Structured Pruning for Semantic Segmentation Models: A Comprehensive Analysis

Arya Saraswathi Potti
Department of Electrical Engineering
North Carolina State University
Raleigh, USA
aspotti@ncsu.edu

Amarnath Shinde
Department of Electrical Engineering
North Carolina State University
Raleigh, USA
avshinde@ncsu.edu

Wujie Wen
Department of Computer Science
North Carolina State University
Raleigh, USA
wwen2@ncsu.edu

***Abstract*—This project investigates the application of structured pruning techniques to enhance the efficiency of real-time semantic segmentation models crucial for computer vision applications. Employing popular architectures such as ICNet [1] and BiSeNetV2 [2], we conducted a comprehensive analysis of the impact of structured pruning on model accuracy, inference speed, and parameter reduction. Our research explores the trade-off between model compression and task-specific performance, considering various pruning ratios and strategies. The findings provide valuable insights into the feasibility and efficacy of structured pruning for optimizing neural networks, offering a roadmap for deploying efficient semantic segmentation models in resource-constrained environments within the evolving landscape of machine learning.**

***Index Terms*—Structured pruning, real-time semantic segmentation,ICNet, BiSeNetV2, model accuracy,model compression.**

I. Introduction

Real-time semantic segmentation holds paramount importance in various computer vision applications due to its ability to provide pixel-level understanding of images, enabling more sophisticated and context-aware decision-making processes. This capability is particularly crucial in scenarios where rapid and accurate scene analysis is essential, such as in autonomous vehicles, robotics, and augmented reality. The significance of real-time semantic segmentation lies in its potential to enhance the efficiency and performance of systems that rely on instantaneous and precise interpretation of visual information. In the context of this project, the choice of real-time semantic segmentation models, including ICNet and BiSeNetV2 is motivated by their proven effectiveness in balancing accuracy and speed.

The efficiency of real-time semantic segmentation models is a critical aspect facing challenges that need to be addressed for the successful deployment of these models in time-sensitive applications. One major challenge lies in achieving a balance between high model accuracy and real-time inference speed. Many state-of-the-art semantic segmentation models, are designed to deliver accurate pixel-level predictions, but this often comes at the cost of increased computational demands, limiting their suitability for real-time applications.

The computational complexity of these models poses challenges for deployment on resource-constrained devices, such as edge computing platforms or embedded systems commonly used in robotics and autonomous vehicles. The demand for real-time processing in these applications necessitates models that can make rapid predictions without compromising the quality of segmentation results.

Additionally, the scalability of real-time semantic segmentation models across diverse datasets and real-world conditions is a challenge. Models optimized for specific scenarios may struggle to generalize effectively, impacting their adaptability to dynamic environments. This limitation is particularly pertinent in applications like autonomous driving, where the model must perform robustly across varied lighting conditions, weather, and terrain. The problem statement, therefore, revolves around devising strategies to enhance the efficiency of real-time semantic segmentation models, specifically addressing the trade-off between accuracy and speed, optimizing for resource constraints, and ensuring robust generalization across diverse operational scenarios.

Model pruning is a technique used to reduce the size (number of parameters) of a neural network by removing certain weights or neurons deemed less important. This process is often done after the initial training of a deep learning model. Pruning is very beneficial for dealing with model complexity because it results in a more compact and computationally efficient network.

The main underlying objective of this project is to conduct a comprehensive quantitative analysis to measure the impact of structured pruning on model accuracy, inference speed, and parameter reduction. This includes perform qualitative assessments through visualizations of segmentation outputs to understand the effect of structured pruning on the task-specific performance of the models. We are also investigating and analyzing the trade-off between model compression achieved through structured pruning and the maintenance of semantic segmentation quality, ensuring a balanced compromise between accuracy and efficiency.

In this project, we are using 'Cityscapes' [4] dataset for training and inference results.

II. Related Work

Semantic segmentation [5] has witnessed significant advancements with the development of various models, each aiming to strike a balance between accuracy and efficiency.

U-Net [6], known for its expansive architecture, has been widely used in medical image segmentation. Its encoder-decoder structure forms the basis for subsequent models. It is build on fully convoluted network.

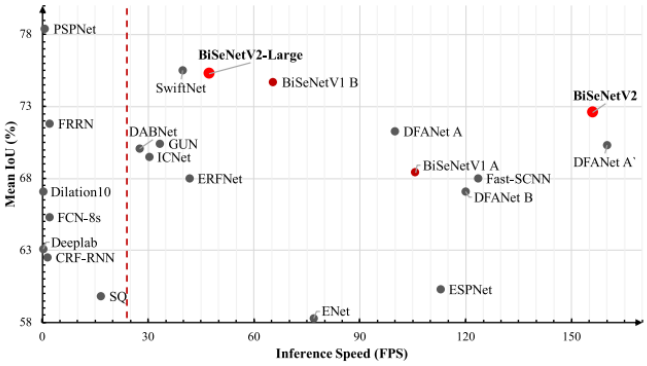

Fig. 1. Speed-accuracy trade-off comparison on the Cityscapes test set

SegNet [15] is structured as a stack of encoders, which is subsequently mirrored by a corresponding decoder stack that feeds into a softmax classification layer. The encoder stack captures hierarchical features from the input image, while the decoders play a crucial role in mapping low-resolution feature maps at the output of the encoder stack to full input image size feature maps. This architectural design specifically addresses a significant drawback observed in previous deep learning approaches.

ENet [7], short for "Efficient Neural Network," introduces a novel deep neural network architecture designed explicitly for tasks demanding low-latency operation. This innovative model exhibits remarkable efficiency, boasting a speed improvement of up to 18 times compared to existing models. Impressively, ENet achieves this enhanced speed while requiring a mere 75 times fewer Floating Point Operations (FLOPs) and employing a significantly leaner parameter count with a reduction of 79 times.

PSPNet [8] employs a pyramid pooling module to capture contextual information at different scales. This enhances the model's ability to understand global context. Similarly, ICNet, which stands for Image Cascade Network, is an innovative deep neural network that incorporates multi-resolution branches was proposed at later stages of research of these segmentation models. BiSeNet, short for Bilateral Segmentation Network, represents a state-of-the-art deep neural network architecture was designed explicitly for real-time semantic segmentation tasks. Subsequently, ESPNet [9], short for Efficient Spatial Pyramid of Dilated Convolutions was introduced as a cutting-edge deep neural network architecture tailored specifically for semantic segmentation tasks.

Nowdays, pioneering methods like DeepLabV3 [10] includes neural network architecture that integrates deep convolutional networks with atrous convolutions and fully connected Conditional Random Fields (CRFs) to achieve precise and context-aware segmentation are used.

Pruning [11] of segmentation models helps to reduce the model size, making it more feasible to deploy on devices with limited resources, such as edge devices or mobile platforms. This results in a sparser model with fewer operations, leading to faster inference times.It enables the deployment of segmentation models on edge devices, extending the reach of these applications to scenarios where cloud-based processing may not be practical.

In the context of pruning the above models, filter-wise pruning [12] approach to FPGA Implementation of Fully Convolutional Network for Semantic Segmentation was introduced firstly as filter-wise pruning, a technique that selectively removes redundant filters from the FCN architecture, thereby reducing computational complexity and memory requirements.

After that, CAP [13] also known as Context-Aware Pruning for Semantic Segmentation proposes a novel approach was important in terms of contextual information during channel pruning for semantic segmentation networks.

Despite evolution in semantic segmentation models and advanced pruning techniques, the trade-off issues between accuracy and size exists. In our method we aimed to maintain that enables efficient deployment while sustaining high segmentation accuracy on various models.

Taking into consideration the speed-accuracy trade-off of different models as seen in Fig 1., in the proposed method the experiments were conducted for ICNet and BiSeNetV2-Large [3] (hereafter referenced as BiSeNetV2).

III. Proposed Method

A computer vision task known as "real-time semantic segmentation" entails instantly classifying each pixel in an input image into one of several item classes. This makes it possible for programs like augmented reality, robotics, and driverless cars to make quick decisions based on their understanding of the scene. Real-time semantic segmentation presents a trade-off between speed and accuracy since obtaining high precision frequently necessitates sophisticated models, which could impair real-time performance. Furthermore, managing heterogeneous environmental variables, such as fluctuating lighting and occlusions, presents a major obstacle to guaranteeing strong and dependable segmentation outcomes in dynamic scenarios.

In this paper, we are proposing a method to deal with the complexity of the model by introducing the concept of pruning into the picture. We propose to prune our model weights using a method called structured pruning to achieve lesser model complexity which would be used for the final predictions.

Fig. 2. [Left Image] Reference image from the Cityscapes dataset [Right Image] Segmented mask predicted from the pretrained ICNet model

A. Pruning

Pruning minimizes the overall complexity of the model by deleting duplicate or less important connections, resulting in benefits like faster inference, decreased memory requirements, and enhanced deployment on resource-constrained devices. There are various pruning techniques, including magnitude-based pruning (removing small-weight connections), structured pruning (removing entire neurons or channels), and iterative pruning methods. The key goal is to maintain or even improve the model's performance while making it more lightweight and efficient, especially in scenarios where computational resources are limited.

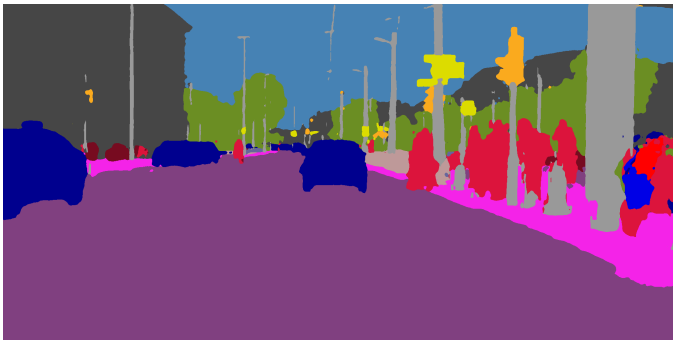

Fig. 3. Segmentation mask for pruned ICNet

B. Structured Pruning for Model efficiency

Structured pruning [14], a technique in neural network compression, focuses on reducing model size by eliminating entire structured components, such as neurons, channels, or layers, rather than individual weights. This method preserves the overall architecture while significantly reducing parameters. The process involves identifying groups of weights or entire filters based on criteria like magnitude or importance during training, resulting in a more streamlined model. Although structured pruning maintains the inherent structure of convolutional neural networks (CNNs), achieving a balance between compression and model accuracy remains a challenge.

In our proposed method, we leverage structured pruning techniques within the PyTorch framework to enhance neural network efficiency. Specifically targeting convolutional and linear layers using PyTorch's pruning module and the 'ln_structured' method, we systematically reduce model complexity. Structured pruning aligns with the broader paradigm of model compression, targeting entire structured components. Using a norm-based approach, exemplified by 'std' (standard deviation), allows the removal of less critical elements, streamlining the neural network architecture. This strategic reduction in parameters aims to balance computational efficiency and model performance, addressing real-time deployment requirements in resource-constrained environments.

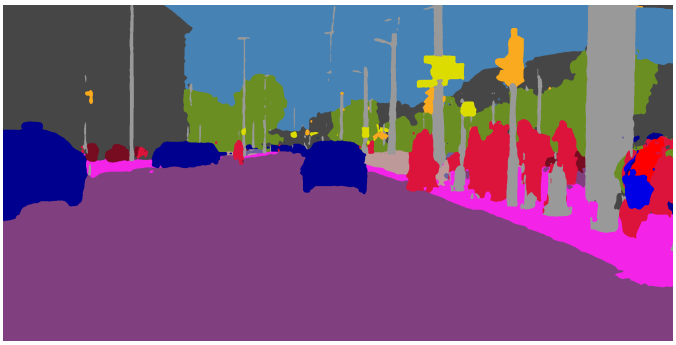

Fig. 4. Segmentation mask for pruned ICNet after fine tuning

C. Implementation

To enhance the efficiency gains, we initiated our approach with a pretrained model, leveraging the knowledge encoded in weights obtained from prior training. Subsequently, we applied structured pruning to this pretrained model, removing non-critical elements and optimizing its architecture. The pruned model was then fine-tuned on the target task, allowing the neural network to adapt to the specific nuances of the dataset. This two-step process, combining structured pruning with fine-tuning, ensures that the model not only retains efficiency gains but also maintains or improves performance on the target task. Our application of structured pruning, as demonstrated in this comprehensive methodology, contributes to the ongoing pursuit of creating compact yet high-performing deep learning models.

IV. Experimental Results

In this section, we present empirical findings from applying our structured pruning and fine-tuning methodology to neural network models. Our experiments assess the impact of structured pruning on model efficiency and subsequent fine-tuning for task-specific adaptation. We provide an overview of the dataset used and its relevance to the chosen neural network architecture. Quantitative and qualitative outcomes of structured pruning are analyzed, including the reduction in model parameters and its influence on inference speed. The fine-tuning phase is explored, highlighting how pruned models adapt to specific semantic segmentation tasks. Comprehensive performance metrics, such as mean IoU, provide a holistic understanding of the trade-offs involved in achieving efficiency gains through structured pruning. These results offer valuable insights into the practical implications and effectiveness of our methodology, guiding a nuanced discussion on the interplay between model compression, task adaptation, and overall performance.

A. Dataset: Unraveling Complexity with Cityscapes

The Cityscapes dataset is a pivotal resource in computer vision research, offering a robust benchmark for semantic segmentation tasks. With 5,000 high-resolution images capturing diverse urban environments across European cities, it provides meticulous pixel-level annotations for crucial object classes like vehicles, pedestrians, and buildings. Focusing on 19 out of 30 available classes in our training, we tailor the model to our target task. The dataset's emphasis on real-world challenges,

Fig. 5. Segmented mask predicted from the pretrained bisenetv2 model

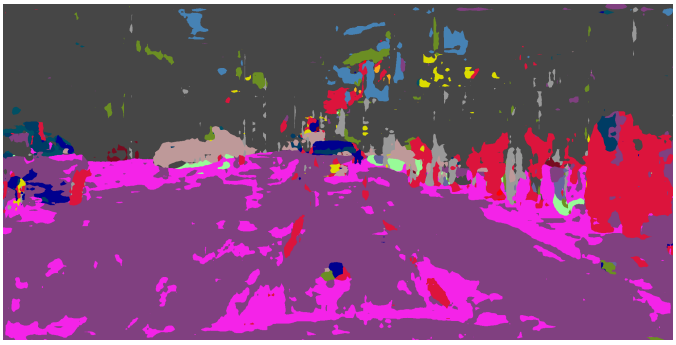

Fig. 6. Segmented mask from the pruned bisenetv2 model

such as occlusions and dynamic elements, makes it invaluable for evaluating computer vision models in autonomous driving and urban scene understanding. Leveraging Cityscapes in our experiments establishes a realistic and comprehensive evaluation framework, enabling a thorough examination of the structured pruning and fine-tuning methodology's performance in complex urban scenarios.

B. Architectures: Navigating Design Complexity

Our evaluation focuses on a targeted comparison between ICNet and BiSeNetV2, both widely used in real-time semantic segmentation. ICNet employs atrous convolutions for multi-scale context, while BiSeNetV2uses attention mechanisms. Employing structured pruning and fine-tuning, we assess their adaptability and effectiveness in real-world scenarios. The chosen architectures, ICNet and BiSeNetV2, bring unique strengths aligned with our semantic segmentation goals. ICNet atrous convolutions suit detailed scene understanding, especially in high-resolution scenarios. Meanwhile, BiSeNetV2attention mechanisms address nuanced features in urban scenes. This selection not only aligns with our research goals but contributes to a broader understanding of diverse approaches to efficient semantic segmentation. Through rigorous evaluation, we aim to unravel their adaptability in real-world scenarios, shedding light on the nuanced interplay between architecture choice, compression techniques, and overall model performance.

Fig. 7. Segmented mask from the pruned bisenetv2 model after fine tuning

C. Mean Intersection over Union (MIoU): Unveiling Model Accuracy

In our evaluation, we employ the Mean Intersection over Union (MIoU) metric to quantify the accuracy of our pruned and fine-tuned models. This metric captures the intersection of predicted and ground truth segmentation masks, normalized by the sum of their areas, providing a comprehensive measure of segmentation quality. By reporting MIoU across different experiments, we offer a robust assessment of the models' ability to accurately delineate object boundaries. This metric serves as a key benchmark for evaluating the success of our structured pruning and fine-tuning methodology in maintaining or improving model accuracy throughout the experiments.

D. Requirements

Hardware: For effective execution of deep learning tasks, ourresearch utilized a high-performance hardware configuration. Featuring an Intel(R) Xeon(R) Silver 4114 CPU with 10 cores and 20 logical processors, coupled with a dedicated NVIDIA Quadro P4000 GPU boasting 1,048,576-byte VRAM, this system accelerated computations significantly. With 64.0 GB of physical memory and SSD-enabled high-speed data access, this hardware setup played a crucial role in enhancing the efficiency of intricate deep learning models.

Software: PyTorch 1.8.1 serves as the primary deep learning framework, complemented by segmentation-models-pytorch for image segmentation, torchmetrics for model evaluation, albumentations for image preprocessing, loguru for logging, and tqdm for informative progress bars. Together, these libraries contribute to the robustness and efficiency of my research project.

E. Qualitative Analysis

In the Qualitative Analysis of the explored architectures, pruning demonstrated a substantial impact on ICNet. As illustrated in Figures 3 and 4, the segmentation masks generated for ICNet after pruning with fine-tuning closely resemble those of the original trained model, indicating that the pruning process effectively maintains the quality of the segmentation results. However, the scenario is notably different for BiSeNetV2. As seen in Fig. 6, the segmentation mask obtained after pruning exhibits a significant loss of crucial information, rendering it

unsuitable for real-world applications. To compensate for this information loss, the fine-tuning step attempts to recover as seen in Fig. 7, by updating the zero parameters, resulting in low sparsity. This observation raises a crucial point – pruning with fine-tuning might not be an optimal technique for models with already low parameter counts. The structured pruning employed removes vital information, and fine-tuning attempts to readjust the model, ultimately diminishing the induced sparsity. This observation underscores the necessity for thoughtful consideration of model characteristics when employing pruning with fine-tuning techniques.

TABLE I
COMPARATIVE PERFORMANCE METRICS FOR ICNET AND BISENETV2

Model	**Sparsity Induced**	**FPS**	**mIoU (Original)**	**mIoU (Ours)**
ICNet	0.11	31	67.7	67.62
BiSeNetV2	0.003	54	73.4	67.49

F. Quantitative Analysis

We conducted a series of experiments on various architectures, including ICNet and BiSeNetV2,focusing on real-time semantic segmentation. The pruning process, set at a sparsity level of 0.25, was followed by fine-tuning, resulting in a final sparsity for each model. Table I provides a comparative overview of the efficiency achieved through our proposed pruning and fine-tuning methodology. Notably, the methodology demonstrated greater effectiveness for ICNet, achieving a model compression with an induced sparsity of 0.11. In contrast, for BiSeNetV2, the induced sparsity was negligible. The impact on mean Intersection over Union (mIoU) revealed that, for ICNet, the speed-accuracy trade-off allowed us to match the original model accuracy. However, for BiSeNetV2, the original model's mIoU proved unrecoverable. Further details on parameter density for both models are outlined in Table II and Table III. These findings emphasize the nuanced trade-offs between model compression and performance metrics, underscoring the importance of carefully tailored pruning strategies for specific architectures.

TABLE II
ICNET MODEL PRUNING RESULTS

Model	**Non-Zero**	**Zero**	**Sparsity**
Original	12881215	0	0
Pruned (s = 0.25)	9666239	3214976	0.25
Fine-tuning	114492688	1431947	0.11

TABLE III
BISENEV2 MODEL PRUNING RESULTS

Model	**Non-Zero**	**Zero**	**Sparsity**
Original	2576409	0	0
Pruned (s = 0.25)	1951209	625200	0.25
Fine-tuning	2568333	8076	0.003

V. FUTURE WORK

Our current research illuminates the implications of structured pruning and fine-tuning in real-time semantic segmentation. Acknowledging the constraints that shaped our exploration, including computational capacity and time limitations, we recognize the need for future work to overcome these challenges. The constrained scope prevented a thorough examination of potential refinements, such as broader evaluations across diverse architectures, extensive exploration of pruning criteria variations, and application to diverse datasets.

Looking forward, increased computational resources and extended time frames can unlock new dimensions for exploration. Firstly, extending the application of our methodology across a broader spectrum of architectures would offer insights into its generalizability. Exploring variations in pruning criteria and fine-tuning strategies tailored to specific model characteristics could enhance the adaptability of the methodology. Moreover, diversifying the evaluation to include datasets beyond Cityscapes would provide a more comprehensive understanding of its robustness in varied urban scenarios. Investigating transferability across different domains and tasks remains an intriguing avenue for future research. Lastly, considering the dynamic nature of real-world environments, evaluating pruned and fine-tuned models under varying conditions—such as lighting changes, weather scenarios, or different urban landscapes—would offer a more realistic assessment of their utility in practical applications.

Addressing these aspects in future research endeavors will contribute to refining and expanding the applicability of structured pruning and fine-tuning methodologies in real-time semantic segmentation and beyond.

VI. CONCLUSION

In conclusion, our research has introduced and explored the application of structured pruning and fine-tuning methodologies in the realm of real-time semantic segmentation. Leveraging these techniques, we aimed to enhance model efficiency without compromising accuracy, addressing the inherent challenges posed by computational constraints in dynamic urban environments. Through rigorous evaluations on architectures like ICNet and BisenetV2, our methodology demonstrated nuanced outcomes, revealing its effectiveness in certain scenarios. However, we acknowledge the impact of resource limitations on the breadth of our exploration. Despite these constraints, our work provides valuable insights into the interplay between model compression, task adaptation, and overall performance. Moving forward, overcoming resource limitations and refining our approach will unlock new avenues for advancing the efficiency and adaptability of deep learning models in real-world applications.

ACKNOWLEDGMENT

We extend our sincere gratitude to our advisor, Dr. Wujie Wen, for invaluable guidance and support. Our heartfelt thanks go to our colleagues for their collaborative efforts and support

throughout this research. We appreciate the institution for providing essential resources, and the open-source community for valuable tools. Special recognition is given to our friends and family for their unwavering understanding and encouragement, contributing significantly to the success of this project

REFERENCES

[1] Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J. (2018). ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11207. Springer, Cham. https://doi.org/10.1007/978-3-030-01219-9_25.

[2] Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., and Sang, N. (2021). Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision, 129, 3051-3068.

[3] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., and Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 325-341).

[4] 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 Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[5] Guo, Y., Liu, Y., Georgiou, T., and Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. International journal of multimedia information retrieval, 7, 87-93.

[6] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19 (pp. 424-432). Springer International Publishing.interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Paszke, A., Chaurasia, A., Kim, S., and Culurciello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147.

[8] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

[9] Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., and Hajishirzi, H. (2018). Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In Proceedings of the european conference on computer vision (ECCV) (pp. 552-568).

[10] Yurtkulu, S. C., Şahin, Y. H., and Unal, G. (2019, April). Semantic segmentation with extended DeepLabv3 architecture. In 2019 27th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[11] Han, S., Mao, H., and Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149.

[12] Shimoda, M., Sada, Y., amd Nakahara, H. (2019). Filter-wise pruning approach to FPGA implementation of fully convolutional network for semantic segmentation. In Applied Reconfigurable Computing: 15th International Symposium, ARC 2019, Darmstadt, Germany, April 9–11, 2019, Proceedings 15 (pp. 371-386). Springer International Publishing.

[13] He, W., Wu, M., Liang, M., and Lam, S. K. (2021). Cap: Context-aware pruning for semantic segmentation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 960-969).

[14] Anwar, S., Hwang, K., and Sung, W. (2017). Structured pruning of deep convolutional neural networks. ACM Journal on Emerging Technologies in Computing Systems (JETC), 13(3), 1-18.

[15] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.