Triplet Attention Enhanced DeepLab V3+ for Semantic Segmentation: Improving Feature Extraction and Fine-Grained Understanding

Zhuoran Li[1,*,†], Xun Shu[2], Yancong Deng[3]

{zhuorali2-c@my.cityu.edu.hk[1], 32021210121@cueb.edu.cn[2], yancongdeng@gmail.com[3]}

City University of Hong Kong, Hong Kong, China[1]
Capital University of Economics and Business, Beijing, China[2]
University of California San Diego, CA, United States[3]
*corresponding author
[†]The author is the first author.

Abstract. Semantic segmentation is a critical task in computer vision that requires assigning class labels to individual pixels for a deeper understanding of visual scenes. This paper explores methods to enhance the DeepLab V3+ model by integrating attention mechanisms, specifically Triplet Attention and CBAM, into the backbone, ASPP module, and decoder. We conducted extensive experiments using the Cityscapes dataset to assess the impact of these attention mechanisms. Our results demonstrate that Triplet Attention, particularly when applied to the backbone, significantly improves segmentation performance with minimal computational overhead, outperforming CBAM in most configurations. This study highlights the effectiveness of attention mechanisms in improving the precision and semantic understanding of segmentation models. Future work will explore additional attention mechanisms and expand their application to broader vision-related tasks.

Keywords: Semantic Segmentation, Triplet Attention Mechanism, DeeplabV3+, Improved Feature Extraction.

1 Introduction

Semantic segmentation is a fundamental task in computer vision, where the goal is to assign a class label to each pixel in an image. A well-designed and trained model can effectively determine the location and extent of pre-defined object classes, allowing for a more comprehensive understanding of visual context and scene structure. Specifically, semantic segmentation aims to achieve dense classification by assigning pixel-wise semantic labels, contributing to a deeper understanding of visual content.

As a high-level vision task, semantic segmentation plays a critical role in scene understanding, which is a core challenge in computer vision. This is essential for a wide range of real-world applications, including autonomous driving, human-computer interaction, computational photography, image search engines, and augmented reality (AR) [1-3].

In recent years, deep learning-based segmentation models have shown remarkable advancements, with the DeepLab family standing out as one of the most effective and

CONF-MLA 2024, November 21, Adana, Turkey

DOI 10.4108/eai.21-11-2024.2354635

representative approaches [4,5]. The DeepLab models incorporate techniques such as conditional random fields, atrous convolutions, and pyramid pooling to efficiently capture multi-dimensional information, enabling precise pixel-level semantic segmentation.

On the other hand, Attention mechanisms, initially popularized in sequence-to-sequence tasks such as natural language processing, have gained traction in computer vision due to their ability to capture long-range dependencies. Introduced by Vaswani et al. [6], attention mechanisms mimic the human cognitive process of focusing on relevant parts of visual and linguistic inputs, thereby improving model performance and generalization [7]. To further enhance segmentation performance, this paper proposes an improved DeepLab v3+ model that integrates the Triplet Attention mechanism proposed by Misra et al. [8]. Triplet Attention is introduced to boost segmentation accuracy with minimal computational overhead. This paper will detail the design, implementation, and experimental evaluation of this enhanced approach.

2 Related Work

2.1 DeepLab Model Family

Semantic segmentation is a critical task in computer vision, aiming to understand and classify pixel-level semantic information. The DeepLab family of models is one of the most prominent architectures for this task. The series began with DeepLab-V1, introduced by Chen et al. [9], which pioneered the use of atrous convolution to expand the receptive field, allowing the model to capture multi-scale image information. DeepLab-V2 enhanced this design by incorporating the atrous spatial pyramid pooling (ASPP) module to efficiently integrate multi-scale features, while also adopting a deeper residual network backbone to replace the earlier VGG architecture for improved feature extraction [10].

DeepLab-v3 further advanced the model by introducing multi-grid techniques for processing higher-resolution images. It also refined the ASPP module by utilizing different dilation rates and adding batch normalization and activation layers to better fuse low-level and high-level features [11].

DeepLab-v3+ represents a major breakthrough in recent years. It employs an encoder-decoder architecture, building upon the ideas of previous designs [12]. In the encoder, the ASPP structure is used to extract visual features after being compressed by four bottleneck modules in the ResNet backbone. These multi-scale features are concatenated and passed to the decoder after a 1×1 convolution for compression. Meanwhile, low-level features extracted from the DCNN backbone, compressed by two bottleneck modules, are processed through a 1×1 convolutional layer and concatenated with the upsampled ASPP features to produce the final feature map. A final 1×1 convolution is applied to adjust the number of channels, followed by upsampling to restore the original image size for pixel-wise prediction.

2.2 Attention mechanisms

Attention mechanisms have become a crucial design element in sequence-to-sequence tasks and have been successfully applied to computer vision due to their ability to focus on important parts of the input while disregarding irrelevant background information. Attention mechanisms

exploit dependencies across different dimensions of the input, such as channel attention in Squeeze-and-Excitation Networks and spatial attention [13,14].

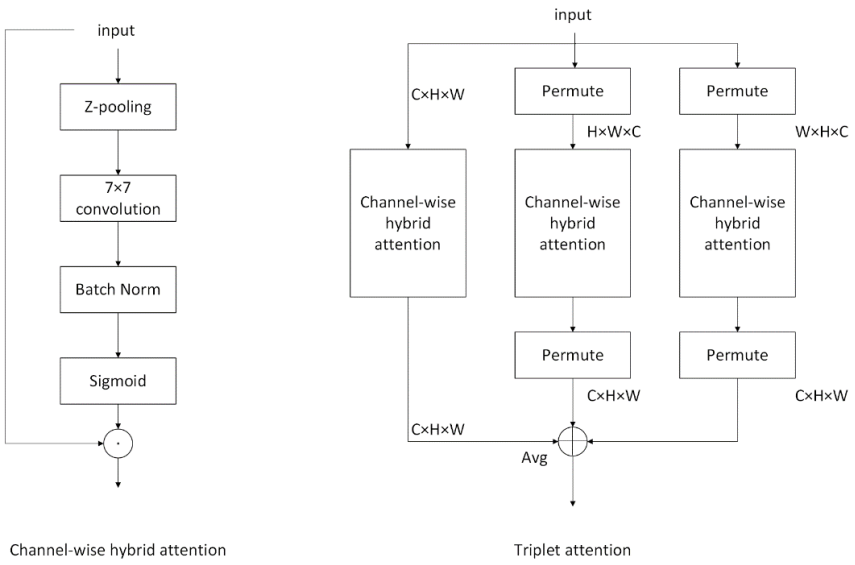

Fig. 1. structure of Triplet attention modules

As shown in Figure 1, this paper focuses on the Triplet Attention mechanism, a lightweight yet effective method that uses a three-branch structure to capture cross-dimensional interactions. Triplet Attention can be easily integrated into traditional backbone networks with negligible computational overhead. The structure consists of three branches, each responsible for capturing interactions across different dimensions of the input tensor. In the first branch, focusing on channel and width (C and W), attention weights are calculated using z-pooling, followed by a 7×7 convolution and Sigmoid activation. Similarly, the other two branches capture dependencies for channel and height (C and H), and height and width (H and W). The resulting attention weights are multiplied by the original input and passed through average pooling to produce the final attention mechanism output.

We also explore the Convolutional Block Attention Module (CBAM), another attention mechanism known for its flexibility in fitting into any CNN architecture with minimal overhead, as proposed by Woo et al. [15,16]. CBAM considers both channel and spatial information, applying max pooling and average pooling along each dimension. The pooled outputs are passed through convolutional layers and a Sigmoid activation function, producing attention weights that are multiplied with the original input to generate the final output.

3 Methodology

3.1 Dataset

Popular semantic segmentation datasets and benchmarks include PASCAL VOC, Cityscapes, and ADE20K, which provide labeled images to train and evaluate segmentation models. In this paper, we selected the Cityscapes dataset [17], which includes 5,000 finely annotated images, due to its appropriate data size and high annotation quality.

Cityscapes is a large-scale benchmark dataset designed for visual semantic understanding in urban street scenes. It provides comprehensive semantic, instance-level, and dense pixel-wise annotations for 30 distinct object classes, which can be grouped into 8 high-level categories: flat surfaces, humans, vehicles, constructions, objects, nature, sky, and void. The dataset comprises approximately 5,000 finely annotated images and an additional 20,000 coarsely annotated samples for various purposes. These images were captured in 50 cities over several months, at different times of the day, and in favorable weather conditions to ensure diversity. The images were extracted from videos, with frames manually selected to include a rich variety of objects, diverse scene layouts, and complex background compositions. This comprehensive dataset is an invaluable resource for computer vision research, particularly in tasks such as semantic and instance segmentation, object detection, and urban scene understanding. The diversity and density of annotations make Cityscapes a prominent benchmark for evaluating the performance and generalization of state-of-the-art computer vision models.

For our research, we used the 5,000 finely annotated images from the Cityscapes dataset and resized them to a resolution of 256×256 to balance computational cost with data quality.

3.2 Model design

Triplet Attention, as discussed earlier, is designed to optimize vision models by helping them focus on crucial parts of the input. Its flexible design allows it to be easily integrated into existing models, typically by inserting it after convolutional layers.

Oue optimized model is based on the original DeepLab V3+ architecture, which consists of three main components: the backbone, the ASPP module, and the decoder, as shown in Figure 3. The backbone uses the classic ResNet-50 architecture to process input image data. As seen in Figure 2, ResNet-50 comprises four convolutional layers, each made up of a series of bottleneck modules. To enhance the backbone, we introduced Triplet Attention after each bottleneck module, following the approach proposed in the original Triplet Attention paper. Alternatively, we experimented with adding attention mechanisms after each backbone layer, inspired by the research of Li et al. [18], which explored improvements in applying DeepLab V3+ to agricultural image segmentation.

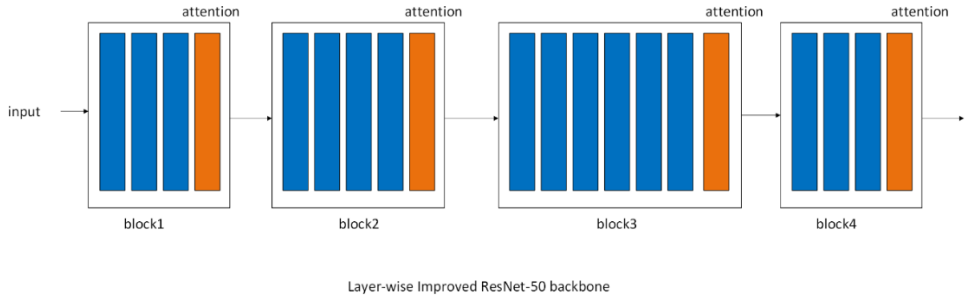

Fig. 2. structure of Layer-wise attention improved ResNet-50 backbone

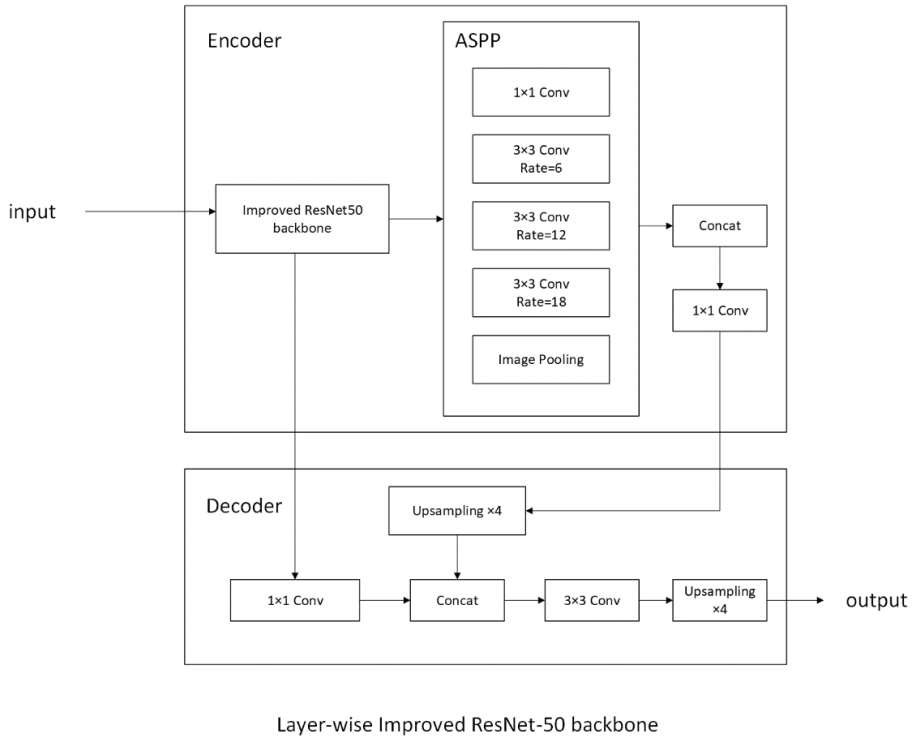

Fig. 3. structure of Layer-wise attention improved DeepLab-v3+ model

For the remaining part of the DeepLab V3+ model, the ASPP module consists of atrous convolution pyramids and a 1×1 convolution for feature mixing. The decoder is responsible for combining low-level and high-level information, followed by a 3×3 convolution for reshaping the feature channels. We explored adding attention mechanisms after relevant convolutional layers in both the ASPP and decoder modules to assess potential performance gains.

To evaluate the effectiveness of Triplet Attention, we also implemented and tested other attention mechanisms, such as CBAM. We applied CBAM at the same locations mentioned above to compare its performance against Triplet Attention.

4 Experiment & Result

4.1 Experiment design

In this section, we present the implementation and performance comparison of the various models discussed in previous sections. The main objective of this study is to evaluate the effectiveness of integrating the Triplet Attention mechanism into the DeepLab-V3+ semantic segmentation model. Specifically, we experimented by inserting Triplet Attention modules at different positions within the original model and evaluated the performance improvements. To further validate the robustness of our approach, we also implemented a comparative study by replacing the Triplet Attention modules with CBAM (Convolutional Block Attention Module) and assessing the resulting performance metrics.

The table below outlines the naming conventions and design details of each model configuration used in the experiment, providing clarity regarding the variations tested:

Table 1. illustration of model adaptation name and their architecture

Model name	Model architecture explanation
Original DeepLab v3+	The original implementation of deeplab v3+ model
ASPP	Applying triplet modules after each ASPP blocks of model
CBAM	Applying CBAM modules after each ResNet layers of model
Layerwise	Applying triplet modules after each ResNet layers of model
Blockwise	Applying triplet modules after each ResNet bottleneck blocks
ASPP+Layer	Applying triplet modules after each ASPP blocks of model

4.2 Experiment result

The performance of the models was evaluated using three primary metrics: training and validation loss, mean Intersection over Union (mIoU) and pixel-wise accuracy. mIoU is a widely accepted metric for semantic segmentation tasks, measuring the overlap between predicted and ground-truth segments, while pixel-wise accuracy provides an aggregate measure of correctly classified pixels across the entire dataset.

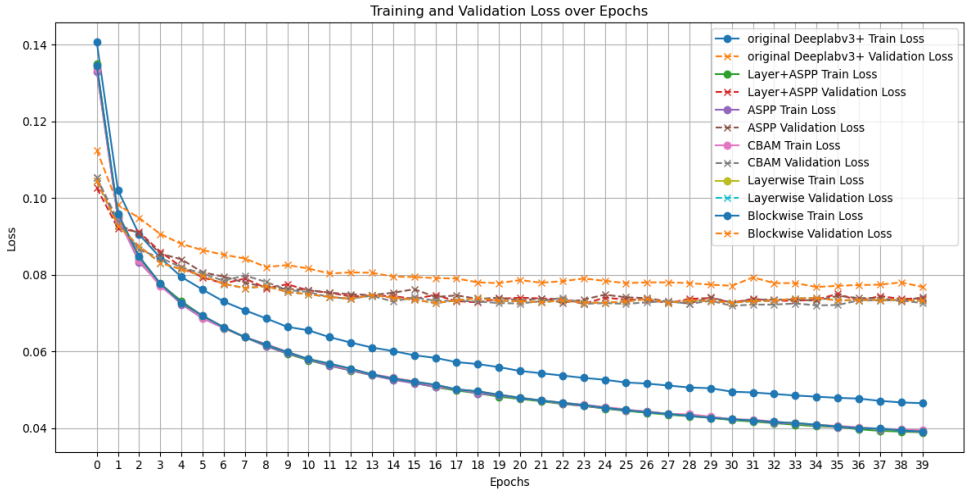

Fig. 4. training and validation losses in training process

Figure 4 illustrates the training and validation loss curves across different models. The losses for all models follow a smooth downward trajectory, indicating stable training across the board. However, the block-wise model exhibits significantly higher losses compared to the other models. This is likely due to inappropriate optimization or an excessive number of attention modules being added, which may have introduced complexity that hindered convergence. These results suggest that further hyperparameter tuning or architectural adjustments are required for the block-wise model.

Figures 5 and 6 provide insights into the maximum mIoU and pixel-wise accuracy achieved by each model. As shown, the layer-wise model consistently outperforms other models in terms of both mIoU and pixel-wise accuracy. To delve deeper into the results, we performed a thorough analysis of each model's performance.

The original DeepLab-V3+ model serves as the baseline for this study. It achieved a maximum mIoU of 50.17% and a pixel-wise accuracy of 0.8366. These values establish a benchmark against which the performance improvements brought by attention mechanisms can be assessed.

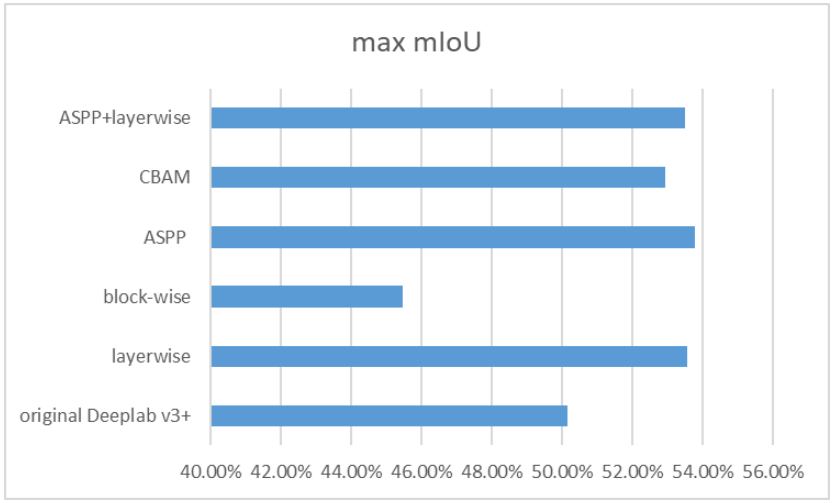

Fig. 5. result mIoU of all models tested

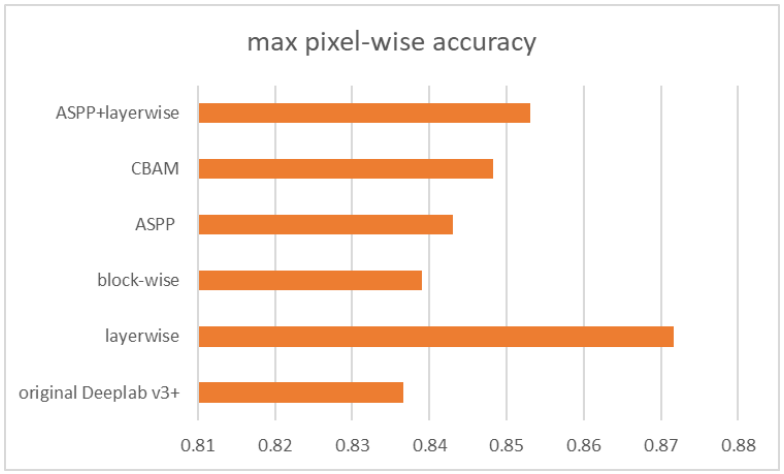

Fig. 6. result pixel-wise accuracy of all models tested

In the layer-wise improvement experiment, we inserted Triplet Attention modules after each ResNet layer in the backbone architecture. The model achieved a significant boost, reaching a maximum mIoU of 53.58% and a pixel-wise accuracy of 0.8716. This represents a notable improvement over the baseline, with only a slight increase in parameter count (from 39,761,331 to 39,761,628 parameters). The effectiveness of Triplet Attention in capturing cross-dimensional dependencies is evident from this considerable lift in performance, which is particularly beneficial in scenarios requiring fine-grained segmentation.

The block-wise improvement involved adding Triplet Attention modules after each bottleneck module in the ResNet-50 backbone. This approach, however, led to suboptimal results, with a maximum mIoU of 45.49% and a pixel-wise accuracy of 0.8391, both of which are lower than

even the baseline model. One possible explanation for this underperformance is the over-complication of the model. The insertion of too many attention modules may have increased the model's complexity, leading to difficulties in optimization and slower convergence during training. This finding underscores the importance of carefully balancing architectural enhancements with training efficiency.

To further enhance feature extraction capabilities, we experimented with ASPP improvements by adding attention modules within each ASPP layer of the encoder. This variant achieved a maximum mIoU of 53.77% and a pixel-wise accuracy of 0.8431. While this shows a moderate improvement compared to the layer-wise enhancement, the increase is relatively modest, suggesting that ASPP enhancement alone may not be as impactful as expected.

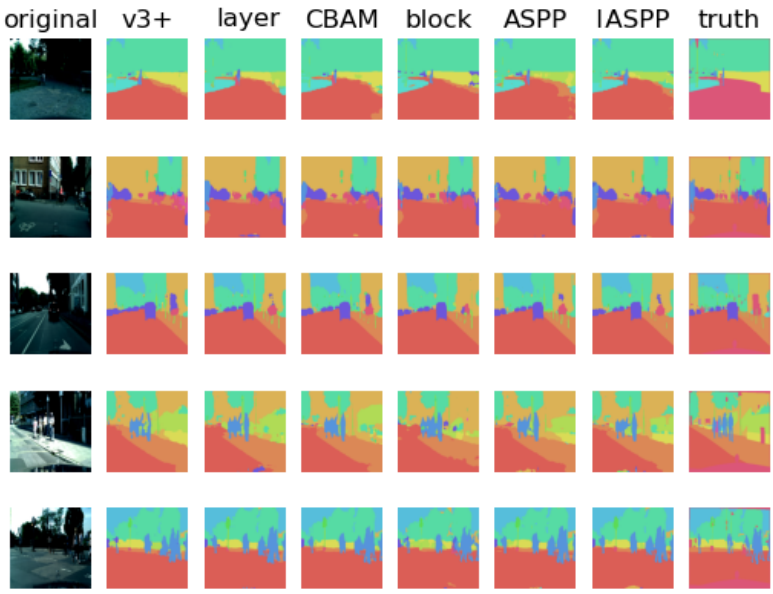

Fig. 7. example of segmentation results of different model comparison

We also explored the combination of Layer-wise and ASPP improvements, applying both techniques simultaneously. The resulting model achieved a maximum mIoU of 53.49% and a pixel-wise accuracy of 0.8531. This indicates that simply combining the two approaches does not necessarily result in a significant cumulative improvement. It seems that both methods contribute to similar improvements in different parts of the model, and their combination does not yield a synergistic effect.

Finally, we evaluated the performance of CBAM modules by replacing Triplet Attention modules in the Layer-wise improved model. The model achieved a maximum mIoU of 52.95% and a pixel-wise accuracy of 0.8438. While CBAM shows a moderate uplift in performance compared to the baseline, it falls short of the performance achieved by Triplet Attention. This reinforces the strength of Triplet Attention in capturing complex spatial relationships in the segmentation task.

In Figure **7**, we provide qualitative visual comparisons of the segmentation results produced by different models. These samples were drawn from the evaluation set and provide a visual demonstration of the differences between the model outputs. For clarity, abbreviations such as "lASPP" represent the combined Layer-wise plus ASPP improvement discussed above. Although the visual differences between models are subtle, the evaluation metrics and segmentation masks consistently indicate the superiority of the Layer-wise Triplet Attention model in producing more accurate and refined segmentations.

5 Conclusion

Semantic segmentation is a foundational task in the field of computer vision, crucial for understanding and analyzing the structure of visual scenes at a pixel level. In this paper, we explored multiple strategies to enhance the performance of the DeepLab V3+ model through the integration of attention mechanisms. Specifically, we introduced and evaluated two prominent attention mechanisms: Triplet Attention and CBAM, integrating them into the model's backbone, ASPP module, and decoder.

Our experimental results demonstrate that attention mechanisms significantly improve the segmentation performance of the DeepLab V3+ model with only a marginal increase in computational resource requirements. Among the various strategies explored, the integration of Triplet Attention in the model's backbone exhibited the most pronounced improvements, highlighting its effectiveness in capturing cross-dimensional dependencies and enhancing feature extraction.

In summary, attention mechanisms have proven to be highly effective in augmenting the semantic understanding and fine-grained detail prediction of segmentation models like DeepLab V3+. This suggests promising avenues for further advancements in the field. Future work will involve exploring additional attention mechanisms and applying these methods to other computer vision tasks, with the aim of continuing to push the boundaries of image understanding.

References

[1] Garcia-Garcia, A. *et al.* (2017) *A review on Deep Learning techniques applied to semantic segmentation*, *arXiv.org*. Available at: https://arxiv.org/abs/1704.06857 (Accessed: 24 November 2024).

[2] Ess, A. *et al.* (2009) 'Segmentation-based urban traffic scene understanding', *Procedings of the British Machine Vision Conference 2009* [Preprint]. doi:10.5244/c.23.84.

[3] Norouzi, A. *et al.* (2014) 'Medical image segmentation methods, algorithms, and applications', *IETE Technical Review*, 31(3), pp. 199–213. doi:10.1080/02564602.2014.906861.

[4] Fahad Lateef *et al.* (2019) *Survey on semantic segmentation using Deep Learning Techniques*, *Neurocomputing*. Available at: https://www.sciencedirect.com/science/article/abs/pii/S092523121930181X (Accessed: 24 November 2024).

[5] Garcia-Garcia, A. *et al.* (2018) 'A survey on Deep Learning techniques for image and video semantic segmentation', *Applied Soft Computing*, 70, pp. 41–65. doi:10.1016/j.asoc.2018.05.018.

[6] Vaswani, A. *et al.* (2017) *Attention is all you need*, *arXiv.org*. Available at: https://arxiv.org/abs/1706.03762v5 (Accessed: 24 November 2024).

[7] Guo, M.-H. *et al.* (2021) *Attention mechanisms in computer vision: A survey*, *arXiv.org*. Available at: https://arxiv.org/abs/2111.07624 (Accessed: 24 November 2024).
[8] Misra, D. *et al.* (2021) 'Rotate to attend: Convolutional triplet attention module', *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* [Preprint]. doi:10.1109/wacv48630.2021.00318.
[9] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014, December 22). *Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs*. arXiv.org. https://arxiv.org/abs/1412.7062.
[10] Chen, L.-C., Papandreou, G., *et al.* (2018) 'DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), pp. 834–848. doi:10.1109/tpami.2017.2699184.
[11] Chen, L.-C. *et al.* (2018) 'Encoder-decoder with atrous separable convolution for Semantic Image segmentation', *Lecture Notes in Computer Science*, pp. 833–851. doi:10.1007/978-3-030-01234-2_49.
[12] Chen, L.-C. *et al.* (2017) *Rethinking atrous convolution for Semantic Image segmentation*, *arXiv.org*. Available at: https://arxiv.org/abs/1706.05587 (Accessed: 24 November 2024).
[13] Liu, T. *et al.* (2022) *Spatial channel attention for deep convolutional neural networks*, *MDPI*. Available at: https://www.mdpi.com/2227-7390/10/10/1750.
[14] Hu, J., Shen, L. and Sun, G. (2018) 'Squeeze-and-excitation networks', *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* [Preprint]. doi:10.1109/cvpr.2018.00745.
[15] Woo, S. *et al.* (2018) 'CBAM: Convolutional Block Attention Module', *Lecture Notes in Computer Science*, pp. 3–19. doi:10.1007/978-3-030-01234-2_1.
[16] Liu, jiajia and huang, chidong (2024) CMANet: Convolutional Multi-channel Fusion Attention Mechanism Network for semantic image segmentation [Preprint]. doi:10.22541/au.171246799.96972852/v1.
[17] Cordts, M. *et al.* (2016) *The cityscapes dataset for Semantic Urban Scene understanding*, *arXiv.org*. Available at: https://arxiv.org/abs/1604.01685 (Accessed: 24 November 2024).
[18] Li, K. *et al.* (2022) *Attention-optimized DeepLab v3 + for automatic estimation of cucumber disease severity - plant methods*, *BioMed Central*. Available at: https://plantmethods.biomedcentral.com/articles/10.1186/s13007-022-00941-8 (Accessed: 26 October 2024).