1

Optimizing Deeplabv3+ with Multi-Scale Attention for Semantic Segmentation

Jiajing Liu and Arfat Ahmad Khan

Optimizing Deeplabv3+ with Multi-Scale Attention for Semantic Segmentation

Jiajing Liu[1] and Arfat Ahmad Khan[2]

[1]**College of Computing, Khon Kaen University, Khon Kaen 40002, Thailand**
e-mail: jiajing.l@kkumail.com
[2]**College of Computing, Khon Kaen University, Khon Kaen 40002, Thailand**
e-mail: arfatkhan@kku.ac.th

Received : October 9, 2025 **Revised :** October 31, 2025 **Accepted :** November 3, 2025

Abstract

DeepLabv3+, a leading model for semantic segmentation, often struggles with high computational costs and inadequate multi-scale representation, leading to blurred boundaries and poor detection of small-scale targets. To overcome these challenges, our work introduces an efficient network built upon the lightweight MobileNetV2 backbone that incorporates three novel modules. First, our DENS-ASPP module replaces the standard ASPP to better capture multi-scale features using a densely connected atrous cascade. These features are then refined by the SEA module, which applies spatial attention for modeling extensive, direction-sensitive contextual information. The last component of our architecture is the DCE module, which enhances the decoder with coordinate attention, embedding positional information to sharpen object details. Our model achieves 72.56% mIoU and 87.28% mPA on the PASCAL VOC 2012 dataset, demonstrating that this integrated framework yields substantial gains in segmentation performance.

Keywords: Attention; Multi-Scale; Segmentation; Efficient Network

Introduction

A foundational challenge in computer vision is the task of image segmentation, a process concerned with partitioning a digital image or

video frame into distinct, semantically meaningful regions (Ding & Qian, 2024). As a cornerstone of visual understanding systems, this capability has wide-ranging applications, impacting fields such as medical diagnostics (e.g., tumor boundary detection and tissue analysis), autonomous navigation (e.g., identifying road surfaces and pedestrians), and interactive technologies like video surveillance and augmented reality (Wang, et al., 2023; Wenkuana & Shicai, 2023). Segmentation can be approached with varying levels of detail, including semantic segmentation, where every pixel is assigned a category label, instance segmentation, and panoptic segmentation. Compared to whole-image classification—a task that assigns a single category to the entire image—segmentation is inherently more complex due to the demand for modeling fine-grained structures, precise object boundaries, and multi-scale context (Jiang et al., 2021; Lili & Jinzhi, 2022).

Over the decades, the field of image segmentation has seen a significant evolution in methodology. Early approaches were dominated by classical techniques including thresholding, histogram-based methods, region growing, and clustering algorithms like k-means and watershed (Li, et al., 2023). More sophisticated formulations later emerged, such as active contours, graph cuts, and probabilistic models like conditional/Markov random fields. In recent years, the advent of deep learning (DL) has triggered a paradigm shift, consistently delivering new state-of-the-art results on major public benchmarks and fundamentally reshaping the field's best practices (Xiang et al., 2024; Lee & Park, 2024). Among the various DL architectures, the DeepLab family has risen to prominence and become particularly influential within semantic segmentation (Azad et al., 2020). Models from the DeepLab family, for instance, effectively enhance boundary localization and detail recovery

1

by employing encoder-decoder architectures and atrous spatial pyramid pooling (ASPP) to aggregate multi-scale context.

Despite these strengths, a key limitation in the DeepLab series is that its multi-scale feature fusion strategy remains relatively simplistic, heavily relying on dilated convolutions. In visually complex or cluttered scenes, this approach can lead to suboptimal cross-scale feature interaction, causing small objects to be overlooked and boundaries to become overly smoothed. Consequently, there is a clear need to improve the representation and fusion of multi-scale information, particularly for handling objects of diverse scales and geometries in challenging contexts.

This work introduces a set of structural enhancements for DeepLabv3+ that target two complementary goals: improving spatial context modeling and refining decoder-level features, all while maintaining computational efficiency. We achieve this by developing modules that provide a more expressive multi-scale representation—enhancing spatial awareness via global and directional pooling (SEA) and embedding coordinate information into the decoder for fine-grained detail recovery (DCE).

A foundational challenge in computer vision is the task of image segmentation, which partitions digital images into semantically meaningful regions (Ding & Qian, 2024). This process plays a crucial role in applications such as medical diagnostics, autonomous driving, and video surveillance (Wang et al., 2023; Wenkuana & Shicai, 2023). Among many models, **DeepLabv3**+ has become one of the most influential frameworks for semantic segmentation due to its encoder-decoder design and use of Atrous Spatial Pyramid Pooling (ASPP) for capturing multi-scale context (Chen et al., 2017).

However, DeepLabv3+ still faces limitations—its multi-scale fusion heavily relies on dilated convolutions, which may lead to weak cross-scale feature interactions and blurred boundaries. Consequently, there is a need

to improve multi-scale representation and decoder accuracy. To address these challenges, this study proposes an optimized version of DeepLabv3+ integrating three innovative modules: DENS-ASPP, SEA, and DCE, all designed to balance feature richness, contextual understanding, and computational efficiency.

Research Objective(s)

The primary objectives of this research are as follows:

1. To design a **DENS-ASPP module** that enhances multi-scale feature extraction using a densely connected atrous cascade and strip pooling.

2. To develop a **Spatial Enhancement Attention (SEA) module** that fuses global and directional pooling for improved spatial context modeling.

3. Finally, propose the DCE (Decoder Coordinate Enhancement) module, which embeds precise positional information into the decoder using coordinate attention to significantly sharpen object boundaries.

Research Methodology

Overview of the DeepLabv3+ Network

Built upon DeepLabv3, the DeepLabv3+ architecture (Yang et al., 2020) incorporates a lightweight decoder, resulting in a classical encoder-decoder structure. In semantic segmentation, this design is widely adopted: the encoder employs a deep convolutional backbone to derive high-level semantic representations (Zhao et al., 2017). On top of the encoder, the ASPP head aggregates multi-scale context using an arrangement of five concurrent streams. Four of these are convolutional: a 1x1 convolution and three 3x3 atrous convolutions with distinct dilation factors (6, 12, and 18). The fifth stream performs image-level global average pooling to provide a holistic feature summary. Together, these branches capture

1

complementary scales and produce robust feature representations. The decoder then progressively restores spatial resolution and refines boundaries: low-level features from the backbone are first projected by a 1×1 convolution and fused with the ×4 upsampled ASPP output; a subsequent 3×3 convolution and a final ×4 upsampling step recover the original resolution and sharpen object edges.

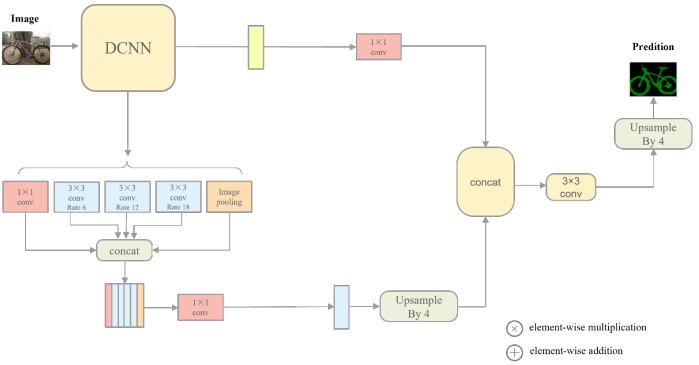

Figure 1. Deeplabv3 plus model
Source: Authors (2025).

Improved DeepLabv3 plus network

Our proposed network architecture is built upon the DeepLabv3+ framework, for which we adopt the lightweight MobileNetV2 as the primary feature extractor to ensure computational efficiency. A core innovation of our model is the replacement of the standard ASPP module with our novel DENS-ASPP, which is designed to capture a richer set of multi-scale features through a densely connected atrous cascade. To further enhance these features, a Spatial Enhancement Attention (SEA) module is positioned subsequently, generating a spatial attention map from global and strip pooling to adaptively refine the feature representation. In the decoder stage, we incorporate a Decoder Coordinate Enhancement (DCE) module. This module leverages coordinate attention to facilitate a more effective integration of high- and low-level features, which is critical for achieving

a precise definition of object contours. The synergistic interplay between these components is depicted schematically in Figure 2.

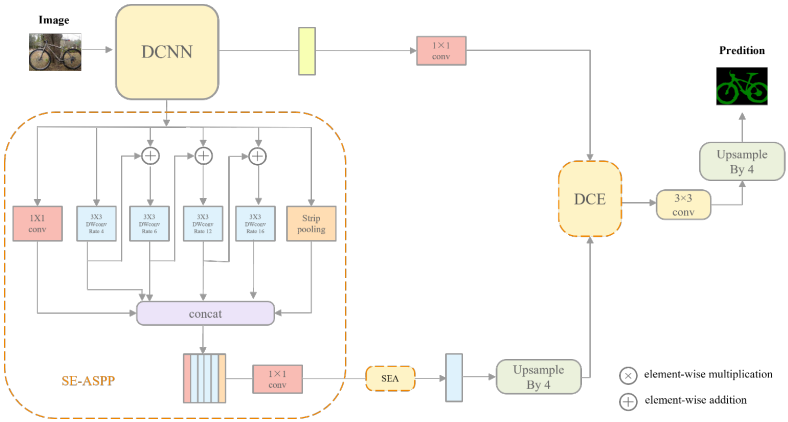

Figure 2. Improved DeepLabv3 plus model
Source: Authors (2025).

DENS-ASPP Module

To provide the network with a rich understanding of multi-scale context in an efficient manner, we propose a novel module named DENS-ASPP, which is inspired by the original DenseASPP (Yang et al., 2018). This module collaboratively extracts features from various dimensions and scales through several integrated parallel branches. At its core is a Densely Connected Atrous Convolution Cascade, which consists of a series of 3x3 depthwise separable atrous convolutions (DWconv) with varying dilation rates (e.g., 4, 6, 12, and 16). Differing from conventional parallel structures, this branch adopts a dense connection scheme: the output of a preceding convolutional layer is element-wise added to its input, and the result is fed into the subsequent layer with a larger dilation rate, thereby constructing a hierarchical feature extraction path. In parallel, a standard 1 × 1 convolutional branch and a Strip Pooling branch operate concurrently. The 1x1 convolutional branch handles the extraction of

basic features, while the Strip Pooling branch is tasked with capturing long-range contextual dependencies by aggregating features horizontally and vertically, compensating for the limitations of atrous convolutions in capturing global information. For comprehensive feature fusion, the outputs from all parallel paths—including the 1 × 1 convolution branch, the features generated by every individual atrous layer in the cascade, and the Strip Pooling branch—are concatenated. Finally, the concatenated features are processed through a terminal 1 × 1 convolution for final feature integration and channel adjustment, yielding the module's ultimate output.

SEA Module

While CNNs dominate semantic segmentation, their receptive fields grow slowly and struggle to capture global context. GAP alleviates this but collapses directional information due to its isotropic pooling. Therefore, combining GAP with anisotropic strip pooling offers a complementary solution to balance global context and directional sensitivity. Strip pooling addresses this limitation by employing anisotropic windows along horizontal and vertical axes, enabling richer global context modeling and stronger sensitivity to structural layouts(Hou et al., 2020).

While the original Atrous Spatial Pyramid Pooling (ASPP) module is designed for the aggregation of multi-scale contextual information, it lacks sufficient capability to capture spatial correlations and long-range dependencies. To overcome this limitation, we propose an enhanced module, named SEA, which integrates global pooling operations and strip pooling to refine spatial awareness and strengthen feature representations. The SEA is placed after the standard ASPP block in the DeepLabv3+ architecture, thereby enriching feature extraction for semantic segmentation.

The proposed SP module incorporates three parallel pooling branches: global average pooling (GAP), which captures the holistic

background context, and strip pooling (SP) along horizontal and vertical dimensions, which models long-range spatial dependencies. To generate the spatial attention map, the feature maps from these parallel branches are first aggregated and then passed through a 1×1 convolutional layer, with a subsequent non-linear activation function producing the final map. This map reweights the input features in an element-wise manner, thereby adaptively enhancing spatially relevant information.

Formally, the relationship between the input and output of the SEA module is defined by the following equation:

$$F_{out} = S\left(F_{in}, \sigma\left(W \cdot (GAP(F_{in}) + H + V)\right)\right) \quad (1)$$

Here, F_{in} and F_{out} are the input and output feature maps.Gap (•) corresponds to global average pooling, while H and V represent the features extracted through horizontal and vertical strip pooling. W denotes the weights of a **1 X 1** convolution, and tands for the Sigmoid activation function. The ⊗ operation indicates an element-wise multiplication, which allows the learned spatial attention map to adaptively refine the input features.

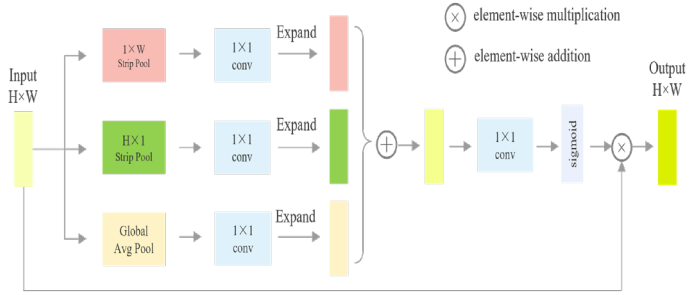

Figure 3. Architectural diagram of the proposed SEA module. The module leverages a combination of global average pooling (GAP) and anisotropic strip pooling (SP) to improve feature representation and refine spatial context.

Source: Adapted from Hou et al. (2020). Created by the authors (2025).

1

Decoder Coordinate Enhancement (DCE) Module

Attention mechanisms in neural networks draw their inspiration from the human visual system's innate ability to selectively focus on salient information while ignoring irrelevant details (Zeng et al., 2020). This principle has been effectively embodied in neural networks as a powerful tool for autonomously learning and extracting critical features from complex data (Chen et al., 2017a). In image analysis, this allows a model to concentrate on regions of interest and ignore extraneous backgrounds, thus optimizing computational resources and boosting performance. Following the rapid advancement of attention algorithms, their integration into deep learning models to enrich feature expression has become widespread and has led to outstanding outcomes (Zhang et al., 2020; Honarbakhsh et al., 2023). Conventional channel attention methods, like the SE module in SENet (Hu et al., 2018), calculate channel weights using 2D global pooling to achieve considerable performance gains. A key limitation, however, is that this approach only models inter-channel relationships and overlooks vital spatial location information. To overcome this, the Coordinate Attention module (Hou et al., 2021) addresses this by factorizing channel attention into two distinct, parallel 1D feature encoding steps, thereby embedding positional information into the attention maps. In our research, we improved the DeepLabv3+ network by introducing a Decoder Coordinate Enhancement (DCE) module into its decoder. Experiments demonstrate that this enhancement yields a significant improvement in performance.

The architecture of our proposed Decoder Coordinate Enhancement (DCE) module is illustrated in Figure 4. It adopts a dual-branch design that integrates spatial coordinate information into decoder features while preserving the original representation through a residual connection. For an input feature map denoted by $\boldsymbol{X}$, with dimensions $\boldsymbol{C} \times \boldsymbol{H} \times \boldsymbol{W}$, the

primary coordinate attention branch first aggregates features along the horizontal and vertical axes using two 1D global average pooling operations. This yields two position-aware feature vectors, Z^h and Z^w, These two vectors are subsequently concatenated and processed by a sequence of operations: a **1** x **1** convolution, batch normalization, and a non-linear activation function δ to encode spatial dependencies:

$$f = \delta(BN(F_1([z^h, z^w]))) \quad (2)$$

This encoded representation is then split and transformed by separate **1** x **1** convolutions and sigmoid functions to generate the final attention weights, g^h and g^w . These weights recalibrate the input features through element-wise multiplication:

$$Y_c(i,j) = X_c(i,j) \times g_c^h(i) \times g_c^w(j) \quad (3)$$

Concurrently, a parallel shortcut path applies a **1** x **1** convolution to the original input *X*. The module's final output is produced by combining this shortcut with the attention-recalibrated feature map *Y* via element-wise addition. This dual-branch architecture allows the module to effectively embed spatial coordinate information, enhancing its capacity for resolving fine-grained details and accurately segmenting small objects with low computational overhead.

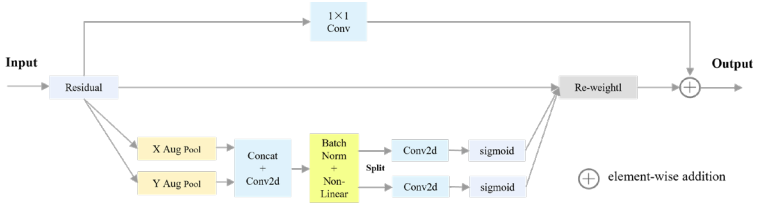

Figure 4. Decoder Coordinate Enhancement (DCE) Module

Source: Adapted from Hou et al. (2021). Created by the authors (2025).

1 Experiments and results

Experiment Datasets

To evaluate our model's performance, the PASCAL VOC 2012 dataset (Everingham et al., 2015) was selected, as it is a widely recognized benchmark for semantic segmentation. This dataset provides comprehensive pixel-level annotations for 21 categories, which are composed of 20 foreground object classes and one background class. For the training phase, we followed common practice and utilized the augmented training set of 10,582 images. The model's performance was subsequently validated and tested on the official validation and test sets, containing 1,449 and 1,456 images, respectively.

Experimental Setup and Evaluation Metrics

The experimental platform consisted of an NVIDIA RTX 3090 GPU, with all code developed in the PyTorch deep learning library. For training, the Adam optimizer was utilized with a learning rate of 0.001, a batch size of 8, and a weight decay parameter of 0.005. The model was trained for 100 epochs on the PASCAL VOC2012 dataset using a two-phase strategy.

The primary metrics for evaluating segmentation performance were Mean Intersection over Union (mIoU) and Mean Pixel Accuracy (mPA). For a total of k classes, these metrics are derived from the following pixel counts:

*P_{ij} The count of pixels belonging to class i but predicted as class ***j;***

*P_{ji} The count of pixels belonging to class *j* but predicted as class ***i;***

*P_{ii} The count of pixels correctly predicted as belonging to class ***i.***

The mathematical expressions for mIoU and mPA are given as follows:

$$MIoU = \frac{1}{K+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij} + \sum_{j=0}^{k}P_{ji} - P_{ii}} \quad (4)$$

$$mPA = \frac{1}{K+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}} \quad (5)$$

DENS-ASPP: Experimental Results and Analysis

To quantify the contribution of our DENS-ASPP module, we conducted a direct comparison with the baseline ASPP, with the outcomes summarized in Table 1. The data shows that substituting the standard ASPP with our module yields considerable improvements across both primary metrics. The proposed DENS-ASPP lifted the mIoU score from 69.84% to 71.31% and the mPA score from 85.51% to 86.37%. These results, representing absolute gains of 1.47 points in mIoU and 0.86 points in mPA, clearly confirm the superior ability of our module's design to aggregate multi-scale context.

Table 1 : Ablation Study: Quantitative Results for Different Modules

Model	MIoU(%)	MPA(%)
ASPP	69.84	85.51
DENS-ASPP	71.31	86.37

SEA Module: Experimental Results and Analysis

As shown in Table 2 below, when using the Spatial Enhancement Attention (SEA), the segmentation result reaches an MIoU of 72.51%. This is 2.67% higher than the baseline without attention, 0.89% higher than using only global attention (GAP), 0.53% higher than using only horizontal strip attention (SP-H), and 0.48% higher than using only vertical strip attention (SP-W). These results demonstrate that by fusing local information from strip attention with global context, the model can better focus on important target information, thereby effectively improving segmentation accuracy.

1

Table 2 : SEA Module: Experimental Results and Analysis

Model	MIoU(%)	MPA(%)
Without Attention	69.84	85.51
ASPP(GAP)	71.62	86.15
ASPP(SP-H)	71.98	86.68
ASPP(SP-W)	72.03	87.05
SEA	72.51	87.21

DCE Module: Experimental Results and Analysis

To specifically assess the effectiveness of our proposed Decoder Coordinate Enhancement (DCE) module, we conducted an ablation study on the PASCAL VOC 2012 dataset. As detailed in Table 3, this study compares the baseline DeepLabv3+ against versions enhanced with either the standard Coordinate Attention (CA) or our DCE module integrated into the decoder.

The baseline model established a performance of 69.84% mIoU and 85.51% mPA. Upon integrating the standard CA module, the scores saw a significant boost to 72.02% mIoU and 86.87% mPA, confirming the benefit of embedding positional information. By further substituting CA with our proposed DCE module, the performance was elevated to a new high of 72.16% mIoU and 87.12% mPA. These results suggest that while standard CA is effective, our DCE module's unique dual-branch architecture provides a more refined method for capturing spatial dependencies and enhancing decoder features, achieving superior performance with minimal added complexity.

Table 3 : Ablation Study of Decoder-Integrated Attention Mechanisms

Model	MIoU(%)	MPA(%)
Baseline	69.84	85.51
+CA	72.02	86.87
+DCE	72.16	87.12

Validation of Module Effectiveness

Table 2 details the ablation experiments conducted to validate our proposed modules on the PASCAL VOC2012 dataset. To establish a benchmark, the baseline model without any new components achieves an MIoU of 69.84%. Subsequently, we observe that the standalone addition of the DENS-ASPP, SEA, and DCE modules substantially improves performance, raising the MIoU to 71.31%, 72.51%, and 72.16%, respectively. The full model, which combines all three components, yields the best performance among all configurations at 72.65% MIoU. This superior result confirms that the modules work in synergy and validates the effectiveness of our overall architecture design.

Table 4 : Evaluating the Individual and Synergistic Contributions of the Proposed Modules on PASCAL VOC 2012.

Group	DENS-ASPP	SEA	DCE	MIoU(%)	MPA(%)
①	×	×	×	69.84	85.51
②	√	×	×	71.31	86.37
③	×	√	×	72.51	87.31
④	×	×	√	72.16	87.12
⑤	√	√	√	72.65	87.28

Category-wise Comparison of IoU

Figure 5 provides a detailed, category-by-category breakdown of the IoU scores to illustrate the impact of our model's modifications. As shown, our enhanced model (mIoU=72.65%) demonstrates superior or equal performance compared to the baseline (mIoU=69.84%) on every single class. The most dramatic improvements are seen in traditionally difficult categories that often contain fine structures. For example, the 'bicycle' IoU jumped by 15 points (from 0.53 to 0.68), while 'pottedplant' and

1

'chair' also saw significant boosts of 13 points each. This consistent, across-the-board improvement confirms that our multi-stage enhancement framework effectively strengthens the model's ability to handle a wide variety of object shapes and sizes, leading to more reliable segmentation.

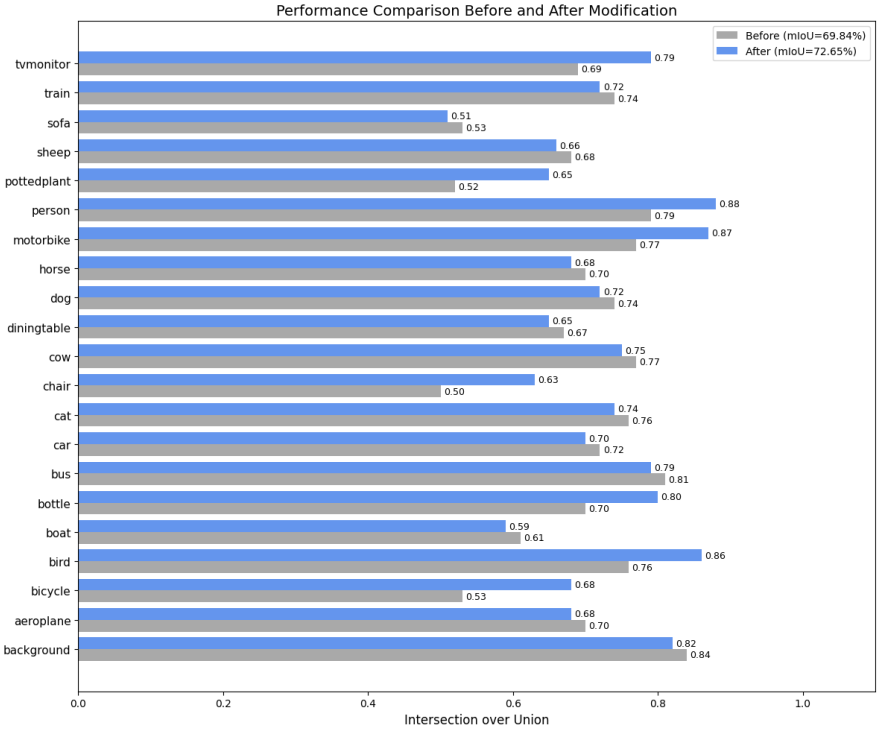

Figure 5. Comparison chart of category segmentation accuracy.
Source: Authors' experiments (2025)

Qualitative Comparison of Segmentation Results

To better illustrate the practical impact of our enhancements, we visually compared the segmentation results from our network with those from the original DeepLabv3+. A representative example is provided in Figure 5, which displays the original image (a), the ground truth (b), the baseline prediction (c), and our model's prediction (d). A close inspection

reveals that the baseline model often yields fragmented boundaries and erroneous classifications, particularly for small or complex objects, reflecting its limitations in multi-scale feature fusion. Conversely, our approach produces significantly sharper boundary delineation and more complete object coverage. These tangible improvements provide strong qualitative evidence that our structural modules successfully boost context-detail awareness in complex scenes.

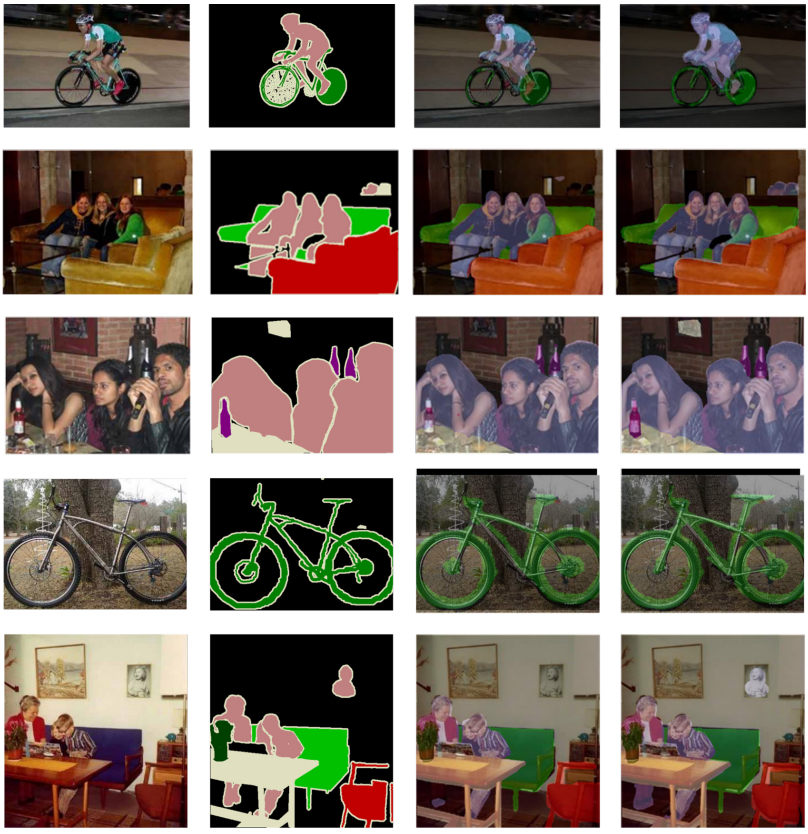

Figure 6. ComPArison of PASCAL VOC 2012 dataset segmentation results. Source: Authors' experiments (2025)

1

Discussion

The experimental results demonstrate consistent improvements in both mIoU and mPA across all configurations. When compared to previous studies such as Light-Deeplabv3+ (Ding & Qian, 2024) and DpNet (Wang et al., 2023), our proposed model achieved higher segmentation accuracy with comparable computational efficiency. While Light-Deeplabv3+ focused primarily on model compression, our work contributes an effective multi-stage attention strategy that strengthens boundary precision and small-object segmentation. Similarly, compared with Attention DeepLabv3+ (Azad et al., 2020), the introduction of the SEA module enables superior direction-aware context aggregation, and the DCE module refines spatial details beyond prior attention-based decoders. These comparisons affirm that our model not only builds upon the strengths of DeepLabv3+ but also extends its capability through lightweight, synergistic attention mechanisms.

Conclusion

This research presented an enhanced DeepLabv3+ network integrating three complementary modules—DENS-ASPP, SEA, and DCE-to improve the model's ability to manage multi-scale context and preserve fine-grained details. Experimental evidence confirmed that the integrated model achieved superior segmentation accuracy and sharper object boundaries compared with the baseline.

Practical Significance and Value: The proposed model offers high performance while maintaining low computational cost, making it applicable for real-time systems such as autonomous vehicles, medical imaging diagnostics, and mobile robotics. Beyond academic contributions, this work provides a practical foundation for deploying efficient semantic segmentation systems in resource-constrained environments. Future research will emphasize model compression, pruning, and knowledge distillation to further optimize accuracy-efficiency trade-offs.

First, at the encoder level, we proposed the DENS-ASPP module, which replaces the standard ASPP with a densely connected atrous cascade and parallel strip pooling to achieve a more powerful aggregation of multi-scale features. Second, these features are further refined by the Spatial Enhancement Attention (SEA) module, which models long-range, direction-aware dependencies through a combination of global and strip pooling. Finally, in the decoder, we introduced the Decoder Coordinate Enhancement (DCE) module, which leverages coordinate attention to optimize the fusion of high- and low-level features, significantly sharpening object boundaries.

Through extensive experiments, our integrated model demonstrated superior performance over the baseline, confirming that the synergistic combination of these modules leads to a more robust and accurate segmentation model. However, we acknowledge that these additions increase the model's parameter count, which could be a consideration for resource-constrained applications. Future work may explore model compression techniques to mitigate this trade-off.

Recommendations

In this paper, we presented a significantly enhanced network based on the DeepLabv3+ architecture, introducing a comprehensive, multi-stage feature enhancement framework. Our approach is composed of three synergistic modules: the DENS-ASPP module for robust multi-scale feature extraction, the SEA module for capturing long-range spatial context, and the DCE module for refining details in the decoder. Experimental results validate that our integrated method achieves substantial improvements in segmentation accuracy, demonstrating a particular aptitude for resolving complex boundaries and accurately identifying small-scale objects in challenging scenes. The principles developed in this work can serve as

a valuable reference for advancing semantic segmentation in domains such as autonomous navigation and medical diagnostics.

Despite these promising results, we acknowledge certain limitations. The integration of our advanced modules inevitably increases the model's computational complexity. Therefore, a primary focus of future work will be on model optimization and generalization. We intend to explore model compression techniques, such as network pruning and knowledge distillation, in conjunction with semi-supervised learning methods. The goal is to develop a model that strikes an effective balance between high accuracy and lightweight efficiency, ensuring robust adaptability for wider real-world deployment.

References

Azad, R., Asadi-Aghbolaghi, M., Fathy, M., & Escalera, S. (2020, August). *Attention Deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation.* In *European Conference on Computer Vision* (pp. 251–266). Springer.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(4), 834–848.

Ding, P., & Qian, H. (2024). *Light-Deeplabv3+: A lightweight real-time semantic segmentation method for complex environment perception. Journal of Real-Time Image Processing, 21*(1), 1.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). *The Pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 111*(1), 98–136.

Honarbakhsh, V., Siahkoohi, H. R., Rezghi, M., & Sabeti, H. (2023). *SeisDeepNET: An extension of Deeplabv3+ for full waveform inversion problem. Expert Systems with Applications, 213*(1), 118848.

Hou, Q., Zhou, D., & Feng, J. (2021). *Coordinate attention for efficient mobile network design.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13713–13722). IEEE.

Hu, J., Shen, L., & Sun, G. (2018). *Squeeze-and-excitation networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7132–7141). Salt Lake City, UT, US.

Jiang, L., Zhou, W., Li, C., & Wei, Z. (2021, March). *Semantic segmentation based on DeeplabV3+ with multiple fusions of low-level features.* In 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) (pp. 1957–1963). IEEE.

Lee, K., & Park, K. S. (2024). *Deep learning model analysis of drone images for unauthorized occupancy detection of river site. Journal of Coastal Research, 116*(SI), 284–288.

Li, L., Zhang, W., Zhang, X., Emam, M., & Jing, W. (2023). *Semi-supervised remote sensing image semantic segmentation method based on deep learning. Electronics, 12*(2), 348.

Lili, G., & Jinzhi, Z. (2022, August). *A lightweight network for semantic segmentation of road images based on improved DeepLabv3+.* In *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)* (pp. 832–837). IEEE.

Wang, J., Zhang, X., Yan, T., & Tan, A. (2023). *Dpnet: Dual-pyramid semantic segmentation network based on improved Deeplabv3 plus. Electronics, 12*(14), 3161.

Wenkuana, D., & Shicai, G. (2023). *Hazy images segmentation method based on improved DeeplabV3. Academic Journal of Computer and Information Science, 6*(5), 21–29.

Xiang, S., Wei, L., & Hu, K. (2024). *Lightweight colon polyp segmentation algorithm based on improved DeepLabV3+. Journal of Cancer, 15*(1), 41–50.

Yang, M., Yu, K., Zhang, C., Li, Z., & Yang, K. (2018). *DenseASPP for semantic segmentation in street scenes.* In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3684–3692). IEEE.

Yang, Z., Peng, X., & Yin, Z. (2020, October). *Deeplab_v3_plus-net for image semantic segmentation with channel compression.* In 2020 IEEE 20th International Conference on Communication Technology (ICCT)(pp. 1320–1324). IEEE.

Zeng, H., Peng, S., & Li, D. (2020, November). *Deeplabv3+ semantic segmentation model based on feature cross attention mechanism. Journal of Physics: Conference Series, 1678*(1), 012106.

Zhang, Z., Huang, J., Jiang, T., Sui, B., & Pan, X. (2020). *Semantic segmentation of very high-resolution remote sensing image based on multiple band combinations and patchwise scene analysis. Journal of Applied Remote Sensing, 14*(1), 016502.

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