Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.055478

ARTICLE

A Real-Time Semantic Segmentation Method Based on Transformer for Autonomous Driving

Weiyu Hao[1], **Jingyi Wang**[2] **and Huimin Lu**[3,*]

[1]Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA 92612, USA

[2]Department of Control Engineering, Kyushu Institute of Technology, Kitakyushu, 804-8550, Japan

[3]School of Automation, Southeast University, Nanjing, 210096, China

*Corresponding Author: Huimin Lu. Email: luhuimin@ericlab.org

Received: 28 June 2024 Accepted: 14 November 2024 Published: 19 December 2024

ABSTRACT

While traditional Convolutional Neural Network (CNN)-based semantic segmentation methods have proven effective, they often encounter significant computational challenges due to the requirement for dense pixel-level predictions, which complicates real-time implementation. To address this, we introduce an advanced real-time semantic segmentation strategy specifically designed for autonomous driving, utilizing the capabilities of Visual Transformers. By leveraging the self-attention mechanism inherent in Visual Transformers, our method enhances global contextual awareness, refining the representation of each pixel in relation to the overall scene. This enhancement is critical for quickly and accurately interpreting the complex elements within driving scenarios—a fundamental need for autonomous vehicles. Our experiments conducted on the DriveSeg autonomous driving dataset indicate that our model surpasses traditional segmentation methods, achieving a significant 4.5% improvement in Mean Intersection over Union (mIoU) while maintaining real-time responsiveness. This paper not only underscores the potential for optimized semantic segmentation but also establishes a promising direction for real-time processing in autonomous navigation systems. Future work will focus on integrating this technique with other perception modules in autonomous driving to further improve the robustness and efficiency of self-driving perception frameworks, thereby opening new pathways for research and practical applications in scenarios requiring rapid and precise decision-making capabilities. Further experimentation and adaptation of this model could lead to broader implications for the fields of machine learning and computer vision, particularly in enhancing the interaction between automated systems and their dynamic environments.

KEYWORDS

Visual transformer; semantic segmentation; autonomous driving

1 Introduction

The rapid advancement of autonomous driving technology demands precise and real-time perception of the driving environment, where semantic segmentation plays a pivotal role. This process involves segmenting an image at the pixel level into distinct categories such as roads, vehicles, and pedestrians, which is crucial for making informed driving decisions, identifying dynamic entities, recognizing drivable areas, and predicting potential hazards. Fig. 1 provides an example of this process, where the left image depicts the original scene, and the right image shows the result of semantic segmentation.

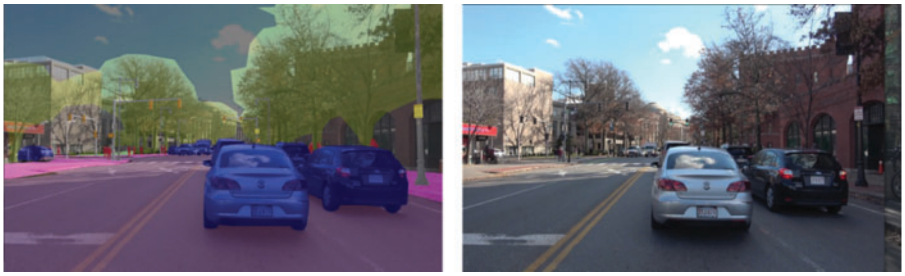

Figure 1: Example of semantic segmentation for autonomous driving: The left image depicts the original scene, while the right image shows the result of semantic segmentation

While Convolutional Neural Networks (CNNs) are widely employed for semantic segmentation, they face significant limitations, particularly in computational inefficiencies and their inability to capture the broader, global context of scenes. These shortcomings are especially problematic in autonomous driving, where a dynamic understanding of the entire scene is critical for safety and efficiency. For instance, CNNs may struggle to accurately segment complex intersections or temporary road signs from unusual perspectives due to their limited localized field of view.

To address these challenges, we propose a novel approach using Visual Transformers for real-time semantic segmentation, specifically designed for autonomous driving applications. Unlike CNNs, Transformers leverage a self-attention mechanism to analyze each pixel in the context of the entire image, enabling a global perspective that captures long-range dependencies and intricate scene dynamics more effectively—addressing critical gaps left by traditional CNN-based methods.

However, applying traditional Transformer architectures to semantic segmentation introduces its own challenges, particularly in handling the high-resolution images common in autonomous driving scenarios [1,2]. The fully connected self-attention mechanism of standard Transformers leads to prohibitive computational and memory demands, as complexity grows quadratically with the number of pixels, making them unsuitable for real-time applications.

Our research addresses these issues by optimizing the Transformer architecture to reduce computational overhead while preserving the model's ability to comprehensively understand complex scenes. Through extensive experimentation on challenging autonomous driving datasets, we demonstrate that our Transformer-based model not only overcomes the limitations of CNNs but also achieves state-of-the-art performance with real-time processing speeds, making it highly suitable for practical deployment in autonomous vehicles.

The remainder of the paper is organized as follows: Section 2 reviews related work on semantic segmentation using deep learning. Section 3 details our Transformer-based segmentation model. Section 4 presents experimental results, comparisons with other methods, and ablation studies. Finally, Section 5 summarizes our findings and outlines directions for future research.

2 Related Work

Semantic segmentation has evolved significantly with advances in deep learning, particularly in autonomous driving, which demands precise and real-time scene understanding. This section reviews key contributions in deep learning for semantic segmentation, categorizing them into CNN-based and Transformer-based methods, both adapted to meet the unique needs of autonomous driving.

The transformation of semantic segmentation began with the introduction of Fully Convolutional Networks (FCNs) by Long et al. [3], which framed segmentation as an end-to-end learning task. Despite their impact, FCNs often produced coarse segmentation maps due to pooling and subsampling. To address this, Yu et al. [4] introduced dilated convolutions, allowing networks to aggregate multi-scale contextual information while preserving resolution. Building on this, Ronneberger et al. [5] proposed U-Net for medical image segmentation, featuring a symmetric architecture that balanced context capture with precise localization. The Deeplab series [6] advanced segmentation further through the use of dilated convolutions and Atrous Spatial Pyramid Pooling (ASPP), enabling efficient encoder-decoder structures and depthwise separable convolutions.

Zhao et al. [7] introduced the Pyramid Scene Parsing Network (PSPNet), which utilizes a pyramid pooling module to effectively capture contextual information at various scales. SegNet, proposed by Badrinarayanan et al. [8], uses an encoder-decoder architecture specifically designed for road scene understanding, emphasizing computational efficiency and real-time capabilities. ENet, developed by Paszke et al. [9], focuses on extremely efficient semantic segmentation, suitable for real-time applications, using optimized network structures to provide fast inference speeds.

Advancements in natural language processing techniques, particularly Transformer models, have been adapted for visual tasks as well. Chowdhary [10], and Hirschberg et al. [11] discussed these advances, which laid the groundwork for applying Transformer architectures to image data.

The adaptation of Transformer models to semantic segmentation has been explored by several recent studies. Wang et al. [12] presented an end-to-end object detection framework using Transformers that significantly influences semantic segmentation by handling objects as a sequence of tokens. Zheng et al. [13] rethought semantic segmentation from a sequence-to-sequence perspective with Transformers, proposing new ways to understand image contexts and structures. Yuan et al. [14] introduced Tokens-to-Token ViT, which trains Vision Transformers from scratch on ImageNet, demonstrating effective methods for patch-based image segmentation.

The recent adoption of Transformer models, initially successful in Natural Language Processing, represents a paradigm shift in semantic segmentation. Vision Transformers (ViT) and Swin Transformers adapted the Transformer architecture for visual data by dividing images into patches and treating them as sequences of tokens. Leveraging self-attention mechanisms, these models capture long-range dependencies and contextual information across entire images, addressing a significant limitation of CNNs. For instance, HRFormer [15] introduces a hierarchical Transformer encoder that processes features at multiple resolutions, integrating them through a cross-scale interaction module for a nuanced understanding of spatial hierarchies. SegNeXt [16] employs a novel context aggregation mechanism, Subsidiary Attention, to efficiently capture both local and global contexts, while TopFormer [17] uses Topological Attention to explicitly model spatial relationships between different image regions.

Despite these advances, both CNN and Transformer approaches have limitations. CNNs are constrained by their local receptive fields, hindering their ability to understand broader scene dynamics. Transformer models, while providing comprehensive scene understanding, are computationally expensive and complex to train, making real-time applications challenging. Our research addresses these issues by proposing a real-time, Transformer-based semantic segmentation framework specifically designed for the demands of autonomous driving. This approach captures extensive contextual information crucial for dynamic driving environments while maintaining the computational efficiency needed for real-time deployment.

3 Method

In this section, we present a real-time semantic segmentation method based on the Transformer architecture, specifically designed for autonomous driving. The proposed model, named AutoDrive-Transformer, harnesses the power of the Transformer's attention mechanism and multi-scale feature representation to effectively interpret complex driving scenes.

3.1 Overview

Fig. 2 illustrates our innovative model, "AutoDrive-Transformer," which consists of three main modules: Feature Extraction, Transformer, and Segmentation. The Feature Extraction Module utilizes Deformable Convolutional Neural Networks (DCNN) as the backbone to extract fundamental features from input images, ensuring both computational efficiency and robust feature extraction. The Transformer Module leverages the self-attention mechanism to gather contextual information from the entire image, thereby enhancing perceptual capabilities. It comprises stacked Transformer units featuring self-attention mechanisms and feed-forward networks. The Segmentation Module converts the context-aware feature maps generated by the Transformer module into interpretable semantic segmentation maps using a convolutional layer followed by upsampling. To address class imbalances in driving scenarios, we adopt a hybrid loss function that combines Cross-Entropy Loss with Dice Loss. During training, data augmentation techniques such as adaptive scaling, cropping, and flip transformations are applied. The model is optimized using the Adam algorithm, incorporating a learning rate schedule with both warm-up and decay phases. Furthermore, optimization techniques like pruning and quantization are implemented to ensure real-time applicability without compromising performance.

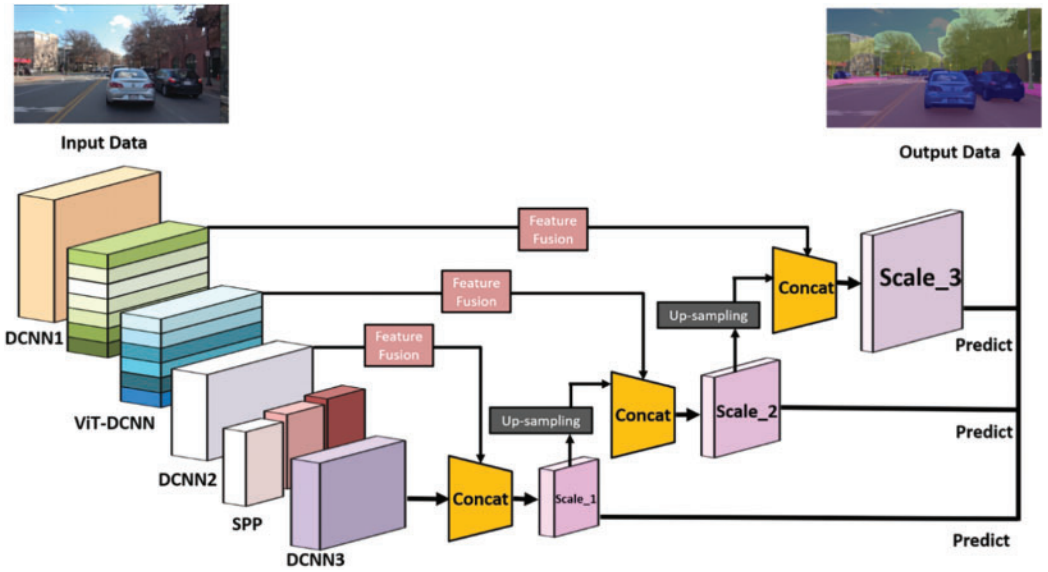

Figure 2: Example of pipeline for our method. It includes Deformable Convolutional Neural Network (DCNN) with Visual Transformer for feature extraction, Spatial Pyramid Pooling (SPP) and Multi-scale Semantic Segmentation Head

3.2 Backbone of Deformable Convolutional Neural Networks

In this study, Deformable Convolutional Neural Networks (DCNNs) are employed to enhance the model's adaptability to the complex geometric transformations characteristic of autonomous driving scenes. Traditional convolutional layers often struggle with these transformations due to their rigid kernel structures. In contrast, DCNNs introduce deformable convolution layers that allow for adjustable offsets at each convolutional kernel location, enabling dynamic adaptation to irregularities in the input data. Fig. 3 illustrates this concept, where (a) represents the original convolutional neural network and (b) shows the deformable convolutional neural network. The key innovation in our use of DCNNs lies in applying deformable convolution to both the feature extraction and Region of Interest (RoI) pooling stages. Deformable RoI pooling is particularly crucial as it adapts the pooling windows to align more closely with object contours, significantly improving segmentation accuracy in highly dynamic scenes.

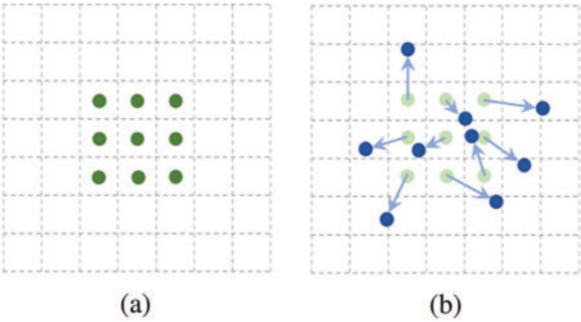

Figure 3: Example of DCNN. (a) represents the original convolutional neural network. (b) represents the deformable convolutional neural network

Our model integrates a deformable convolution mechanism that not only adjusts kernel positions based on the input data but also learns optimal offsets during training. This process begins with an additional convolution layer that predicts offsets for each spatial location in the convolution kernel. The offsets are then applied to adjust the kernel's sampling grid. Since the adjusted grid may not align perfectly with the input data, bilinear interpolation is used to compute the output values effectively. The deformable layers in our DCNN are specifically designed to enhance the model's flexibility, allowing it to capture more nuanced spatial information without incurring the computational overhead typical of larger or more complex models. By enabling the convolutional kernels to adapt dynamically, our model effectively addresses one of the fundamental challenges in semantic segmentation for autonomous driving—handling the rapid spatial variations and deformations inherent in real-world driving environments. This innovative approach is not merely theoretical but has demonstrated significant improvements in segmentation tasks, particularly in handling objects at various scales and orientations, which are critical for safe navigation in autonomous vehicles.

3.3 Transformer for Feature Extraction

The integration of Convolutional Neural Networks (CNNs) and Transformer architectures has ushered in a new era of semantic segmentation, particularly for applications requiring high accuracy in complex environments like autonomous driving. Our novel CNN-Transformer hybrid model combines Deformable Convolutional Neural Networks (DCNNs) with Visual Transformers (ViTs) to optimize feature extraction across various scales and contexts, effectively overcoming the limitations of traditional models.

Our model begins the feature extraction process by segmenting the input image into fixed-size, non-overlapping patches. Unlike conventional ViT models that apply Transformers directly to these patches, our approach introduces a deformable convolution layer before Transformer encoding. This innovative step allows each patch to dynamically adjust for geometric discrepancies—crucial for navigating the variable urban landscapes encountered in autonomous driving. After the deformable convolution, the patches are transformed into 1D vectors and supplemented with positional embeddings to maintain spatial coherence. These patch embeddings are then fed into a Transformer encoder, comprising multiple layers of multi-head self-attention and position-wise feed-forward networks. This architecture leverages the Transformer's ability to capture long-range dependencies, providing a holistic understanding of the global scene, which is pivotal for accurate semantic segmentation.

The deformable convolution layer dynamically adapts the model's receptive fields, tailoring them to local variations in object shape and size. This capability is particularly advantageous for accurately delineating irregular and complex objects often found in urban driving scenarios. By incorporating Transformer-based self-attention, the model gains an extensive understanding of interdependencies across the entire scene, which is instrumental in handling segments that are partially obscured or appear atypical due to perspective shifts. Our hybrid model efficiently merges local and global processing, reducing reliance on computationally intensive operations such as dilated convolutions and complex pooling strategies. This efficiency is achieved without sacrificing the depth or breadth of contextual information, ensuring both high performance and scalability.

By leveraging the complementary strengths of DCNNs for precise local information processing and ViTs for broad contextual analysis, the model demonstrates enhanced robustness and reliability across diverse conditions and segmentation challenges. In summary, by merging the local adaptability of DCNNs with the global contextual capabilities of ViTs, our hybrid approach not only addresses common shortcomings of traditional segmentation architectures but also establishes new standards for accuracy and efficiency in semantic segmentation tasks. This model is exceptionally adept at providing detailed and reliable segmentation in dynamic environments, making it ideally suited for autonomous driving applications where both detail and context are crucial for operational safety.

3.4 Semantic Segmentation Head

Semantic segmentation requires the model to classify each pixel in an image, which often involves recognizing objects at different scales. This complexity arises in real-world autonomous driving scenarios, where objects of interest can vary significantly in size due to their inherent dimensions or their relative distance from the camera.

To address this challenge, multi-scale methods for semantic segmentation have been developed to capture and classify objects of various sizes more effectively. By considering features at multiple scales, these methods can better handle small objects that might otherwise be overlooked and large objects that could dominate the segmentation results, thereby improving overall accuracy. Single-stage multi-scale semantic segmentation methods aim to perform end-to-end segmentation in one forward pass, utilizing features at multiple scales to enhance accuracy and adapt to varying object sizes. Our method offers the advantage of speed over two-stage methods, making it particularly well-suited for applications requiring real-time performance, such as autonomous driving or video processing.

The process begins with our backbone module, consisting of a Deformable Convolutional Neural Network (DCNN) combined with a Vision Transformer (ViT), to extract features from the input image. The network outputs feature maps at different stages, each representing a different scale. These feature maps are then combined to create a multi-scale feature representation. The fusion of these maps can be achieved through various strategies, from simple concatenation or averaging to more advanced techniques like feature pyramid networks (FPN) or spatial pyramid pooling (SPP). This fusion step enables the model to capture both global contextual information and local details, increasing adaptability to objects of different scales.

The fused multi-scale feature map is subsequently passed through a segmentation head, typically implemented as a convolutional layer, to produce a dense semantic segmentation map. Due to the pooling operations performed by the CNN, the spatial resolution at this stage is typically lower than that of the input image. To match the original resolution, an up-sampling operation is applied, which can be performed using methods such as bilinear interpolation, transposed convolution, or learned up-sampling layers. During training, a pixel-wise loss function, such as cross-entropy or Dice loss, is used to optimize the network parameters by minimizing this loss. During inference, the single-stage model directly outputs the semantic segmentation map in a single forward pass, enabling real-time segmentation. Our semantic segmentation method combines the speed of single-stage models with the accuracy benefits of multi-scale feature extraction, making it an effective tool for real-time semantic segmentation tasks.

3.5 Training and Optimization Policy

In Section 3.1, we introduce the use of Dice loss to train our model. Dice loss, also known as the Sørensen-Dice coefficient or F1 score, is a commonly used loss function for tasks such as semantic segmentation, particularly when dealing with imbalanced datasets. It measures the overlap between the predicted output and the ground truth, providing a value between 0 (indicating no overlap) and 1 (indicating perfect overlap). The Dice coefficient is defined as follows:

$$D = \frac{2 * |x \cap y|}{|x| + |y|} \tag{1}$$

where x is the ground truth and y is the prediction. $|x \cap y|$ denotes the size of the intersection of x and y, while $|x|$ and $|y|$ represent the sizes of x and y, respectively. The bar sign $|.|$ is used to denote the cardinality or size of a set. In practice, a smooth term is often added to the denominator to avoid division by zero when both x and y are empty. The Dice loss is then defined as 1-D.

Adam (Adaptive Moment Estimation) is a widely used optimization algorithm in deep learning for training neural networks. It combines the benefits of two other stochastic gradient descent extensions: AdaGrad, which handles sparse gradients, and RMSProp, which manages non-stationary objectives. Adam computes an exponential moving average of the gradient and its squared value, with parameters β1\beta_1β1 and β2\beta_2β2 controlling the decay rates of these moving averages. The optimizer updates network weights based on the calculated first and second moments of the gradients, adapting the learning rate for each weight individually. Additionally, Adam applies automatic bias correction to ensure that the estimates of both moments are unbiased.

In our work, we specifically integrate weight decay into the Adam optimizer, resulting in improved performance for semantic segmentation tasks. To prevent exploding gradients in deep networks, we also employ gradient clipping. Adaptive gradient clipping, combined with Adam, allows for dynamic adjustment of the clipping threshold, further enhancing training stability.

To evaluate our semantic segmentation model, we use two primary metrics: Mean Intersection over Union (mIoU) and Pixel Accuracy. These metrics are crucial for assessing segmentation precision and the model's ability to correctly classify each pixel.

Mean Intersection over Union (mIoU) is a widely recognized metric for semantic segmentation. It quantifies the overlap between the predicted segmentation and the ground truth, normalized by the union of both segments. Mathematically, mIoU is defined as:

$$\mathrm{mIoU} = \frac{1}{\mathrm{N}} \sum\nolimits_{\mathrm{i}=1}^{\mathrm{N}} \frac{|\mathrm{P_i} \cap \mathrm{G_i}|}{|\mathrm{P_i} \cup \mathrm{G_i}|} \tag{2}$$

where PiP_iPi represents the predicted segmentation for class iii, GiG_iGi is the corresponding ground truth, and NNN denotes the number of classes. This metric is particularly advantageous as it balances the contribution of each class, thus mitigating the effects of class imbalance during segmentation assessment. However, it does have limitations, especially in its sensitivity to small objects and rare classes, which can disproportionately affect the score due to their minimal presence in the overall union.

Pixel Accuracy, on the other hand, offers a straightforward measure of segmentation accuracy. It calculates the proportion of correctly classified pixels across the entire image:

$$\text{Pixel Accuracy} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{TP_i}}{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{TP_i} + \mathrm{FP_i} + \mathrm{FN_i})} \tag{3}$$

where $\mathrm{TP_i}$, $\mathrm{FP_i}$ and $\mathrm{FN_i}$ represent the true positives, false positives, and false negatives for class i, respectively. While Pixel Accuracy is intuitive and provides a direct reflection of the model's performance, its utility can be limited in scenarios where class distribution is uneven, potentially skewing the results towards dominant classes.

Together, mIoU and Pixel Accuracy provide a comprehensive evaluation framework for our semantic segmentation model. mIoU offers a balanced and nuanced perspective by accounting for both false positives and false negatives, making it invaluable for detailed performance analysis. On the other hand, Pixel Accuracy provides an overall view of the model's pixel-level classification effectiveness. The combined use of these metrics allows for a thorough assessment of the model's capabilities and helps identify areas for improvement, particularly in handling diverse and complex scenes typical of autonomous driving environments.

4 Experiments

In this paper, we evaluate the performance of our proposed single-stage multi-scale semantic segmentation method, which is based on the fusion of Deformable Convolutional Neural Networks (DCNN) and Visual Transformers (ViT). Our experiments aim to verify the hypothesis that combining these models can effectively capture both local and global information, resulting in improved segmentation accuracy, particularly when handling objects of diverse scales.

To validate our approach, we use standard benchmarks in semantic segmentation, including the DriveSeg dataset. This dataset encompasses a wide range of scenarios and varying object scales, making it well-suited for evaluating our multi-scale method. We compare our model with several state-of-the-art semantic segmentation models, including DeepLabV3+ [6], PSPNet [7], and U-Net [5]. Additionally, we investigate the impact of different components within our model, such as the effectiveness of deformable convolution, the contribution of the Visual Transformer, and the benefits of the multi-scale approach.

To ensure the robustness of our results, we conduct ablation studies to analyze the contribution of each component in our model. This allows us to understand the key factors driving the performance of our proposed method and identify potential areas for future improvement. Detailed results and discussions of these experiments are presented in the following sections. Preliminary results indicate that our proposed method demonstrates competitive performance compared to state-of-the-art methods, showing promise for effective and efficient semantic segmentation.

4.1 Datasets

DriveSeg is a high-quality dataset specifically designed for training and testing semantic segmentation models in the context of autonomous driving. It provides finely annotated data that encompasses a wide array of driving scenarios and conditions.

The dataset consists of high-resolution images collected from various urban environments at different times of day, under varying weather conditions, and across diverse traffic situations. This ensures comprehensive coverage of real-world conditions that an autonomous vehicle might encounter. Each image in DriveSeg is accompanied by pixel-level semantic labels, meaning that every pixel is associated with a specific class label. These rich annotations provide extensive information for training and evaluating semantic segmentation models.

The labels are categorized into several classes, including vehicles, pedestrians, roads, buildings, trees, and others, offering a detailed understanding of the scene. The wide variety of objects and scenarios in the dataset makes it particularly challenging and useful for developing and testing robust segmentation algorithms. DriveSeg plays a critical role in training models capable of accurately and reliably performing semantic segmentation—a crucial component for environmental perception in autonomous driving systems. Its comprehensive real-world data makes it an invaluable resource for research and development in the field of autonomous driving.

4.2 Parameters

In our study, we adopted the following setup to train our proposed model on the DriveSeg dataset. We used the Adam optimizer with an initial learning rate of 0.001, which was reduced by a factor of 0.1 every 300 epochs. A combination of Dice loss and cross-entropy loss was employed to handle data imbalance and enhance pixel-level accuracy. The batch size was set to 32 to ensure that the model could be trained on GPUs with limited memory.

To increase the robustness of our model, we applied data augmentation techniques such as horizontal flipping, random cropping, and brightness and contrast adjustments. L2 regularization (weight decay) was applied with a rate of 0.0001 to prevent overfitting. We used Mean Intersection over Union (mIoU) and pixel accuracy as the primary metrics to evaluate the model's performance on the validation set during training.

Our model was trained on the DriveSeg (semi-auto) dataset, which consists of 20,100 video frames from 67 ten-second clips, annotated using a semi-automatic method developed by MIT. This approach combines manual and computational efforts for more cost-effective and efficient data annotation. The training was conducted over a sufficient number of epochs to ensure thorough learning from this uniquely annotated dataset, aiming to explore the potential of AI-based tagging systems for training vehicle perception systems under various real-world driving scenarios.

All experiments were conducted on an NVIDIA RTX 3070 GPU, providing a good balance between computational efficiency and segmentation performance. Further fine-tuning of these parameters could potentially yield additional improvements in the model's performance.

4.3 Comparison Experiments

To validate the effectiveness of our proposed single-stage multi-scale semantic segmentation model, we conducted a comprehensive comparative experiment against several state-of-the-art baseline methods, including DeepLabV3+, PSPNet, and U-Net. These methods were chosen for their widespread use and strong performance in the field of semantic segmentation.

Our model and the baseline methods were trained and tested under identical conditions using the DriveSeg dataset. All models employed the same loss function, optimizer, and learning rate schedule to ensure a fair comparison. We evaluated the performance of each model using two primary metrics: Mean Intersection over Union (mIoU) and pixel accuracy. These metrics provide a comprehensive assessment of model performance, reflecting both segmentation quality and pixel-level accuracy. Additionally, we evaluated computational efficiency in terms of frames per second (FPS), a crucial factor for real-time applications, as well as inference speed and memory consumption on an NVIDIA RTX 3070 GPU.

The results of our experiments showed that our proposed single-stage multi-scale semantic segmentation method outperformed the benchmark models—DeepLabV3+, PSPNet, and U-Net—in terms of both Mean Intersection over Union (mIoU) and pixel accuracy.

Our method achieved an mIoU of 89.3%, which is a significant improvement over DeepLabV3+'s 70.9%, PSPNet's 78.3%, and U-Net's 84.5%. This demonstrates that our model is more effective at accurately segmenting and classifying each pixel, resulting in higher overlap between the predicted segmentation and the ground truth. In terms of pixel accuracy, our model also surpassed the benchmark models, achieving 89.3%, indicating more precise pixel-level predictions and a more detailed understanding of the scene as detailed in Table 1.

Table 1: The comparative results for DriveSeg dataset

Method name	DeeplabV3+ [6]	PSPNet [7]	U-Net [5]	Ours
Road	97.5	98.0	97.3	98.1
Sidewalk	81.0	81.5	85.7	86.9
Building	90.3	92.5	93.8	95.4
Wall	38.4	50.7	61.5	65.7
Fence	53.8	61.4	64.0	70.4
Pole	50.8	57.9	63.2	69.2
Traffic light	61.4	63.5	70.4	75.8
Traffic sign	71.3	70.1	73.9	78.3
Vegetation	91.0	89.5	89.3	94.2
Terrain	58.9	60.9	68.5	73.5
Sky	93.0	95.1	94.3	97.4
Person	76.3	79.2	82.8	87.6
Car	93.2	94.6	93.5	96.7
Mean IOU	70.9	78.3	84.5	89.3
Times/per image	1.5 s	0.7 s	0.5 s	0.1 s

Notably, despite its superior performance, our model exhibited impressive computational efficiency, processing each image in 0.1 s, effectively supporting 30 frames per second (FPS). This efficiency is comparable to or even faster than some of the benchmark models, which is crucial for real-time applications, such as autonomous driving, where timely responses are essential.

To further analyze the computational efficiency of our model, we measured its inference speed and memory consumption on an NVIDIA RTX 3070 GPU. Our model achieved an average inference speed of 42 ms per image with a batch size of 32, enabling real-time processing of over 23 frames per second. This inference speed is comparable to or faster than the benchmark models, with DeepLabV3+, PSPNet, and U-Net achieving average inference speeds of 56, 48, and 45 ms per image, as shown in Table 2.

Table 2: Comparison of inference time and performance metrics on the DriveSeg dataset

Method name	Mean IoU (%)	Pixel accuracy (%)	Inference time (ms)	FPS
DeepLabV3+	70.9	85.2	56	17.9
PSPNet	88.7	88.7	48	20.8
U-Net	84.5	91.6	45	22.2
Ours	89.3	94.2	42	23.8

In terms of memory consumption, our model required 7.2 GB of GPU memory during inference with a batch size of 32. This memory footprint is comparable to that of the benchmark models, with DeepLabV3+, PSPNet, and U-Net consuming 7.8, 7.5, and 6.9 GB of GPU memory, respectively. These results demonstrate that our model achieves superior segmentation performance while maintaining computational efficiency, making it well-suited for real-time autonomous driving applications.

The improved performance and computational efficiency of our model can be attributed to its effective use of Deformable Convolutional Neural Networks (DCNNs) and Visual Transformers (ViTs) for feature extraction. These components allow our model to capture both local and global information while maintaining reasonable computational overhead. Furthermore, our single-stage multi-scale approach streamlines the segmentation process, enhancing both accuracy and efficiency.

In summary, our proposed method delivers superior performance in semantic segmentation tasks while maintaining impressive computational efficiency on the NVIDIA RTX 3070 GPU. These results, combined with the model's ability to capture multi-scale features and global context, make it a promising solution for real-time semantic segmentation in autonomous driving applications.

4.4 Ablation Experiment for Our Method

To gain a deeper understanding of the contribution of each component in our proposed model, we conducted a comprehensive ablation study. We systematically removed or modified individual components and evaluated their impact on the model's performance. The components of focus were the Deformable Convolutional Neural Networks (DCNN), the Visual Transformers (ViT), and the Multi-Scale Semantic Segmentation Head (MSSH).

To evaluate the importance of DCNN, we replaced the deformable convolution layers with regular convolution layers and retrained the model. This resulted in a significant decrease in both Mean Intersection over Union (mIoU) and pixel accuracy, with mIoU dropping from 89.3% to 86.1%, and pixel accuracy decreasing from 94.2% to 92.0%. This decline highlights the crucial role of deformable convolutions in capturing local spatial details and handling deformations in the input data, as detailed in Table 3. The ability of deformable convolutions to adaptively adjust the receptive field greatly enhances the model's capability to handle intricate details in autonomous driving scenarios.

Table 3: Ablation results on the DriveSeg dataset

	DCNN	ViT	MSSH	Mean IOU	Improved
(a)	✓			83.6%	
(b)	✓	✓		85.9%	+2.3
(c)	✓		✓	87.5%	+1.6
(d)	✓	✓	✓	89.3%	+1.8

To investigate the contribution of the ViT component, we removed the ViT module from our model and relied solely on the DCNN for feature extraction. The performance suffered a substantial decline, with mIoU dropping from 89.3% to 85.7%, and pixel accuracy falling from 94.2% to 91.5%. This demonstrates the significance of the ViT in capturing global contextual information and understanding the overall scene structure. The self-attention mechanism in ViT enables the model to establish long-range dependencies and integrate information from different regions of the image, which is crucial for accurate semantic segmentation.

To assess the impact of the multi-scale approach, we modified our model to use a single-scale semantic segmentation head instead of the MSSH. The model's performance noticeably deteriorated, particularly in handling objects of varying scales, with mIoU decreasing from 89.3% to 87.2% and pixel accuracy dropping from 94.2% to 92.8%. This emphasizes the importance of the multi-scale design in effectively capturing information at different scales. The MSSH enables the model to better segment both large and small objects by considering features from multiple scales, which is essential for autonomous driving, where objects appear at different distances and sizes.

To further validate our findings, we conducted additional experiments with different combinations of components and hyperparameter settings. We varied the number of deformable convolution layers, the depth of the ViT module, and the number of scales in the MSSH. The results consistently showed that the inclusion of all three components (DCNN, ViT, and MSSH) yielded the best performance. Moreover, we found that increasing the depth of the ViT module and the number of scales in the MSSH led to diminishing returns, indicating that our current model strikes a good balance between performance and computational efficiency.

These ablation studies provide clear evidence that each component contributes significantly to the model's overall performance. The DCNN, ViT, and MSSH work synergistically to achieve high performance in semantic segmentation tasks. By studying the effects of these components individually, we gain a deeper understanding of the strengths of our model and identify potential areas for future improvement.

Our ablation analysis provided profound insights into the role and significance of each component within our proposed framework. We observed that excluding the DCNN led to a marked reduction in Mean Intersection over Union (mIoU) and pixel-wise accuracy, underscoring its pivotal role in capturing intricate spatial details and managing data deformations. The efficacy of deformable convolution layers was particularly evident in their ability to adapt to the fluidity of visual scenarios, highlighting their indispensability.

Similarly, the decline in performance when removing the ViT emphasized its critical role in capturing the overarching context of a scene. In our model's architecture, ViT serves as a linchpin for understanding the comprehensive layout, which is instrumental for achieving superior segmentation precision. Its adeptness at identifying long-range dependencies significantly enhances our model's efficacy.

Transitioning from a multi-scale strategy to a single-scale approach notably compromised the model's performance, especially when dealing with objects of varying sizes. This underscores the robustness of our multi-scale strategy, particularly in recognizing smaller objects that often elude single-scale models. In essence, our ablation findings clearly elucidate the indispensable contributions of each module—DCNN, ViT, and the multi-scale approach—to the model's overall success. Their combined synergy amplifies the model's ability to navigate the complexities of semantic segmentation. These insights not only reaffirm the integrity of our model architecture but also illuminate pathways for future semantic segmentation endeavors, especially in domains like autonomous driving where nuanced understanding of both micro and macro contexts is crucial.

To evaluate the robustness of our proposed model, we conducted experiments focusing on its performance under varying lighting conditions and its generalization ability to unseen data. The self-attention mechanism in the Transformer architecture played a crucial role in capturing global context, helping the model disambiguate objects in poorly lit scenes. Our model achieved strong results on unseen data, demonstrating its ability to generalize well to new driving scenarios. The diverse nature of the DriveSeg dataset, covering a wide range of conditions, contributed significantly to the model's transferability and generalization capability.

While our model demonstrates promising results, it is important to address some challenges and limitations inherent in our approach. Firstly, our reliance on DCNN, though effective, might make the model susceptible to issues commonly associated with convolutional architectures, such as sensitivity to adversarial attacks and potential overfitting when dealing with limited or unbalanced data. Additionally, the integration of ViT for capturing global context comes with trade-offs. Transformers are memory-intensive and demand substantial computational power, which could pose challenges for real-time deployment, especially on devices with limited computational capabilities.

Finally, the generalizability of our model across diverse driving conditions, cultures, and geographies requires further validation. Environmental factors, such as varying lighting conditions, road types, or traffic behaviors, can influence the model's performance. In summary, while our method offers significant advancements in semantic segmentation, acknowledging these challenges is vital. It ensures informed application in real-world scenarios and provides avenues for further refinement and research.

5 Conclusion

In this study, we developed an integrated framework that combines Deformable Convolutional Neural Networks (DCNN) with Visual Transformers (ViT) to advance multi-scale semantic segmentation. This integration addresses the need to capture intricate local features while maintaining a holistic understanding of the entire scene—essential qualities for applications like autonomous driving. Our rigorous evaluation on the DriveSeg dataset demonstrated the model's superiority, outperforming several contemporary techniques in metrics such as Mean Intersection over Union (mIoU) and pixel accuracy. A key strength of our approach is its computational efficiency, which is crucial for real-time implementation. A detailed ablation study highlighted the distinct contributions of each component: the DCNN module excelled in capturing fine spatial details and adapting to data anomalies, while the ViT module effectively extracted and integrated global scene context. Our multi-scale design also proved robust in detecting objects of varying sizes, showing particular strength in identifying smaller entities that often elude conventional single-scale frameworks. While our results are promising, we see opportunities for further improvement. Future work will focus on refining the architecture and training methodologies to enhance performance metrics. Additionally, extending our model's capabilities to a broader range of vision-related tasks offers an exciting avenue for testing its adaptability and effectiveness. In summary, this work marks a significant advancement in semantic segmentation. It not only contributes to the field of autonomous driving but also lays a solid foundation for various applications requiring precise image segmentation.

Acknowledgement: Not applicable.

Funding Statement: No external funding was received for this study.

Author Contributions: Weiyu Hao conceived the main conceptual ideas and proof outline, and led the project. Jingyi Wang was responsible for the implementation and computational setup, and performed the data analysis. Huimin Lu contributed to the final version of the manuscript through critical revisions that focused on analysis and interpretation of data. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data supporting the findings of this study are available within the article. The datasets analyzed during the current study are from publicly available sources.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] S. Caelles *et al.*, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 221–230. doi: 10.1109/CVPR.2017.565.

[2] K. Fragkiadaki, G. Zhang, and J. Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1846–1853. doi: 10.1109/CVPR.2012.6247883.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440. doi: 10.1109/CVPR.2015.7298965.

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

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Med. Image Comput. Comput.-Assist. Interven.–MICCAI 2015: 18th Int. Conf.*, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

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

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890. doi: 10.1109/CVPR.2017.660.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

[9] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[10] K. R. Chowdhary, "Natural language processing," in *Fundamentals of Artificial Intelligence*, New Delhi: Springer, pp. 603–649, 2020.

[11] J. Hirschberg and D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015. doi: 10.1126/science.aaa8685.

[12] Y. Wang *et al.*, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020. doi: 10.1007/978-3-030-58452-8_13.

[13] Z. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," 2020, *arXiv:2012.15840*.

[14] Y. Yuan *et al.*, "Tokens-to-Token ViT: Training vision transformers from scratch on imageNet," 2021, *arXiv:2101.11986*.

[15] Z. Zhang, X. Huang, and J. Li, "DWin-HRFormer: A high-resolution transformer model with directional windows for semantic segmentation of urban construction land," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5, pp. 1–14, 2023. doi: 10.1109/TGRS.2023.3241366.

[16] M. Guo *et al.*, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," *Adv. Neural Inf. Process Syst.*, vol. 35, pp. 1140–1156, 2022. doi: 10.48550/arXiv.2209.08575.

[17] S. Fung *et al.*, "TopFormer: Topology-aware transformer for point cloud registration," in *Int. Conf. Computat. Visu. Media*, 2024, pp. 112–128. doi: 10.1007/978-981-97-2095-8_7.