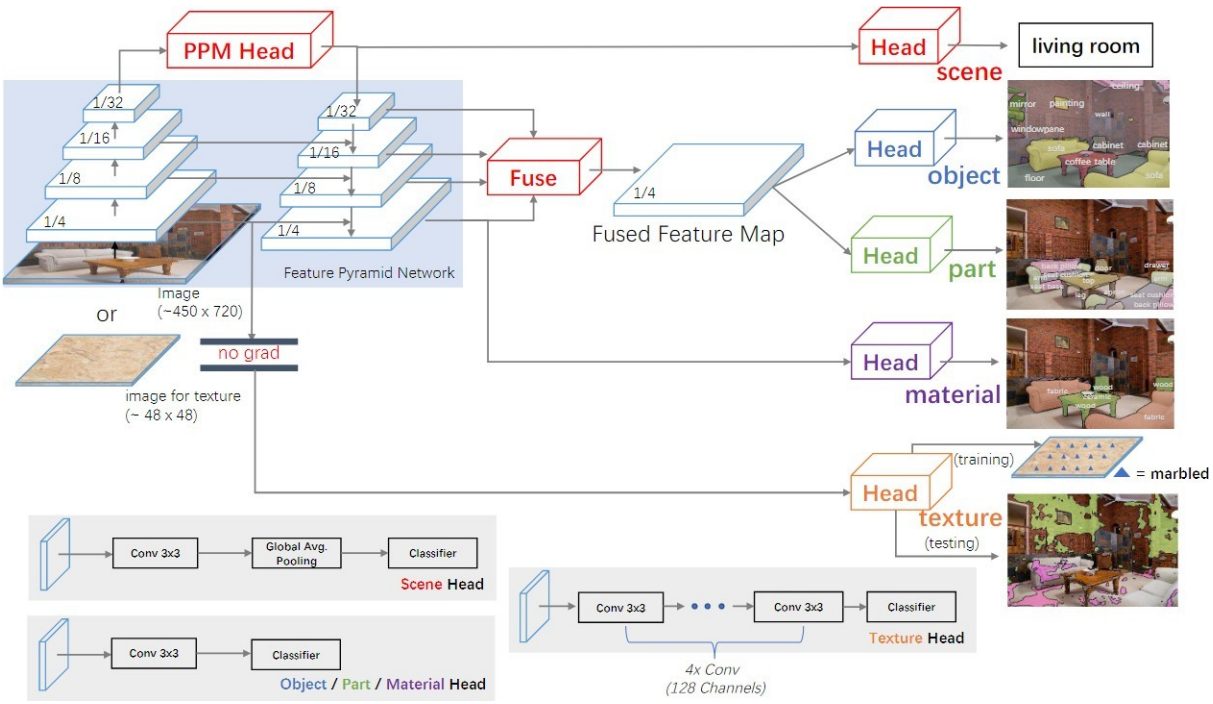

Fig. 2: The structure of the UperNet [29]

B. Multi-layer Feature Fusion

In the field of deep learning, feature fusion and multi-scale feature extraction are key techniques to improve the performance of target detection and semantic segmentation. Feature pyramid network [8] significantly improves the detection performance of objects of different sizes by constructing a multilevel feature pyramid that effectively integrates features of different scales. FPN utilizes the inherent multiscale nature of deep convolutional networks to construct a network structure that can provide rich high-level semantic features through lateral connectivity and top-down architecture. This structure not only significantly enhances the ability of the semantic segmentation model to detect small objects, but also maintains the accuracy of detecting large objects. However, Path Aggregation Network (PANet) points out that in the feature pyramid structure of FPN, the lower level features may experience information loss in the process of passing to the higher level, resulting in the loss of detailed information. To solve this problem, PANet [31] proposes an innovative solution by designing a direct path from the low level to the high level, which not only preserves the integrity of the information in the low level, but also provides more details and localization information, thus improving the model's ability to capture details.

Unlike FPN and PANet, Pyramid Scene Parsing Network (PSPNet) employs a unique pooling technique to generate multi-scale feature representations. PSPNet [32] captures rich image context information by pooling the feature map at different resolutions. This strategy not only constructs a feature pyramid, but also enables the network to synthesize global and local visual cues through feature fusion, which significantly improves the segmentation performance for objects at different scales.

The evolution of the DeepLab series of models in the field of semantic segmentation not only demonstrates the continuous innovation and advancement of deep learning technology, but also becomes an important milestone in understanding and interpreting visual scenes in the field of computer vision. The launch of DeepLab v1 introduces the Atrous Convolution technique for the first time, which effectively enlarges the sensory field by adjusting the null rate in the convolutional process, and at the same time maintaining computational efficiency and significantly improving segmentation accuracy, especially when dealing with images with complex textures and details [9]. DeepLab v2 further refines the segmentation edges by introducing Conditional Random Fields (CRF) as a post-processing step, which takes into account pixel interrelationships using a probabilistic graph model to optimize the smoothness and accuracy of the edges. Meanwhile, v2 employs the Depthwise Separable Convolution technique, which reduces the model parameters and computation by separately performing spatial depth convolution and point-by-point convolution on the input channels, realizing feature extraction effects similar to those of the standard convolution, while reducing the model complexity [15]. DeepLab v3 further pushes the technology forward by introducing the Skip Fusion decoder architecture, which realizes an effective combination of deep encoder features and shallow decoder features, allowing the network to simultaneously utilize deep semantic information and shallow spatial details, further improving segmentation accuracy, especially when dealing with complex scenes and images with multilevel structures [16], as shown in Fig. 3. DeepLab v3+ further extends the concept of multi-scale feature fusion by fusing multi-scale features in the encoder-decoder architecture, which enhances the segmentation capability of objects at different scales, enabling the network to more comprehensively understand and segment small objects and large backgrounds in an image, improving the ability to capture details as well as enhancing the grasp of the overall scene [17].

UNet employs the idea of feature fusion in the decoding stage, where high-resolution features in the encoder are directly connected to the corresponding layers in the decoder through skip connections, effectively fusing more detailed information [23]. UNet++ further redesigns these skip connections to allow the aggregation of features at different semantic scales in the sub-network of the decoder, realizing a more flexible scheme. A more flexible feature fusion scheme is realized. However, the effectiveness

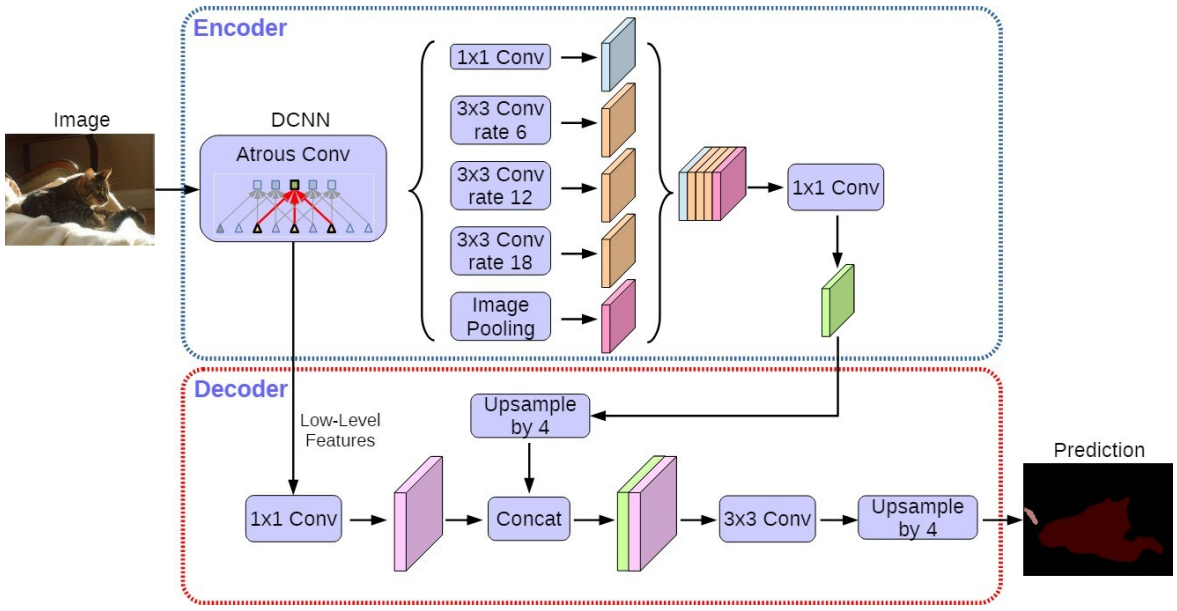

Fig. 3: The structure of the deeplab v3+ [17]

of feature fusion is somewhat dependent on the quality of the extracted features [33]. As the depth of the network increases, the features may suffer from loss during the delivery process, resulting in the loss of detailed information. TransUNet utilizes Transformer's global self-attention mechanism to encode image chunks from a CNN feature map, and then up-samples the encoded features through a decoder and combines them with a high-resolution CNN feature map for accurate localization [34]. This approach aims to overcome the limitations of U-Net in explicitly modelling remote dependencies while addressing the issue of low-level detail information that Transformer may lack in medical image segmentation.

In order to solve this problem, the RefineNet introduces a multi-scale feature fusion and refinement strategy. It effectively solves the problem of loss of detail information during feature transfer in traditional convolutional neural networks by maintaining high-resolution branches at each stage of the network. RefineNet further enhances the feature representation through Refine Modules, which are capable of extracting finer detail information from low-level features and combining it with high-level features, thus improving the segmentation accuracy. These modules can extract finer detail information from the lower level features and combine it with the higher level features, thus improving the accuracy and robustness of segmentation [35]. HRNet effectively solves the problem of loss of detail information in the feature- passing process of traditional convolutional neural networks by maintaining high-resolution branching at

each stage of the network. This design not only preserves rich detail information, but also significantly improves segmentation accuracy during feature fusion, especially in scenes requiring fine edge detection and small object recognition [10]. Nevertheless, existing structures may not fully consider the balance between multi-layer feature fusion. To solve this problem, AttaNet proposes a new multi-layer feature fusion method. It realizes the balance between the information inputs of different layers in the feature fusion process by interacting between the information of different layers, calculating their respective weight coefficients, and then multiplying these coefficients with the corresponding features [14]. This approach not only enhances the network's ability to integrate multi-scale information, but also further improves the performance of the segmentation task, especially when dealing with images with complex structures and multi-scale features.

C. Attention Mechanism

The integration of attention mechanisms has brought breakthrough enhancements in feature extraction and characterization for deep learning models. SENet [36] plays a pioneering role in this field by innovatively introducing the channel attention mechanism. This mechanism intelligently strengthens the interactions and synergies among channels by dynamically adjusting and re-calibrating the feature responses of each channel. This adaptive adjustment not only enhances the model's ability to capture key features, but also significantly improves the overall representation performance of the network. By accurately identifying and amplifying useful information while suppressing unimportant features, SENet optimizes the model's decision-making process, resulting in a quantum leap in performance across a wide range of visual tasks.

CBAM [37] further enhances the network's ability to capture features through well-designed channel attention modules and spatial attention modules. The channel attention module utilizes maximum pooling (maxpool) and average pooling (avgpool) to capture feature information in different dimensions, where maximum pooling is responsible for highlighting salient regions in the feature map, while average pooling enhances the representativeness of the channel, as shown in Fig. 4. This combination of max-pooling and average-pooling not only preserves the texture features of the image, but also preserves more background information, which enables the network to capture the key information more accurately and improves the performance effectively.BAM (Bottleneck Attention Module) fuses the channel attention and spatial at- tention through its innovative parallel processing architecture, effectively capturing the multi-dimensional features of the image [38]. BAM through its innovative parallel processing architecture, combines channel attention and spatial attention to effectively capture the interactions

between multi-dimensional features

and achieve comprehensive optimization of features. This design not only improves the expressive power of features and the perceptual ability of the model, but also maintains high computational efficiency, which enables BAM to be widely used in a variety of visual tasks, especially in scenarios such as small-target detection, which significantly improves the accuracy and robustness of detection.

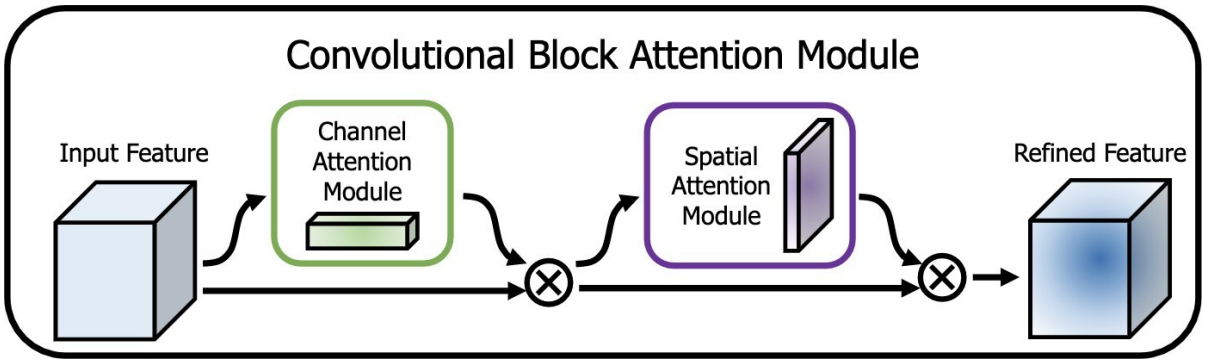

Fig. 4: The structure of the CBAM [37]

GAM (Global Attention Mechanism) enhances the performance of deep neural networks in small target detection tasks through its global attraction mechanism [39]. It simplifies and enhances the interaction of global features by reducing the information, especially emphasizes the importance of cross-dimensional information, and amplifies the channel-space interactions, which significantly improves the model's recognition accuracy of small targets. This mechanism provides a new perspective for the field of small target detection, and combines with other enhancement strategies such as data augmentation and multi-scale learning to jointly promote the development of small target detection technology. The Triplet module, through its innovative channel transformation technique, successfully breaks the constraint that the tradi- tional attention mechanism is only limited to single-channel internal interactions. This breakthrough not only facilitates the exchange of information between different channels, but also significantly enhances the expressive capability of features. The design of the Triplet [40] module cleverly utilizes the strategy of channel blending, through which it is able to capture more subtle and integrated features, thus providing a completely new perspective on feature fusion techniques. The design of this module allows the network to integrate information from different channels more flexibly during feature extraction, enhancing the model's comprehension of complex scenes. The parallel processing capability of the Triplet module ensures that both local and global contextual information can be taken into account during the feature fusion process, further enhancing the model's comprehensive perception of image content.

The breakthrough contribution of the Non-local Net [41] is to emphasize for the first time the central role of long-range cross-pixel communication in feature fusion, an innovation that goes beyond spatial, channel, and dimensional attention mechanisms that focus only on localization. The network brings a qualitative leap in deep learning performance in visual tasks by introducing an innovative long-range attention mechanism that enables the network to capture and correlate pixel features that are far apart in an image, thereby significantly improving the understanding of the overall scene. Following Non-local Net, CCNet [42] has further advanced the field by skillfully combining the mechanisms of long- and short- range attention, and by synergizing these two types of attention, feature information is comprehensively refined. This fusion not only effectively improves the richness and accuracy of feature representation, but also enhances the network's understanding of both local and global contexts and improves its adaptability to complex scenes. The design of CCNet enables the network to model the intrinsic structure of visual data in a more refined way by simultaneously considering direct interactions between pixels as well as indirect contextual relationships. This strategy of integrating long-range and short-range dependencies provides deep learning models with a more comprehensive view of features, leading to more accurate performance in visual tasks such as image classification, target detection, and semantic segmentation.

Axial-DeepLab [43] and CSWin [44] significantly improve the segmentation performance of the networks by introducing an advanced strip-attention mechanism, an innovation that focuses on optimizing the feature fusion process. Axial-DeepLab's Axial Attention Module is specifically designed to accurately capture long-range dependencies of an image along a particular direction, which not only optimizes the understanding of texture and structural features but also achieves finer boundary alignment in image segmentation tasks, and achieves finer boundary alignment, especially when dealing with images with complex textures and subtle structural variations. CSWin, on the other hand, employs an innovative convolutional window design, which enhances the localized and directional representation of the feature maps in a particular direction, making the network more sensitive to edge and texture variations. This sensitivity is crucial for accurately segmenting the detailed parts of an image, thus further improving the performance of the segmentation task.CSWin's directional focus is particularly suitable for processing image content that has distinct directional features, such as streamlined textures, aligned edges, and angularly specific shapes, which play a key role in image content understanding. With this directional feature fusion, both Axial-DeepLab and CSWin are better able to handle spatial contextual information in images, which is crucial for improving segmentation accuracy and robustness.

The importance of global contextual scene information for deep learning models should not be un-

derestimated, it provides the model with a comprehensive scene concept, which greatly facilitates the

model's deep understanding of the image foreground and background. This deep understanding plays a crucial role in accurately recognizing and separating objects in complex scenes. EncNet (Encoder-Decoder with Attention Mechanism) is based on such a concept, and significantly improves the network's ability to capture and utilize global contextual information by introducing a contextual attention mechanism [45]. By deeply interacting with contextual information, this mechanism not only facilitates effective differentiation of information, but also strengthens the model's grasp of the relationships among the elements in the scene, thus realizing a significant performance improvement in segmentation tasks.The design of EncNet cleverly fuses the encoder-decoder architecture with the attention mechanism, enabling the model to focus more on key information during feature extraction and reconstruction. This fusion strategy not only improves the model's sensitivity to local details, but also enhances its grasp of the global structure, resulting in more accurate pixel-level predictions when processing images with rich semantic information.

The innovation of OCNet [46] lies in its unique strategy of alternately applying global and local attention, a strategy that revolutionizes feature extraction and fusion for deep learning models. By flexibly switching between global and local features, this strategy greatly enhances the network's ability to process features at different scales, thus achieving significant improvements in the efficiency of feature fusion as shown in Fig. 5. Specifically, OCNet's global attention mechanism captures the macrostructure of an image, such as the overall layout and shape, which helps the model understand the global context of the image. The local attention mechanism, on the other hand, focuses on the microscopic details of the image, such as edges, textures, and small objects, which are crucial for accurate recognition of details. With this strategy, OCNet is not only able to handle large-scale features, but also able to pay attention to subtle feature variations, realizing the comprehensive capture and integration of macro-structures and micro-details in an image. OCRNet (Object-Contextual Representations) introduces an innovative fusion strategy in the deep learning model, which combines regional information with contextual information to achieve deep feature representation [12]. The core of this design is to emphasize the interrelationships between objects and their contexts, and through the understanding of such relationships, OCRNet is able to capture richer and semantic features, which can significantly improve the performance in tasks such as image segmentation. In addition, OCRNet's fusion strategy is not limited to improving segmentation performance; it also enhances the model's understanding of patterns and relationships implicit in the image, which is crucial for improving the model's generalization ability and adaptability to a wide range of visual tasks.

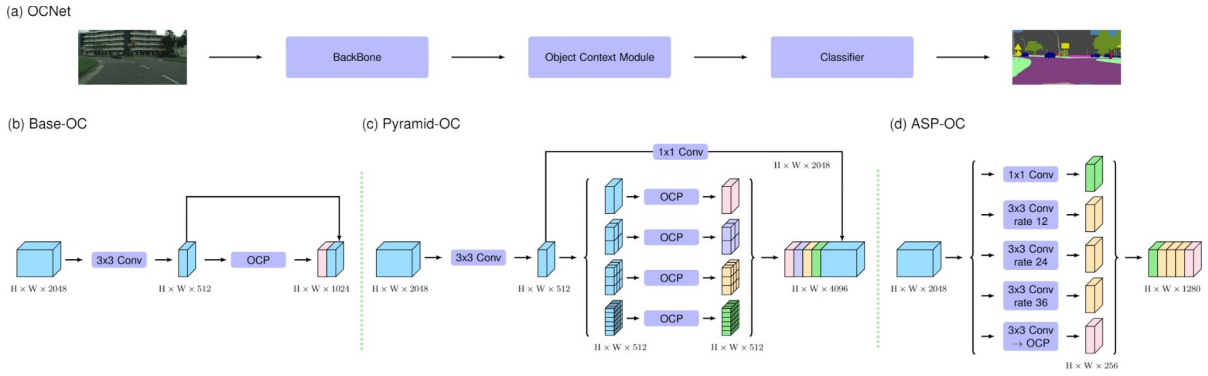

Fig. 5: The structure of the OCNet [46]

D. Transformer

Transformer [47] model has revolutionized the field of neural networks through its innovative attention mechanism. The core of this mechanism lies in the Self-Attention layer, which allows the model to process data in a way that not only focuses on local information, but also captures the global context, significantly improving the ability to understand sequential data. In particular, the Multihead Attention mechanism, which further enhances this capability of the model by processing information in parallel in multiple representation subspaces, allows the model to capture features of the data from different perspectives and scales simultaneously. Vision Transformer [48] is based on this principle and applies the Transformer architecture to image processing. ViT efficiently processes image data by segmenting the image into small chunks (i.e., Image Chunks or Patch) and then applying the attention mechanism. This approach breaks through the limitations of traditional CNNs and demonstrates a completely new way of image representation and processing. ViT demonstrates excellent performance in computer vision tasks, such as image classification, target detection, and segmentation, proving the potential of the Transformer model in processing visual information. Although ViT has achieved remarkable results in terms of performance, its huge demand on computational resources has limited its wide application in high-resolution image processing. Swin Transformer [49] is innovative in that it significantly improves computational efficiency by implementing the self-attention mechanism within locally non-overlapping windows and cleverly introducing cross-window connections. This design skillfully integrates the model's computational efficiency. This design cleverly links the computational complexity of the model to the image size, realizing a linear growth instead of the square growth in the traditional Transformer model, which effectively relieves the computational burden of processing high-resolution images.

Transformer architectures have been a catalyst for innovation in the field of computer vision, especially in semantic segmentation tasks. The breakthrough innovation of SETR (Semantic Encoding with Transformer) is that it abandons the traditional CNN-based encoder-decoder architecture in favor of a SETR (Semantic Encoding with Transformer) is a breakthrough innovation in that it abandons the traditional CNN-based encoder-decoder architecture in favor of a Transformer architecture to directly process images [50]. SETR is able to efficiently capture long-range dependencies between pixels by decomposing an image into a series of image chunks and applying a global self-attention mechanism, which transforms an image segmentation problem into a sequence-to-sequence prediction task, a method that is uniquely suited for capturing the global contextual information.SETR is designed to allow the network to maintain a high spatial resolution at each layer of the encoder to maintain a high spatial resolution, avoiding the downsampling operation in traditional CNNs, thus making the feature learning process more efficient and able to capture the detailed information of the image in greater detail. However, this maintenance of high resolution features also poses a computational challenge, especially when processing high resolution images. To address this challenge, SegFormer (Segmentation Transformer) introduces a hierarchical Transformer encoder, which is capable of efficiently outputting multi-scale feature maps.The design of SegFormer cleverly simplifies the decoding process, and optimizes the use of computational resources by efficiently fusing deep features from different layers. SegFormer is designed to simplify the decoding process by effectively fusing deep features from different layers, which not only optimizes the use of computational resources, but also significantly improves the model's adaptability to complex scenes [51]. This hierarchical fusion strategy not only reduces the computational burden, but also enhances the model's ability to capture features from different scales, thus realizing both accuracy and efficiency improvement in semantic segmentation tasks. In addition, SegFormer's advantage of multi-scale feature capture makes it perform well in processing images with rich details and complex structures.

DETR [52] cleverly circumvents the necessary candidate region generation and complex post-processing operations such as non-maximal suppression (NMS) in traditional target detection algorithms by revo- lutionizing the target detection problem by redefining it as an ensemble prediction task. This innovative approach not only simplifies the detection process, but also realizes an end-to-end learning path from the input image to the target category and bounding box prediction.The proposal of DETR provides a new perspective and solution to the target detection problem within the field of deep learning, and lays the foundation of a new paradigm for subsequent research and applications. MaskFormer v1 [53] represents an innovative instance segmentation field step, which draws on the

seminal ideas of DETR to reconceptualize the instance segmentation problem as an ensemble prediction task, as shown in Fig. 6. This

novel approach effectively simplifies the segmentation process by avoiding the complex components of traditional instance segmentation algorithms, such as region proposal networks (RPNs) and subsequent candidate region refinement steps.MaskFormer v1 enables the model to generate accurate pixel-level masks directly in an end-to-end training framework through the introduction of the mask prediction header, an innovative component. This design not only improves the accuracy of the masks generated by the model, but also maintains the efficiency and simplicity of the learning process, bringing significant performance gains to the instance segmentation task. Subsequently, MaskFormer v2 [54] was further improved and optimized based on v1. The v2 version enhances the model's ability to handle objects of different scales and complexity through a more refined design. v2 employs a more advanced attentional mechanism in the mask prediction header, which allows the model to capture the detailed information in the image more efficiently, especially when dealing with small objects or objects with a complex backgrounds. In addition, the v2 version may also include improvements to the loss function, more efficient feature fusion strategies, or further optimizations to the model architecture, all of which are aimed at improving the model's generalization ability and segmentation accuracy.

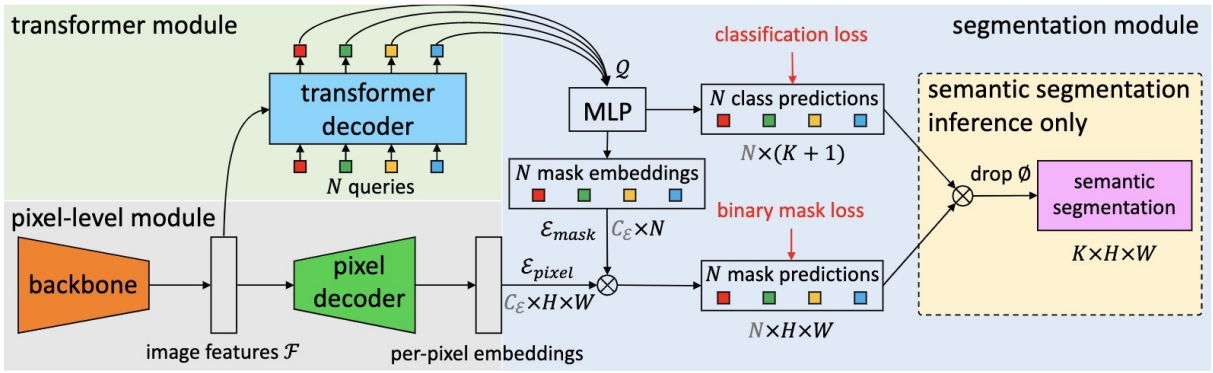

Fig. 6: The structure of the MaskFormer [53]

OneFormer [55] achieves multi-threaded parallel processing of key visual tasks such as semantic segmentation, instance segmentation, and target detection by proposing an innovative unified framework, an architecture that achieves significant achievements in model generalization capabilities. This multi-task learning approach not only optimizes the design of the multi-task visual system and reduces the redundancy of the model through the shared representation and feature extraction mechanism, but also significantly improves the efficiency and performance of the model in processing different visual tasks.OneFormer's multi-task learning strategy also facilitates the migration of the model's knowledge

and

information across tasks, and this migration learning capability further improves the model's adaptation to new tasks and reduces the reliance on large amounts of labeled data.

Since the introduction of Transformer technology into the field of computer vision, it has paved the way for the construction of large-scale visual-linguistic models. CLIP exemplifies this advancement, which is trained using a contrast learning approach to efficiently capture the deep semantic connections between images and text. In addition, CLIP's training process relies on a wide range of unlabeled image and text resources, which not only reduces the need for labeled data, but also enhances the model's generalization ability [56]. The pre-trained CLIP demonstrates excellent zero-sample learning capabilities and can be directly applied to diverse visual and linguistic tasks without additional task-specific fine-tuning, significantly enhancing the model's flexibility and utility. By incorporating the advanced multimodal capabilities of the CLIP model, SAM implements an innovative cue-based approach that greatly enhances the intuitiveness and flexibility of user interaction. Users can now specify exactly which objects they want the model to recognize and segment through a series of simple interactions such as clicks, drawing boxes and question inputs [57]. This cue-based interaction not only enhances the model's adaptability, allowing it to flexibly respond to a variety of complex visual tasks, but also significantly improves the user experience, making the model more intuitive and user-friendly to use.

III. Experimental Design

A. Philosophy

In the field of computer vision, semantic segmentation is a crucial task that requires a system to be able to accurately classify each pixel in an image into the correct category. This capability is critical for high-risk applications such as autonomous driving and medical imaging analysis, which rely on an accurate understanding of the scene to make safe and effective decisions. However, existing semantic segmentation models often face challenges in dealing with details and complex backgrounds in images, leading to the omission or misrecognition of such critical information, which can have serious consequences.

To address this challenge, we design an innovative multi-layer feature fusion module that integrates an attention mechanism to significantly enhance the model's ability to learn features at different layers. This design not only fixes possible misinformation during feature fusion and up-sampling, but also significantly improves the accuracy and robustness of semantic segmentation. Our module specifically utilizes the global perception capability of the attention mechanism to ensure that the model captures and understands every detail in the scene, thus enabling finer segmentation in complex scenes.

In building this multilayer feature fusion module based on the attention mechanism, we adopt the concept of multiscale feature learning, which effectively combines deep (semantically rich but lower resolution) and shallow (less semantic but higher resolution) feature maps through cross-layer connectivity techniques. This fusion not only preserves the details of the image, but also greatly enhances the semantic expressiveness of the features. In addition, the introduced attention mechanism enables the module to adaptively adjust the importance of each region in the feature fusion process, further enhancing the model's ability to integrate multi-scale features.

In the experimental phase, we integrated this multilayer feature fusion module into several mainstream semantic segmentation networks and conducted a comprehensive evaluation on several standard datasets. The experimental results show that our module significantly improves the segmentation accuracy, especially excelling in processing images with complex backgrounds and containing fine objects. In addition, through ablation studies, we validate the contribution of individual components in the module to the overall performance improvement, as well as the central role of the attention mechanism in the feature fusion process.

Our work not only advances semantic segmentation techniques, but also provides new ideas for other tasks in computer vision, such as target detection and instance segmentation. Through this innovative combination of multi-layer feature fusion and attention mechanisms, we expect to contribute to building

smarter and more reliable visual understanding systems, which in turn will enable deeper visual perception and decision support in a wide range of application scenarios. With the continuous progress and validation of the technology, we believe that this module will bring far-reaching impact to the field of computer vision.

B. Shortage

In the research field of semantic segmentation, multi-layer feature fusion technique has become a core strategy to enhance the performance of models. By integrating feature information from different layers, this technique significantly enhances the model's ability to detect and recognize small targets. For example, FPN (Feature Pyramid Network) effectively enhances the model's ability to perceive multi-scale objects by fusing multi-resolution feature maps.PSPNet (Pyramid Scene Parsing Network) further extends this concept by capturing multi-scale feature representations through pooling layers, deepening the understanding of the global context. DeepLab V3+, on the other hand, adopts an encoder-decoder architecture, which realizes comprehensive feature fusion from shallow to deep layers and greatly improves the recognition accuracy of object details.AttaNet introduces the concept of feature fusion coefficients, which emphasizes the importance of considering different layers of features during feature fusion, and these coefficients help the model to filter out the more discriminative features, further improving the model's performance. The development of these models demonstrates the great potential of multi-layer feature fusion in capturing fine-grained semantic information, laying a solid foundation for improving the accuracy and robustness of semantic segmentation tasks.

However, although these fusion strategies achieve a certain degree of performance improvement through up-sampling and feature map superposition, they do not fully address the information distortion that may occur during up-sampling, such as edge blurring and detail loss. Such distortions may accumulate during multi-scale feature synthesis, affecting the detail quality of segmentation results, especially in the recognition of small targets and complex backgrounds. In the long run, this may weaken the accuracy and robustness of the model in semantic segmentation tasks. To address these challenges, researchers need to explore more advanced feature fusion techniques that should be able to reduce information distortion while improving the model's ability to integrate multi-scale and multi-detail features, leading to higher quality semantic segmentation results. Future research will focus on developing more refined feature fusion methods to optimize model performance and ensure robustness in various complex scenarios.

C. Approach

In the research field of attention mechanism, SENet [36] adaptively adjusts the features of channels accordingly by self-attention mechanism, focusing on highlighting the important feature channels. CBAM [37] integrates the channel attention mechanism to enhance the features by maximum pooling and average pooling, which highlights the representation of important regions and channels in the feature map, respectively. Triplet [40], on the other hand, enhances the representation of features by focusing on the interactivity between channels externally and increasing the network's attention to the interactions between different channels.

This study is dedicated to the development of an innovative attention mechanism module that aims to break through the limitations of existing techniques in feature fusion. We propose a unique multi-layer feature fusion attention module by drawing on existing attention module designs. The module is especially designed for fine optimization of the feature maps generated by up-sampling through the attention mechanism after the up-sampling step. It is able to identify and enhance key elements in the image while suppressing irrelevant or interfering features to effectively correct possible distortions during the upsampling process.

This strategy not only significantly improves the prediction accuracy of the model, but also enhances the model's ability to recognize complex backgrounds and diverse objects. Through a well-designed attention mechanism, our module is able to capture and integrate information at different scales, thus achieving more accurate semantic segmentation at the feature level. The architectural details of the module are comprehensively demonstrated in Fig. 7, which elaborates its core components and workflows, showing its potential and advantages in enhancing semantic segmentation performance.

To further enhance the model performance, we also introduce a multi-scale feature fusion technique, which enhances the model's ability to represent features at different scales by combining deep semantic information with shallow texture details. This fusion strategy not only enriches the hierarchical structure of features, but also improves the model's recognition accuracy for small targets and fuzzy boundaries.

Extensive experimental validation of our module will be conducted, including evaluation on multiple public datasets and comparison with other state-of-the-art methods. The experimental results show that our module achieves excellent performance in a variety of scenarios, especially when dealing with images with complex backgrounds and fine objects, and significantly improves the accuracy and robustness of segmentation.

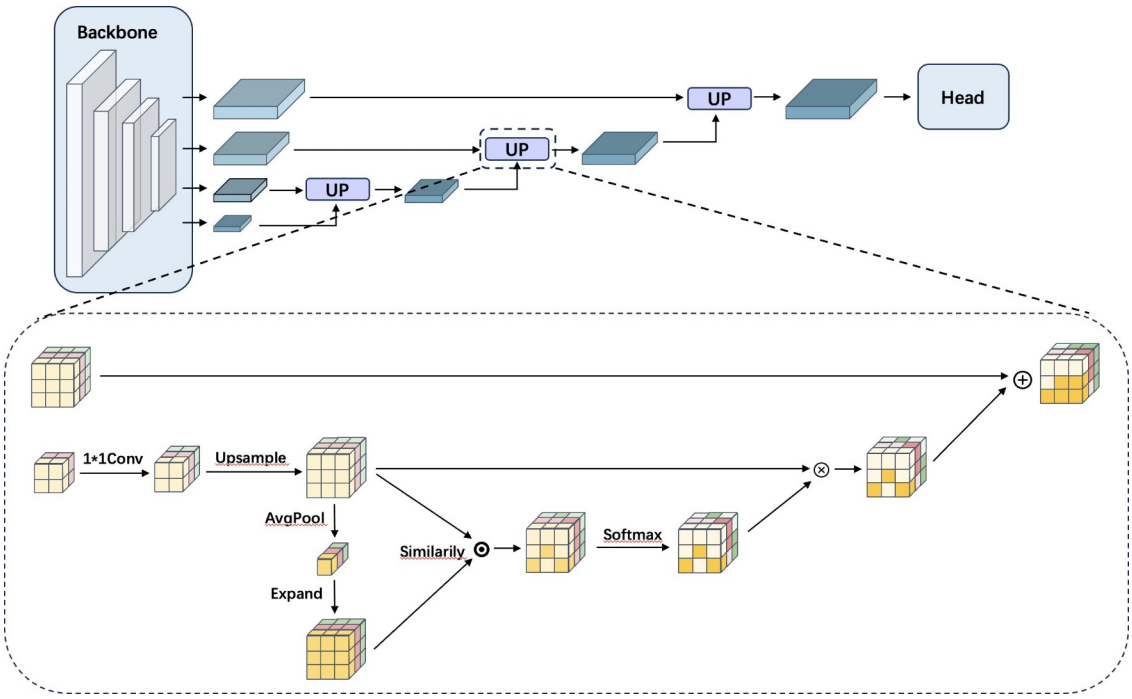

Fig. 7: Attention-based feature fusion network architecture [58].

D. Algorithm

In order to accurately capture and retain key feature information during the up-sampling process, we implement a series of well-designed optimization strategies. We first dimensionally extend the feature vector x_t of the deep network layer t by 1x1 convolution, which not only significantly improves the feature representation, but also lays the foundation for effective alignment with the features of the previous network layer t-1 by increasing the number of feature channels. Next, we use the incoming bilinear upsampling technique, which restores the features to their original spatial dimensions while effectively reducing artifacts and blurring using advanced interpolation algorithms, thus ensuring high quality and fidelity of feature information in the delivery process.

$$x_{expand} = f_{1*1Conv}(x_t) \tag{1}$$

$$x_{up} = up(x_{expand}) \tag{2}$$

To further improve the performance of the model in the up-sampling phase, we introduce an attention mechanism to correct possible misinformation and enhance the representation of features. Specifically,

we

first apply an average pooling operation to the up-sampled features x_{up} to identify and extract the critical features among them. This step helps us focus on those features that are critical to the performance of the model.

Subsequently, we dimensionally expand these key features x_{avg} to restore their spatial dimensions to what they were before average pooling, ensuring the spatial consistency of the features. To further enhance the model response to these important features, we compute the similarity between x_{up} and x_{avg}, a process that not only highlights the important features in x_{up}, but also effectively suppresses those unimportant features.

We assigned each feature a probability coefficient reflecting its importance by applying the softmax function. This coefficient not only guides the model to pay more attention to those features that have the greatest impact on the prediction results, but also enhances the model's ability to adapt to complex scenarios and generalize by dynamically adjusting the feature weights.

$$x_{avg} = avgpool(x_{up}) \tag{3}$$

$$x_{similarily} = softmax(similarily(x_{up}, x_{avg})) \tag{4}$$

After successfully identifying and assigning feature importance coefficients, we take the crucial step of integrating this information to fix potential errors in the up-sampling process. Specifically, we multiply the computed feature similarity coefficients $x_{similarity}$ with the up-sampled features x_{up} on an element-by-element basis. This operation not only directly utilizes the importance assessment of the features, but also effectively suppresses those features that contribute less to the final result through weight adjustment, while enhancing the influence of those critical features, so that erroneous information in the up-sampling can be repaired.

$$x_t = x_{similarily} * x_{up} \tag{5}$$

With this multiplication operation, we are actually performing a kind of adaptive feature filtering and enhancement, which enables the model to focus more on those information that are more important for a specific task. This strategy not only improves the efficiency of feature utilization, but also further improves the prediction accuracy and robustness of the model by reducing the interference of noise and irrelevant features.

In addition, this restoration method based on the attention mechanism provides the model with an ability to dynamically adjust the processing strategy of features flexibly according to different task requirements and data characteristics. This flexibility and adaptability are extremely valuable features in the design of modern deep learning models, which help the models achieve better performance in diverse application scenarios.

In summary, through this approach of comprehensively considering the importance of features and performing adaptive repair, we not only optimize the feature processing in the up-sampling process, but also lay a solid foundation for the overall performance improvement of the model.

E. Data Preparation

In this study, we intend to utilize a series of well-recognized, diverse and complex publicly available semantic segmentation datasets in order to train and validate our designed multi-layer feature fusion module. These datasets have been selected because they are widely known and recognized in the academic community.

1) Cityscapes: The Cityscapes dataset is a resource crafted specifically for the semantic understanding of urban streetscapes, bringing together thousands of high-resolution urban streetscape images [59]. Not only is the number of these images huge, but each one has been labeled at a fine pixel level to ensure the high quality and accuracy of the data . A wide range of categories are covered, from buildings, vehicles, and pedestrians to traffic signs and natural landscapes, with as many as 30 different categories, as shown in Fig. 7. The Cityscapes dataset carefully includes approximately 5,000 finely labeled images, which are subdivided into 2,975 training images, 500 validation images, and 1,525 test images, each with a high resolution of 1024x2048 pixels.

Fig. 8: Some samples of Cityscapes training, validation and testing [59]

2) ADE20K: The ADE20K dataset is a large scene parsing repository containing more than 20,000 images, with 20,210 images for the training set, 2,000 images for the validation set, and 3,000 images planned to be released for the test set. It covers 150 finely labeled scene categories, covering a wide range

of visual content from indoor to outdoor, from natural landscape to urban environment [60]. It provides rich visual content and data support for the research and development of scene parsing technology.

Fig. 9: Some samples of the ADE20K dataset. [60]

3) PASCAL VOC: The PASCAL VOC dataset, which occupies a pivotal position in computer vision research, is a widely adopted standardized dataset. The dataset carefully defines 20 object categories covering humans, animals (e.g., birds, felines, cows, dogs, horses, sheep, etc.), transportation (e.g., airplanes, bicycles, boats, buses, cars, motorcycles, and trains), and indoor objects (including bottles, chairs, dining tables, plants, sofas, and TVs/displays) [58]. As shown in Fig. 10, these categories are not only rich and diverse, but also highly representative. The PASCAL VOC dataset contains approximately 5,000 training images and 5,000 test images, each with a resolution of 500x500 pixels, providing researchers with a high-quality resource to support research and development in a variety of visual recognition tasks.

F. Evaluation Metrics

Intersection-to-Union (IoU) and Mean Intersection-to-Union (MIoU) are key metrics for measuring the performance of semantic segmentation models. These metrics provide an accurate way to evaluate the model's recognition ability on different categories by measuring the degree of overlap between the predicted segmented regions and the actual labeled regions. They not only provide a comprehensive picture of the overall performance of the model, but are also particularly suitable for segmentation tasks involving multiple categories. In addition, IoU and MIoU provide guidance for model training and optimization, enabling researchers to identify and improve the model for underperformance on specific

categories.

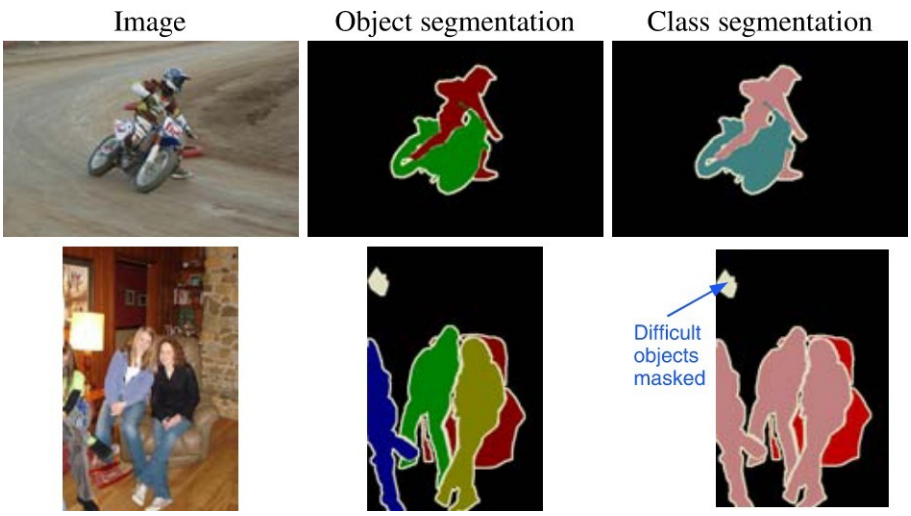

Fig. 10: Images of some PASCAL VOC samples compared to real labels [58].

1) IoU: The intersection and union ratio (IoU) is an important metric to quantify the consistency of the predicted image segmentation region with the actual labeled region. It is widely used for performance evaluation in the field of image segmentation.IoU is calculated as follows where A is the predicted segmentation region, B is the actual segmentation region, $A \cup B$ denotes the intersection of A and B, and $A \cup B$ denotes their concatenation.The value of IoU ranges from 0 to 1, where 1 denotes the perfect segmentation, i.e., the predicted region is completely overlapped with the actual region.The closer the IoU is to 1, the better is the segmentation effect

$$IoU = \frac{\text{Area } (A \cap B)}{\text{Area } (A \cup B)} \tag{6}$$

2) MIoU: When performing semantic segmentation of multiple categories, an intersection and merger ratio (IoU) value can be computed for each individual category. The Mean IoU (MIoU) is the arithmetic average of these IoU values, which provides a comprehensive assessment of the overall segmentation effectiveness of the model. This approach allows us to compare the performance of different models or techniques on semantic segmentation tasks.MIoU is calculated as follows

$$MIoU = \frac{1}{N} \sum_{i=1}^{N} IoU_i \tag{7}$$

G. Resources

1) Hardware Resources: In this study, we deployed a set of high-performance hardware resources, including a server equipped with an Intel Core i7-11700 processor (2.50 GHz main frequency), 128 GB of RAM, and two NVIDIA RTX 3090 GPUs. These resources provide powerful computing power for the experiments and ensure the smooth progress of the study.

2) Software Resources: In this study, we used the following development tools to improve coding efficiency and ease of remote collaboration:

1) Integrated Development Environment (IDE): we chose Visual Studio Code (VS Code), version 1.92. Known for its powerful code editing, smart hints, and rich ecosystem of plug-ins, VS Code is our preferred tool for writing and debugging code.
2) Terminal Management Software: We used Termius, version 9.2.0, as our terminal software for connecting to the server remotely and managing maintenance tasks. Termius greatly facilitates our remote operations with its intuitive user interface and efficient connection performance.

The selection of these tools ensured efficient code development and smooth remote collaboration during our research.

IV. RESULTS

In the experimental part of this study, we adopt widely recognized datasets such as Cityscapes [59], ADE20K [60], and PASCAL VOC [58] for model training, which ensures a rich set of image scenes and object classes to provide a solid training foundation for the model. The experimental framework is based on the PyTorch [61] deep learning library, supplemented with mmsegmentation and timm tools to enhance the flexibility and efficiency of the experiments. CrossEntropyLoss was chosen for the loss function and AdamW [62] was used for the optimization algorithm, the combination of which demonstrated stability and efficiency on large-scale datasets. Our learning rate is optimized using a polynomial decay strategy to improve the convergence speed and final performance of the network. The training process was set to 160,000/80,000/40,000 steps to achieve full learning and convergence of the model. In addition, we carefully selected cutting-edge backbone network structures such as ResNet [20], HRNetV2 [10], and Swin Transformer [49], which are known for their superior feature extraction capabilities. On top of that, we further integrate deeplab v3 [16] , UperNet [29] and OChead [46] as the head of the model. These architectures are designed for image segmentation tasks, which effectively enhance the model's performance in terms of accuracy and efficiency. All our experiments were performed with 2*RTX 3090.

A. Module design

In order to construct the optimal network architecture, we have carefully designed three multi-layer feature fusion attention mechanism modules based on different pooling strategies, aiming to enhance the model's representational ability and segmentation accuracy through diverse feature integration methods. The following are the details of the three modules we designed:

1) Average Pooling (AvgPool) Attention Mechanism Module: this module utilizes the average pooling operation to aggregate the responses on the feature map, by which it is able to efficiently capture the global contextual information in the image. Average pooling acts as a smoothing operation that helps to reduce the variance of the features, making the model more robust to noise and small variations.

2) Max Pooling (MaxPool) Attention Mechanism Module: unlike Average Pooling, the Maximum Pooling module focuses on extracting the most salient features in the feature map. This strategy allows the model to focus on the most discriminative features, thus maintaining higher sensitivity and recognition accuracy when dealing with image regions with significant visual differences.

3) Fusion Pooling (AvgPool + MaxPool) Attention Mechanism Module: this module is the highlight of our design, which combines the advantages of average pooling and maximum pooling. With this fusion

strategy, the model is not only able to acquire global contextual information, but also able to highlight

local key features. This two-pronged approach is expected to enhance the model's responsiveness to important visual cues while maintaining the comprehensiveness of feature information.

In designing these modules, we also considered how they can be integrated in different network architectures and how they can work with existing convolutional, activation, and normalization layers to achieve optimal network performance. In addition, we have carefully tuned the parameters of these modules to ensure that they maximize the expressive power of the model while adding the least computational burden. In order to evaluate the performance of these modules in real-world applications, we chose to perform rigorous testing on the cityscapes dataset. During the testing process, we adopted a training size of 512x512 pixels, set 40,000 training steps, used a fixed learning rate of 1e-2, and introduced the Online Hard Example Mining (OHEM) [63] strategy to improve the training efficiency and model performance.

OHEM is a method that dynamically selects hard-to-classify samples for training. method, which helps the model converge faster and improves its generalization ability.

After intensive training and careful comparison of these modules, our results have been exhaustively summarized in Table 1. After exhaustive data analysis, we found that the Attention Mechanism module with the average pooling strategy excels in all performance metrics. Based on these findings, we decided to choose this design as the final version of our model. This design not only demonstrates significant advantages in feature fusion, but also shows higher accuracy and robustness when dealing with complex scenes and diverse objects. In addition, it significantly improves the model's ability to capture critical information and enhances its sensitivity to details, resulting in more accurate and reliable results in a variety of visual recognition tasks. With this decision, we are confident that our model can better adapt to different application requirements and provide users with even better performance.

Model	**OCNet**	**OCNet + avg**	**OCNet + max**	**OCNet + avg + max**
val result(%)	76.88	**79.33**	76.73	78.34

TABLE I: Comparison of OCNet and its variants with different attention mechanisms.

B. Generalizability Test

In order to comprehensively evaluate the generalization capabilities of our model, we took a systematic approach and conducted in-depth tests on three challenging semantic segmentation datasets: the ADE20K, Cityscapes, and Pascal VOC 2012. these datasets are considered to be important benchmarks

for measuring

the performance of a model due to their diverse scenarios, complex backgrounds, and rich categories. Our goal is to not only validate the competitiveness of our model by comparing it with state-of-the-art methods in the field, but also to highlight its unique strengths and innovative points among similar techniques.

For the Cityscapes dataset, we chose different initialized learning rates based on the complexity of the model and previous experience. For example, for the Swin Transformer model, we used an initialized learning rate of 6e-5, while for DeepLab V3, we used an initialized learning rate of 0.01. All models were optimized for learning rate scheduling using a polynomial decay strategy, with a view to achieving more stable performance gains during training. The images were uniformly cropped to a resolution of 512*1024, the batch size was flexibly set to 4/8/16 depending on the model and hardware resources, and the number of iterations was uniformly set to 160k to ensure that the models could fully learn the features of the dataset.

For the ADE20K dataset, we adopt a similar strategy to Cityscapes, choosing different initialized learning rates according to the model features and adopting a polynomial decay strategy. The images are cropped to a resolution of 512*512, and the settings of batch size and number of iterations are kept consistent with Cityscapes, which are designed to ensure the model's ability to generalize and capture details in complex scenes.

For the Pascal VOC 2012 dataset, although its difficulty is relatively low, we still adopt rigorous experimental settings to ensure that the model can also show excellent performance on simpler datasets. The initialized learning rate and optimization strategy are the same as the previous two datasets, the images are cropped to a resolution of 512*512, and the batch size and number of iterations are also set identically.

In addition, we performed exhaustive tuning and optimization of different components and hyperparameters of the model, including but not limited to learning rate, batch size, resolution, etc., to explore the impact of various configurations on the model performance. We also recorded the changes in loss function values, accuracy, and other key metrics during the training process to facilitate the analysis of the model's learning ability and convergence behavior.

Through these comprehensive experiments and analyses, we expect to gain a comprehensive understanding of the model's performance on different datasets and make fair comparisons with other state-of-the-art methods in the field. We believe that through these rigorous tests and evaluations, our model will demonstrate its superior performance and generalization ability on semantic segmentation tasks, bringing new value and insights to the field of computer vision.

1) cityscapes: To ensure that our performance comparisons are both fair and comprehensive, we used a uniform and consistent training and testing protocol to evaluate the performance of our models against other state-of-the-art models. We selected OCNet and HRNetV2-W48 as benchmark models, both of which are built on top of the pre-trained ResNet101 and HRNetV2 architectures on the ImageNet dataset. This choice was made not only because of their state-of-the-art within the semantic segmentation domain, but also because of their innovations in multi-scale feature fusion and representation learning.

In Table 2, we show the detailed comparison results. Our replicated OCNet and HRNetV2-W48 achieve an accuracy of 76.88% and 81.10%, respectively, a result that is highly consistent with those reported in the existing literature, validating the reliability of our experimental setup. On this basis, our model achieves an accuracy of 79.33% and 82.30%, respectively, by employing an innovative multi-layer feature fusion attention mechanism. Compared to OCNet, our model improves by 2.45 percentage points, showing significant performance improvement. Compared to HRNetV2-W48, we also achieve an accuracy improvement of 1.20 percentage points, further demonstrating the effectiveness of our model in processing high-resolution images.

Model	**Backbone**	**Head**	**mIoU (%)**
OCNet	resnet101	OCHead	76.88
OCNet	resnet101	OCHead + avg	**79.33 (+2.45)**
HRNetV2-W48	HRNetV2		81.1
HRNetV2-W48	HRNetV2	avg	**82.3 (+1.2)**

TABLE II: Model Performance Comparison

In order to demonstrate more clearly the model's ability to generalize over different sizes of backbone, we conducted in-depth testing and analysis of the Swin Transformer and ResNet families. In these experiments, we adopted a unified HEAD structure and applied it to the UperNet and DeepLab v3 frameworks, further integrating our innovative modules. This consistent approach ensures that our comparisons are fair and meaningful, while also allowing us to accurately assess the specific contribution of our modules to the performance of different backbone networks.

Table 3 shows the results of these comparisons. On the mIoU metric, our replicated Swin Transformer series performs as follows: swin_tiny, swin_small, and swin_base achieve 78.69%, 81.22%, and 81.87% accuracy, respectively. Similarly, the replicated ResNet series - resnet18, resnet50, and resnet101 exhibit 68.1%, 70.54%, and 71.3% accuracy, respectively. When our modules are integrated into these models,

the performance is significantly improved: in the Swin Transformer series, the accuracies of swin_tiny, swin_small, and swin_base are improved to 79.05%, 81.99%, and 82.09%, respectively; in the ResNet series, resnet18, resnet50 and resnet101 have improved their accuracy to 70.64%, 71.21% and 74.13%, respectively. These results clearly demonstrate that our module is able to significantly improve the performance of the model regardless of the size of the backbone, proving its excellent generalization ability and adaptability.

Model	Backbone	Head	mIoU (%)	mIoU (ms+flip) (%)
Swin Transformer	swin_tiny	upernet	78.69	79.92
Swin Transformer	swin_tiny	upernet + avg	**79.05 (+0.36)**	**80.19 (+0.27)**
Swin Transformer	swin_small	upernet	81.22	82.54
Swin Transformer	swin_small	upernet + avg	**81.99 (+0.77)**	**83.10 (+0.56)**
Swin Transformer	swin_base	upernet	81.87	83.19
Swin Transformer	swin_base	upernet + avg	**82.09 (+0.22)**	**83.31 (+0.12)**
ResNet	resnet18	deeplab v3	68.10	69.72
ResNet	resnet18	deeplab v3 + avg	**70.64 (+2.54)**	**72.60 (+2.88)**
ResNet	resnet50	deeplab v3	70.54	71.63
ResNet	resnet50	deeplab v3 + avg	**71.21 (+0.67)**	**71.76 (+0.13)**
ResNet	resnet101	deeplab v3	71.30	73.03
ResNet	resnet101	deeplab v3 + avg	**74.13 (+2.83)**	**75.28 (+2.25)**

TABLE III: Model Performance in cityscapes with and without Multi-Scale and Flip Testing

In order to further improve the test performance, we adopt some advanced technical strategies during the evaluation process. Specifically, we implemented a multi-scale test and an image flip test to enhance the generalization ability of the models and verify their performance under different conditions. In the multi-scale test, we resized the input images with the following scaling: 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, and 2.0. This approach allows the model to learn and predict at multiple scales to better adapt to inputs of different resolutions and to ensure that the model is able to deal with a variety of image sizes without loss of accuracy.

An image flip test was also performed to test the robustness of the model to spatial variations by flipping the image horizontally or vertically. This test helps to assess whether the model is able to accurately recognize and process mirrored or rotated objects, thus verifying the stability of the model in the face of real-world changes. In Table 4, we show the positive impact of these testing techniques on the model performance. The results show that even after these additional testing challenges, our model

still exhibits

a significant performance improvement, which proves its stability and reliability under different testing conditions.

2) ADE20K: On the ADE20K dataset, we used consistent test configurations to ensure the reliability and comparability of the experimental results. In Table 4, we present results showing that swin_small gets a significant performance improvement with 49.49% mIoU after integrating our module. This result not only outperforms the 48.13% of swin_base without integrating our module, but is also very close to the performance after employing the multi-scale and flip test (ms+flip), which has an mIoU of 49.72%. This finding suggests that the segmentation performance of the model can be significantly improved by well-designed modules even without adding additional computational complexity.

Model	Backbone	Head	mIoU (%)	mIoU (ms+flip) (%)	Params (M)	GFLOPs
Swin Transformer	*swin_tiny*	upernet	44.51	45.58	59.94	236.15
Swin Transformer	*swin_tiny*	upernet + avg	**45.73 (+1.22)**	**46.96 (+1.38)**	59.94	236.15+0.08
Swin Transformer	*swin_small*	upernet	47.64	49.47	81.26	259.33
Swin Transformer	*swin_small*	upernet + avg	**49.49 (+1.85)**	**50.96 (+1.49)**	81.26	259.33+0.08
Swin Transformer	*swin_base*	upernet	48.13	49.72	121.28	297.28
Swin Transformer	*swin_base*	upernet+opm	**49.55 (+1.42)**	**51.61 (+1.89)**	121.28	297.28+0.06
ResNet	resnet50	deeplab v3	39.29	40.54	68.21	270.25
ResNet	resnet50	deeplab v3 + avg	**39.91 (+0.62)**	**41.33 (+0.79)**	68.21	270.25+0.06
ResNet	resnet101	deeplab v3	39.49	40.98	87.21	348.15
ResNet	resnet101	deeplab v3 + avg	**40.33 (+0.84)**	**41.76 (+0.86)**	87.21	348.15+0.06

TABLE IV: Model Performance in ADE20K with and without Multi-Scale and Flip Testing

It is worth noting that although the number of covariates for swin_small is only 81M, which is much lower than the 121M of swin base, its performance demonstrates a level comparable to that of much larger models. This result highlights the significant effect of our module in improving model performance and demonstrates superiority in parameter efficiency. This is particularly important for application scenarios where there are constraints on model size and computational resources.

In the design of our module, we focus on the balance between parameter and computational efficiency. Our module does not introduce additional parameters, which ensures the compactness of the model. Even when dealing with images of 512x512 pixels, the module adds only 0.08 GFLOPs to the computation of the Swin family of models, which is almost negligible compared to 259.33 GFLOPs for swin small

and 297.28 GFLOPs for swin_base. Similarly, in the comparison with the ResNet family of models, the increase in computation is only 0.06 GFLOPs, which has a negligible impact compared to 270.25 GFLOPs for resnet50 and 348.15 GFLOPs for resnet101. These data confirm that our module significantly improves model performance while maintaining high computational efficiency.

3) Pascal VOC2012: On the Pascal VOC 2012 dataset, we adopted a uniform test configuration to ensure consistency and comparability of the evaluation results, and thus accurately assess the performance of different models. By carefully analyzing the data in Table 5, we observe that on smaller scale models, such as the tiny version of Swin Transformer (swin_tiny), the smaller version (swin_small), and ResNet-50, performance degradation occurs under this test configuration. This contrasts with our test results for the larger model, which did not seem to be affected by this configuration and showed more consistent performance.

Model	Backbone	Head	mIoU	mIoU (ms+flip)
Swin Transformer	swin_tiny	upernet	76.06	76.68
Swin Transformer	swin_tiny	upernet + avg	75.5(-0.56)	74.17(-2.51)
Swin Transformer	swin_small	upernet	81.24	80.76
Swin Transformer	swin_small	upernet + avg	80.72(-0.52)	81.88(-0.73)
Swin Transformer	swin_base	upernet	80.76	82.61
Swin Transformer	swin_base	upernet + avg	**81.01(+0.57)**	**82.65(+2)**
ResNet	resnet50	deeplab v3	70.39	71.99
ResNet	resnet50	deeplab v3 + avg	68.92(-1.47)	71.26(-0.73)
ResNet	resnet101	deeplab v3	66.45	68.82
ResNet	resnet101	deeplab v3 + avg	**66.81(+0.36)**	**69.16(+0.34)**

TABLE V: Model Performance in Pascal VOC2012 with and without Multi-Scale and Flip Testing

This phenomenon may indicate that our current test configuration is not entirely suitable for the Pascal VOC 2012 dataset, or that there may be some inappropriateness for small models. This mismatch may have led to instability and performance fluctuations during model training, especially when confronted with the category and image complexity of the Pascal VOC 2012 dataset. Small-scale models may struggle to capture all features in the dataset due to parameter limitations, which becomes particularly evident with uniform test configurations.

V. ANALYZE

A. Visualization

In order to visually demonstrate the significant improvement in model performance by our module, we will carry out a comparative analysis of model outputs. We choose OCNet as the benchmark model because of its proven performance in the field of image segmentation. Based on this, we integrate our module into OCNet to create an improved version of the model.

We will ensure that both models are trained under the same training configuration to ensure a fair comparison. We will select the model that performs best under these conditions for the final comparison. Next, we will conduct experiments using the Cityscapes dataset, which is a challenging dataset containing rich city street scenes and is well suited for evaluating the performance of image segmentation algorithms. In our experiments, we will input images from the Cityscapes dataset into both models and collect the segmentation maps they generate. These segmentation maps will serve as the basis for our comparative analysis, allowing us to directly observe and evaluate the specific contribution of our modules to the performance of the models. With this approach, we expect to be able to clearly demonstrate the important role of our module in improving segmentation accuracy and overall model performance.

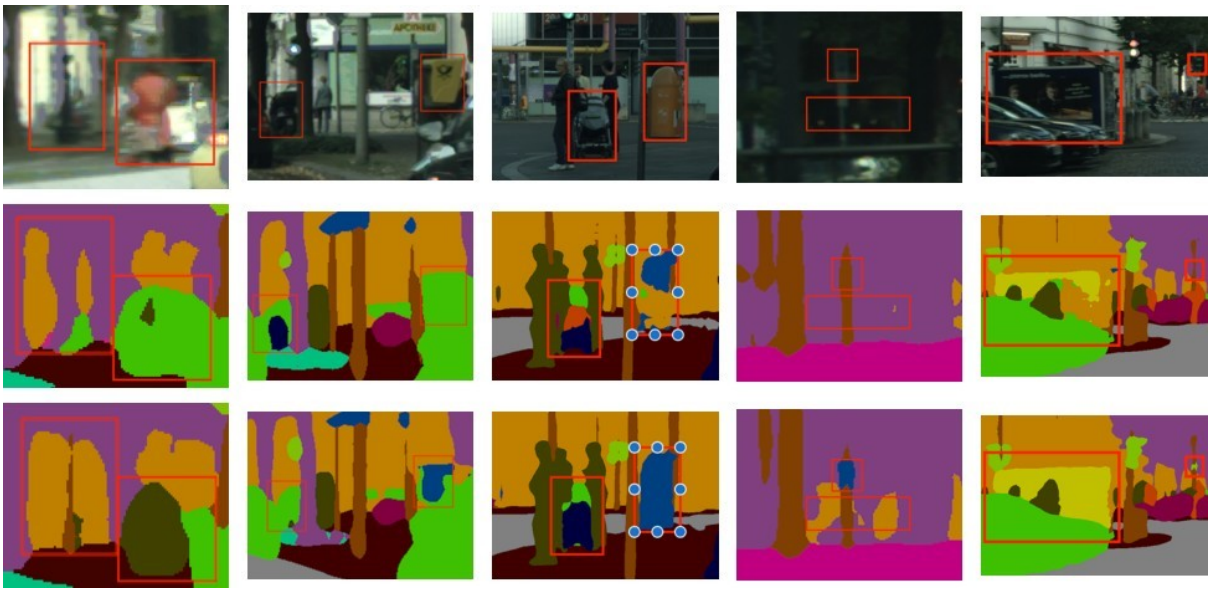

Fig. 11: Comparison of model output results

As shown in Fig. 11, the first row shows the original real image, the second row immediately after

presents the prediction results of the baseline model, while the third column shows the predicted output of our attention mechanism model based on average pooling (avgpool). The red boxes highlight where our model performs better than the original model, and where our module performs better in some details. For example, the motorcycle on the right side of the first image, the original model barely sees the motorcycle while our model accurately segments the motorcycle, and the utility pole on the left side of the first image, the original model does not recognize the utility pole at all while our model recognizes the utility pole. The motorcycle on the left of the second image, the original model incorrectly recognizes the model car as a baby carriage, while our model completely recognizes it as a motorcycle and does not misidentify it. For the mailbox on the right side of the third image, we found that the original model only recognized the top half of the mailbox, while our model recognized the entire mailbox completely. For the signage in the fourth picture, the original model only recognized the pole and ignored the signage on the pole, while our model recognized both the pole and the signage. In the fifth picture of the billboard, the original model did not recognize the whole billboard completely and some noises appeared in the background, while our model recognized the billboard completely and without any noises. Through comparative analysis, we can clearly observe that our model is on par with the baseline model in terms of segmentation of the overall background, while showing higher accuracy at the detail level.

Our improvements not only confirm the absence of negative impact of the Attention Mechanism module on model performance, but also significantly enhance the ability to capture image details. In particular, our model shows excellent performance when dealing with small objects, which are often overlooked or misjudged in conventional models. With a multi-layer fusion strategy through the attention mechanism, our model shows high sensitivity to details in complex scenes, which brings significant performance gains for semantic segmentation tasks.

B. Training Process

In order to get a full picture of the experimental process and to assess the impact of our modules, we have thoroughly documented and visualized the key performance metrics in our experiments. These metrics include mean Intersection over Union (mIoU), mean Accuracy (mAcc) and mean Accuracy (aAcc). These metrics allow us to quantify the performance of the model on image segmentation tasks. In addition, we specifically compare the model that integrates the averaging (avg) module with the model that does not integrate this module. This comparison not only demonstrates the contribution of the avg module in improving the performance of the model, but also reveals its potential advantages in the training process. Through this comparative analysis, we are able to gain a deeper understanding of

the importance of the avg module in improving model stability and accuracy.

In this experiment, we constructed a baseline model using ResNet-50 as the backbone network (backbone) for feature extraction and incorporating DeepLab as the head network (head) for segmentation task. We chose the Cityscapes dataset as a testbed, which is a high-resolution image dataset containing complex urban street scenes widely used to evaluate the performance of semantic segmentation algorithms.

Our training strategy is to perform a total of 80,000 iterations of the training process. To ensure the generalization ability of the model, we perform validation every 8000 iterations. During the validation process, we strictly ensure that the validation dataset is completely independent from the training dataset to avoid overfitting phenomenon and ensure the reliability and validity of the validation results.

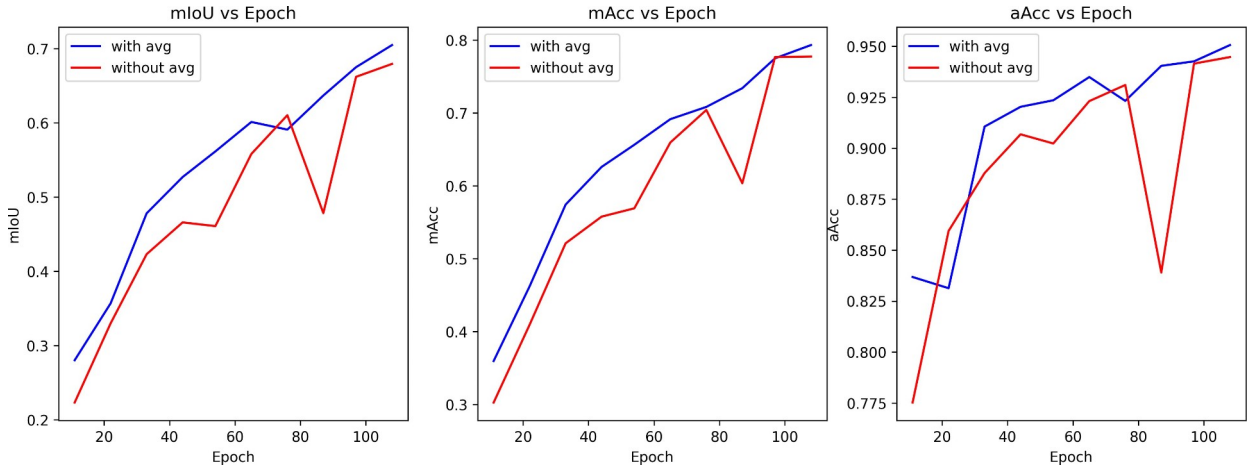

Fig. 12: Comparison of training processes

The three subgraphs, as presented in Fig. 12, correspond to three key metrics of model performance: average intersection and merger ratio, average precision, and average accuracy. In these graphs, the blue curve represents the performance of the model with the averaging (avg) module integrated, while the red curve represents the performance of the baseline model.

Looking at the graphs it is clear that the blue curve lies above the red curve for all metrics, indicating a significant performance advantage for the model that introduces the avg module. The height of the curve directly correlates to the model's performance on the corresponding metric; the higher the curve, the better the model's performance on that metric.

Further analyzing the graphs, we notice that the red curve shows large fluctuations in the later stages of training, which may indicate that the baseline model encountered overfitting or underfitting problems

during training, resulting in unstable performance. In contrast, the blue curve maintains a high level

of stability and consistency, which not only confirms the effectiveness of the avg module in improving performance, but also demonstrates its role in improving the stability of model training.

C. Quantitative analytics

In order to fully evaluate the impact of the avg module, we not only focused on the model performance improvement, but also deeply analyzed the number of covariates and the computational complexity of the model. Our goal is to ensure that the avg module improves performance without negatively impacting the scalability and efficiency of the model.

We conducted an exhaustive comparative analysis of the performance of the Swin Transformer family of models at 512x512 resolution, aiming to explore the impact of different model configurations on the final performance. As shown in Figure 13, the left-hand chart demonstrates the standard mean Intersection over Union (mIoU) performance, a key metric for measuring the accuracy of semantic segmentation models. The right-hand charts, on the other hand, demonstrate the enhanced mIoU performance, i.e., the application of multi-scale and flip enhancement strategies after model prediction, which further improves the robustness and accuracy of the model.

In these graphs, the X-axis represents the amount of computation in terms of FLOPs (floating-point operations times), reflecting the computational resources required by the model for forward propagation, and the Y-axis represents the accuracy, i.e., the percentage of mIoU, which is directly correlated to the model's prediction performance. The size of each circle, on the other hand, visualizes the number of parameters of the model, with larger circles indicating that the model has more parameters. This visual presentation allows us to quickly identify the trade-off between the number of parameters, computation and accuracy for different models.

The analysis results show that the models after the introduction of the avg module all outperform the original model in terms of accuracy, which indicates that the avg module plays a positive role in improving performance. At the same time, these improved versions of the models also show a good balance in terms of maintaining the number of parameters and computational effort, indicating that they successfully improve the segmentation accuracy without significantly increasing the model complexity. This optimization is particularly important for application scenarios that require deployment of models on resource-constrained devices, such as mobile devices, embedded systems, etc.

In addition, we also observe that as the number of model parameters increases, the accuracy improvement shows a marginal decreasing trend. This suggests that we need to consider a reasonable allocation of the number of parameters in model design to achieve an optimal performance-to-resource

ratio. Further analyses also consider the generalization ability of different models on different datasets and different classes, as well as their robustness to factors such as occlusion, light variations, and background interference.n.

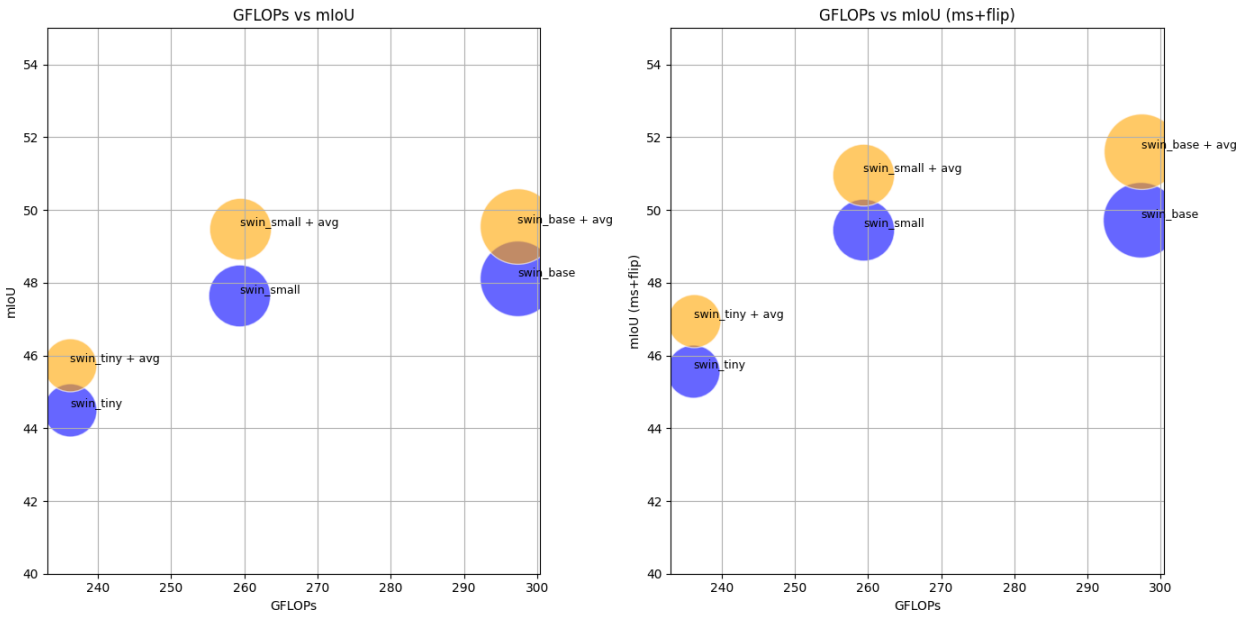

Fig. 13: Comparison of parametric quantities and operations of different models

We present an exhaustive comparative analysis of the computational requirements of two different families of models, Swin Transformer and ResNet, at different resolutions. Focusing on the performance evaluation from 480×480 to 2048×2048 resolutions, we aim to explore the impact of resolution increase on the computational efficiency of the models. As shown in Fig. 14, we systematically compare the additional computation of the model at each resolution to evaluate the performance of the model at different input sizes.

The results show that across all tested resolutions, our module maintains a low computational effort even in the face of a significant resolution increase, thanks to its efficiently designed algorithms. This finding not only confirms the efficiency of the module, but also highlights its clear advantage in reducing resource consumption, especially in high-resolution image processing. This has important practical implications for application scenarios where large-scale images need to be processed.

We find that although high-resolution images bring more detailed information, they also increase the computational complexity. Our module effectively balances this challenge through a well-designed network structure and parameter configuration, achieving the optimization of accuracy and computation

at high resolution.

To summarize, our comparative analysis reveals the additional computation of the modules at different resolutions, showing their potential and advantages in high-resolution image processing. With the continuous development and optimization of the technology, we believe that these modules will bring more efficient and low-consumption solutions to the field of computer vision and promote the use of related technologies in a wider range of application scenarios.

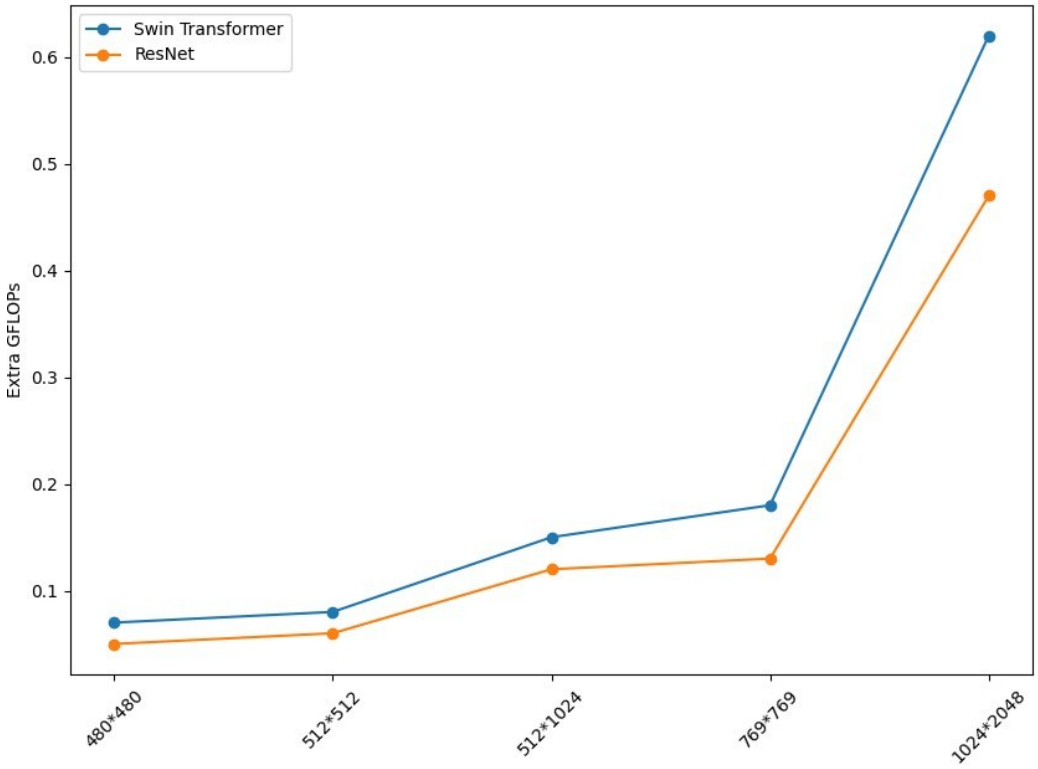

Fig. 14: Extra GFLOPs by Resolution for Different Models

In order to fully evaluate the performance of the model, we first performed a frame rate (FPS) comparison test at two different resolution settings - 512x512 and 512x1024. After completing the resolution comparison, we selected one of the resolutions and performed a further detailed performance evaluation of the model. For this evaluation, we ran 200 consecutive iterations, recording the runtime of each iteration. These data allowed us to accurately analyze the average processing speed and performance stability of the model at the selected resolution.

Fig. 15 shows a comparison of the frame rate (FPS) of the model with and without the averaging (avg) module at different resolutions (512x512 and 512x1024). The left graph corresponds to 512x512 resolution and the right graph corresponds to 512x1024 resolution. In each graph, the blue bar represents the FPS without avg module, while the orange bar represents the FPS with avg module. the height of the bar indicates the average processing speed of the model, and the higher the bar, the faster the processing speed. Through comparative analysis, we find that although the introduction of the avg module slightly reduces the average processing speed of the model, this performance loss is not significant. This suggests that the avg module has a limited impact on the inference speed of the model while improving performance.

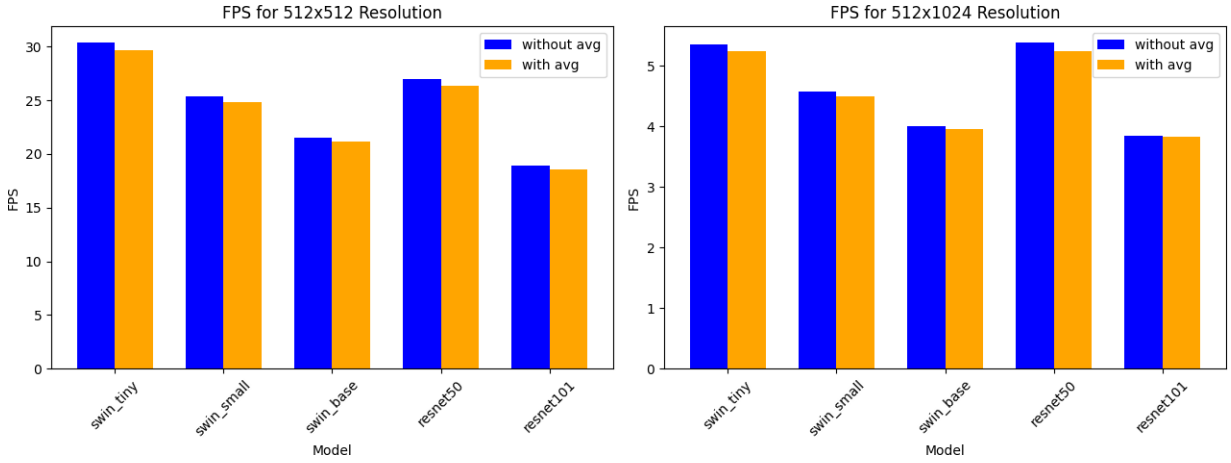

Fig. 15: FPS of different models at different resolutions

D. Analysis of segmentation results

1) Analysis of ADE20K: In our study to deeply analyze the performance of the Swin Base + Avg model on the ADE20K dataset, we pay special attention to the differences in the performance of the model on different categories. Since the ADE20K dataset contains a large number of categories, we select the top 10 categories with better performance and the bottom 10 categories with worse performance for our study, with a view to revealing the model's recognition ability on different categories.

In Table 6, we list the recognition accuracies for these categories and the performance differences between the Swin Base + Avg model and the Swin Base model. Among the better performing categories, the Swin Base + Avg model performs well on the category "bridge" with an accuracy of 77.95%, which

is 41.46% higher than the 36.49% of the Swin Base model. This indicates that the Swin Base + Avg model significantly improves the recognition accuracy by the introduction of average pooling layer when dealing with complex structures such as bridges.

Similarly, the “copy” category has an accuracy of 40.26% in the Swin Base + Avg model, while the accuracy of the Swin Base model is only 11.4%, with a difference of 28.86%. This finding further confirms the superiority of the Swin Base + Avg model in handling categories with complex textures and structures.

However, in Table 7, we observe some interesting phenomena. For example, the accuracy of the Swin Base + Avg model for the category “ship” is only 17.94%, while the accuracy of the Swin Base model is as high as 62.59%, with a difference of -44.65%. This may indicate that the Swin Base model already has a high initial performance on some specific categories, and the introduction of the average pooling layer did not bring the expected gain, but instead may have led to a decrease in performance.

Furthermore, the accuracy of the Swin Base + Avg model for the category “microwave” is 71.94%, while the accuracy of the Swin Base model is 82.31%, a difference of -10.37%. This further emphasizes that the average pooling layer may not be the best choice for improving model performance on certain categories.

class	with avg	without avg	Difference
bridge	77.95	36.49	41.46
canopy	40.26	11.4	28.86
escalator	52.41	28.55	23.86
flag	53.99	33.89	20.1
skyscraper	65.6	47.01	18.59
arcade machine	50.78	33.69	17.09
buffet	43.35	27.85	15.5
base	39.34	24.76	14.58
sand	54.89	40.65	14.24
bar	39.56	27.41	12.15

TABLE VI: The better 10 classes

class	with avg	without avg	Difference
ship	17.94	62.59	-44.65
truck	25.37	43.16	-17.79
oven	41.24	52.06	-10.82
microwave	71.94	82.31	-10.37
plaything	25.78	34.83	-9.05
barrel	53.52	61.9	-8.38
apparel	36.17	43.76	-7.59
wardrobe	45.65	51.48	-5.83
boat	42.54	48.19	-5.65
fountain	25.22	30.76	-5.54

TABLE VII: The worse 10 classes

By analyzing this data, we can conclude that the combined performance of the Swin Base + Avg model on the ADE20K dataset is positive, especially when dealing with categories with complex structures. However, the performance improvement of the model on some specific categories is not significant, and

there is even a risk of performance degradation. This finding suggests that in practical applications, we need to flexibly choose the model structure according to the characteristics of specific categories to achieve optimal performance. Future research will continue to explore more optimization strategies to further improve the performance of the model on all categories.

2) Analysis of Pascal VOC: In Table 8, when we analyze in depth the performance of the swin base + avg model on the Pascal VOC dataset, we can observe some significant trends. The model shows excellent performance in handling some categories, while it leaves something to be desired in some other categories. Overall there are more categories that perform better than those that do not. The following is an expansion of this paragraph based on the contents of the table:

By analyzing the Pascal VOC dataset in detail, we can see that the swin base + avg model achieves higher accuracy than the swin base model in several categories. For example, when recognizing the categories “airplane” and “background”, the model achieves an accuracy of 94.10% and 95.79%, respectively, which indicates that the model has a very high recognition ability in these scenes. In addition, the accuracy for the “person” category is also relatively high at 89.57%, which may be attributed to the model’s sensitivity to the complex features of the human form.

However, the model did not perform as well on some categories. For example, the accuracy of the “cat” (cat) category dropped from 90.44% to 88.14%, which is a significant decrease. Similarly, a similar drop was observed for the “dog” (dog) category, from 84.51% to 80.06%. This may indicate that the model is experiencing some difficulty in recognizing these animal categories, possibly due to the higher diversity and complexity of the images of these categories in the dataset.

It is worth noting that the difference in accuracy for some categories is very small, such as “bicycle” and “tvmonitor”, which suggests that the model’s performance on these categories is very close to its potential optimal performance. For categories such as “boat” and “diningtable”, although the model’s accuracy decreases, the differences are very small, -0.12% and -0.17% respectively, which may imply that the images in these categories have high feature similarity in the dataset, making it difficult for the model to differentiate between them.

Overall, the performance of the swin base + avg model on the Pascal VOC dataset shows its versatility and complexity on different categories. The high accuracy of the model on some categories proves its strong recognition ability, while the drop on other categories hints at the room for improvement of the model in these areas. With further optimization and tuning, we can expect the model to provide a more balanced and comprehensive performance in the future.

Class	swin base + avg	swin base	diff
sofa	50.87	47.74	3.13
bottle	81.98	79.72	2.26
cow	86.99	84.90	2.09
car	87.28	86.13	1.15
horse	90.22	89.17	1.05
bus	92.16	91.35	0.81
chair	34.01	33.19	0.82
sheep	90.97	90.50	0.47
train	84.82	84.44	0.38
motorbike	86.63	86.03	0.60
aeroplane	94.10	93.91	0.19
bicycle	81.15	80.96	0.19
background	95.79	95.68	0.11
person	89.57	89.49	0.08
tvmonitor	78.28	78.21	0.07
boat	79.12	79.24	-0.12
diningtable	62.95	63.12	-0.17
bird	92.22	92.59	-0.37
pottedplant	64.36	65.98	-1.62
cat	88.14	90.44	-2.30
dog	80.06	84.51	-4.45

TABLE VIII: Comparison of swin base and swin base + avg with their differences.

VI. Conclusions and Recommendations

In this study, we identify some limitations of current multilayer feature fusion techniques in the field of semantic segmentation through an in-depth literature review and experimental analysis. In particular, we note the error that is easily introduced during the feature up-sampling process, which may affect the final segmentation accuracy. The introduction of such errors not only affects the model's ability to capture details, but may also lead to a degradation of the model's generalization performance in complex scenarios. For example, in the field of autonomous driving, the model's recognition results are crucial for vehicle navigation and decision making. Incorrect prediction results not only lead to navigation errors of self-driving vehicles, but also may cause traffic accidents, resulting in personal and property losses. In addition, automatic driving systems need to be highly adaptive and robust in the face of changing road conditions and traffic environments, and current semantic segmentation technology still needs to be improved in these aspects.

In the field of medical imaging, accurate medical image segmentation is of great significance for the diagnosis and treatment of diseases. High-quality segmentation results can help doctors more accurately identify lesion areas and thus formulate more precise treatment plans. In addition, automated image segmentation techniques can also reduce the workload of doctors and improve diagnostic efficiency, especially when dealing with large-scale medical image data. However, current semantic segmentation models still face challenges when dealing with medical images with high similarity or complex backgrounds. This requires us to further optimize the feature fusion strategy and improve the model's adaptability to different types of medical images to meet the needs of clinical applications.

To address the problem of detail loss and error accumulation due to up-sampling in semantic segmentation tasks, we propose an innovative hypothesis: by introducing an attention mechanism, the error generated during up-sampling can be effectively repaired and the model's ability to focus on key features can be enhanced. Based on this hypothesis, we design and implement a novel fusion module based on the attention mechanism, aiming to enhance the performance of the model on semantic segmentation tasks. Our fusion module utilizes the adaptive property of the attention mechanism to dynamically adjust the importance weights of features at different levels. This dynamic adjustment not only optimizes the feature integration process, but also reduces the loss of information and accumulation of errors during the up-sampling process. By giving the network the ability to self-optimize feature fusion, our module is able to reconstruct high-resolution, detail-rich segmentation results more accurately.

In addition, the design of the module takes into account the balance between computational efficiency

and the number of parameters, ensuring that the introduction of the attention mechanism is not overly burdensome to the overall complexity of the model. This is particularly important in resource-constrained deployment environments, such as real-time semantic segmentation in mobile devices or embedded systems.

In order to test our hypothesis that attention mechanisms can improve the performance of multilayer feature fusion, we conducted a series of systematic experiments. The first step of the experiments was to design three different types of attention modules: the average (avg) module, the maximum (max) module, and their combination (avg+max) module. These modules were designed to explore the effects of different attentional strategies on the effectiveness of feature fusion.

We chose to integrate these modules into two state-of-the-art semantic segmentation models: the OCNet and HRNetV2-W48. These two models were chosen because they are representative of the semantic segmentation field and can provide a solid foundation for our experiments. Subsequently, we performed an initial validation on the Cityscapes dataset, a demanding dataset containing complex city streetscapes, which is well suited for evaluating model performance.

The experimental results show that the average (avg) module performs the best in terms of performance, which may be attributed to its ability to balance the contributions of different feature maps, thus reducing the error in the up-sampling process. Based on this finding, we further explore the generalization and robustness of the avg module. To this end, we apply the module to several standard public datasets, including Cityscapes, ADE20K, and Pascal VOC 2012.

The avg module shows good performance on all these diverse datasets, which not only confirms our hypotheses but also demonstrates the applicability and effectiveness of the avg module in different scenarios. These results provide strong support for our module and demonstrate its potential and value in semantic segmentation tasks.

In order to visually assess the performance improvement of our model, we performed a comparative analysis of the prediction results between the OCNet baseline model and the OCNet model integrated with our improved module. Through this comparison, we observe that our model effectively repairs the segmentation information of the objects and significantly reduces the errors introduced during the up-sampling process. This suggests that our module plays a key role in improving the quality of feature fusion and up-sampling.

In addition, we provide an in-depth analysis of several key aspects of the model, including the stability of the training process, the efficiency of the model parameters, and the speed of inference. This multi-perspective evaluation approach allows us to gain a comprehensive understanding of the model's

performance and identify potential strengths and weaknesses.

By analyzing the training process, we ensured the convergence and generalization ability of the model. The analysis of model parameters helped us understand the impact of modules on model complexity and how to optimize parameters for better performance and efficiency. The evaluation of inference speed, on the other hand, allowed us to assess the usefulness of the model in real-world applications, especially in scenarios that require real-time processing.

During these analyses, we also learned that although our model does not introduce additional parameters and has very few operations, the inference speed is indeed slower, which also provides a clear direction for us to improve in the future. We believe that through this comprehensive and careful evaluation, our model will be able to provide users with higher quality semantic segmentation results and play an important role in various application scenarios.

VII. Reflection

This was my first venture into the field of Artificial Intelligence, and it marked not only a significant milestone in my graduate career, but also a major success in my journey of academic exploration. In the process, I have immersed myself in a sea of knowledge, read a large number of cutting-edge academic papers, navigated through a multitude of exquisite code implementations, and thought deeply about a multitude of complex scientific issues.

Whenever I came across a marvelous paper, it always made me stop for a long time and think deeply about the rationale and innovation behind it. I would ask myself: why does this idea work so well? What's so amazing about it? These questions inspired me to keep exploring and moving forward.

Similarly, I am fascinated by the intelligence and creativity of the code I browse through. I think: Why is this code so elegant and efficient? What is the logic and structure behind them? This appreciation of beauty and pursuit of excellence motivated me to keep learning and imitating in order to reach a higher technical level.

Although I have accumulated a great deal of theoretical knowledge during my academic journey, I often feel confused and clueless when I try to apply this knowledge to actual code implementations. This experience made me realize deeply that there is a huge gap between theory and practice. Although the theoretical knowledge provided me with a solid foundation, I faced many challenges in translating it into practical applications. For example, I encountered a lot of problems in the process of designing my model, what should I base on as the foundation of my research, and how should I design my model, all of which I encountered during the project and which often made me think for a long time.

I gradually realized that practice is the best way to theorize. In practice, I need to constantly debug, test and optimize, which is a process full of trial and error and learning. Every failure is a re-understanding and internalization of knowledge, and every success is a consolidation and enhancement of skills.

I learned to start paying more attention to the accumulation of practical experience. I realized that only through continuous practice could I really understand the depth and breadth of theories and transform abstract concepts into concrete solutions. This is where the success of my experiments lies, I have conducted many experiments and experienced many failures, and it is on the basis of so many failed experiments that I have successfully completed the module I designed.

Therefore, I will more actively seek practical opportunities to combine theoretical knowledge with practical problems through projects, internships or individual research. I believe that with experience, I will be able to apply my knowledge and skills more confidently and skillfully to solve more complex and challenging problems.

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[2] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[4] K. He, G. Gkioxari, P. Dolla´r, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[7] Z. Cui, W. Chen, and Y. Chen, "Multi-scale convolutional neural networks for time series classification," *arXiv preprint arXiv:1603.06995*, 2016.

[8] T.-Y. Lin, P. Dolla´r, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[10] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[11] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "Ocnet: Object context for semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 8, pp. 2375–2398, 2021.

[12] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 173–190.

[13] L. Ke, Y.-W. Tai, and C.-K. Tang, "Deep occlusion-aware instance segmentation with overlapping bilayers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4019–4028.

[14] Q. Song, K. Mei, and R. Huang, "Attanet: Attention-augmented network for fast and accurate scene parsing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 2567–2575.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[16] ——, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going

deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[22] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[24] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[25] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[26] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 963–11 975.

[27] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[28] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9308–9316.

[29] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418–434.

[30] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in neural information processing systems*, vol. 34, pp. 17 864–17 875, 2021.

[31] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[32] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[33] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[34] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[35] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[38] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[39] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," *arXiv preprint arXiv:2112.05561*, 2021.

[40] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3139–3148.

[41] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[42] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[43] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *European conference on computer vision*. Springer, 2020, pp. 108–126.

[44] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 124–12 134.

[45] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[46] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "Ocnet: Object context for semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 8, pp. 2375–2398, 2021.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[49] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[50] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[51] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[52] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[53] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in neural information processing systems*, vol. 34, pp. 17 864–17 875, 2021.

[54] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[55] ——, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[56] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[57] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, “Oneformer: One transformer to rule universal image segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2989–2998.

[58] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[59] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[60] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, “Semantic understanding of scenes through the ade20k dataset,” *International Journal of Computer Vision*, vol. 127, pp. 302–321, 2019.

[61] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

[62] I. Loshchilov, F. Hutter *et al.*, “Fixing weight decay regularization in adam,” *arXiv preprint arXiv:1711.05101*, vol. 5, 2017.

[63] A. Shrivastava, A. Gupta, and R. Girshick, “Training region-based object detectors with online hard example mining,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 761–769.

APPENDIX

Below is the project timeline, and the Gantt chart is shown in Figure 11.

1) Background (02.19 - 03.17), 4 weeks were spent on collecting and analysing.
2) Proposal Development (03.18 - 05.12), 8 weeks were spent on developing the research proposal.
3) Model Selection (05.13 - 06.22), 6 weeks were spent on selecting suitable models.
4) Model Design (06.23 - 07.13), 3 weeks were spent on designing the selected model.
5) Experimentation (07.14 - 07.28), 2 weeks were spent on conducting the experiments.
6) Data Analysis (07.29 - 08.10), 2 weeks were spent analysing the experimental data.
7) Writing (08.11 - 08.25), 2 weeks were spent on writing the report.