Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063507

ARTICLE

A Lane Coordinate Generation Model Utilizing Spatial Axis Attention and Multi-Scale Convolution

Duo Cui* **and Qiusheng Wang**

School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100191, China

*Corresponding Author: Duo Cui. Email: zy2203601@buaa.edu.cn

Received: 16 January 2025; Accepted: 31 March 2025; Published: 09 June 2025

ABSTRACT: In the field of autonomous driving, the task of reliably and accurately detecting lane markings poses a significant and complex challenge. This study presents a lane recognition model that employs an encoder-decoder architecture. In the encoder section, we develop a feature extraction framework that operates concurrently with attention mechanisms and convolutional layers. We propose a spatial axis attention framework that integrates spatial information transfer regulated by gating units. This architecture places a strong emphasis on long-range dependencies and the spatial distribution of images. Furthermore, we incorporate multi-scale convolutional layers to extract intricate features from the images. The two sets of feature maps are concatenated and subsequently transformed into an input sequence for the decoder, with the lane marking coordinates considered as a target sequence for coordinate generation. This decoder can directly segment multiple lane markings, eliminating the need for additional post-processing algorithms, thereby significantly streamlining the lane recognition process. The proposed method demonstrates a high degree of accuracy in recognizing lane markings and exhibits robust capabilities in differentiating between occlusions and objects resembling lanes. It shows exceptional performance on the TuSimple and CULane datasets.

KEYWORDS: Lane detection; autonomous driving; self-attention

1 Introduction

The lane detection task represents a fundamental aspect within the domain of computer vision, extensively used in the context of autonomous driving. The main aim of the lane detection task is to recognize and monitor the lane markings on the roadway to ensure proper vehicle navigation. In light of the intricate conditions of the traffic environment, lane markings, which are defined as elongated strip-like features, frequently face disruptions from occlusions, stains, comparable traffic signs, and various other elements in real-world situations, thereby making the task of lane marking detection notably difficult.

The lane detection task typically consists of two main components: feature extraction and post-processing, and can be understood as an encoder-decoder architecture [1]. Within the encoder component, image processing modules, including Convolutional Neural Networks (CNN) [2–5] and Vision Transformers (ViT) [6], are utilized to extract image semantic information. However, CNNs can only capture local information, which is prone to false positives and missed detections due to numerous occlusions in complex traffic conditions. The attention mechanisms can capture global attention information, while introducing significant computational overhead, which may not meet the real-time requirements for lane detection. Within the decoder component, lane detection methods depend on a series of complex post-processing algorithms, including fitting [7], edge detection [8], and image segmentation [9]. These methods require the

establishment of independent sub-tasks and loss functions, and during the computation process, they may introduce significant errors.

This paper presents a novel model designed to generate coordinate sequences for lane markings, which is termed ACP2Seq (Attention Multi-Scale Convolution Parallel to Sequence). In the encoder component, we develop an image feature extraction framework that operates in conjunction with spatial axis attention [10] and multi-scale convolution. This design enables the model to focus on the intricate details associated with lane markings, thus mitigating the potential for confusion between lane markings and analogous traffic signs that may be difficult to differentiate. This framework maintains the advantages of the attention mechanism in focusing on global information while keeping computational costs low.

In the decoder component, feature map information is employed as the input sequence, while the coordinates of lane markings act as the target sequence, facilitating a text generation task aimed at sequentially producing the coordinates of lane marking points. In a manner similar to the Pix2Seq model [11], the ACP2Seq model generates coordinate information for targets directly from images. This decoder consists of only one single layer of transformer, which not only features fast computation speed but also facilitates the easy implementation of multi-lane segmentation, thereby avoiding the introduction of additional errors.

To summarize, the primary contributions of this paper are outlined as follows: 1) the proposal of a sequence generation-based lane recognition model, ACP2Seq; 2) the introduction of an image feature extraction module that integrates attention layers and convolutional layers in parallel; 3) the development of a spatial axis attention model that incorporates spatial information transfer; 4) the demonstration of the proposed model's exceptional performance on multi-lane datasets, affirming the substantial recognition capabilities of the ACP2Seq model under diverse complex road conditions.

The structure of this paper is as follows: Section 2 will review the relevant literature, focusing on existing lane markings recognition techniques and their limitations. Section 3 will provide a detailed description of the proposed ACP2Seq model, including its architecture, key components, and innovative aspects. Section 4 will describe the experimental setup and present the experimental results, comparing them with existing advanced methods to validate the effectiveness of the proposed approach. Finally, Section 5 will summarize the main findings and contributions of the research.

2 Related Work

Image processing tasks are typically categorized into two components: the encoder and the decoder. The encoder component is primarily utilized for feature extraction tasks, typically employing convolution, attention mechanisms, and their enhancement methods. The decoder component typically performs subsequent processing tasks in accordance with the specified task objectives.

A variety of methodologies for the extraction of features from lane images are already present within the encoder component. In addition to widely recognized models such as U-Net [12] and ENet [13], numerous researchers have implemented specific enhancements to address the distinctive semantic characteristics of long-lane markings. This includes models such as SCNN [14], RESA [15], and CondLaneNet [16], which employ spatial convolution techniques to facilitate the propagation of information across rows and columns. These methods significantly enhance the capability of lane line recognition, thereby affirming the importance of spatial characteristics in the task of lane markings detection. Nonetheless, the ability of CNN to extract local features via the application of convolutional kernels is constrained in modeling long-range dependencies. Their recognition capability diminishes when confronted with targets that exhibit multiple occlusions.

In response to the aforementioned challenges, the transformer model [17] has been proposed. The transformer model is capable of capturing global dependencies across all positions within the input sequence, which allows it to demonstrate significant processing capabilities for images characterized by long-range dependencies. Dosovitskiy et al. [6] transformed images into patches to enable global attention computation, whereas Liu et al. [18] validated the improved capabilities of attention mechanisms as fundamental networks in image processing. Nevertheless, the computational expense associated with attention mechanisms is considerably elevated when dealing with large input images. In order to decrease the quantity of model parameters, local attention constraints [19] and axis attention mechanisms [10,20] have been introduced sequentially. The axis attention model incorporates an axial structure that highlights long-range dependencies, making it particularly well-suited for lane segmentation tasks. Wang [21] utilized axis attention to conduct position-sensitive attention operations, showcasing robust segmentation performance across various datasets, while also confirming the reliance of images on axial information. This feature extraction method relies on axis attention calculations and is highly suitable for lane recognition tasks. However, due to the lack of complete information from the images, its ability to infer the orientation of lane markings remains insufficient. This paper presents an improved spatial axial attention model based on axial attention, which has a lower computational complexity and incorporates global information.

In the decoder component, traditional lane recognition tasks typically involve the semantic segmentation of lane markings. However, with the introduction of the Pix2Seq model [11], Dr. Chen and others have transformed target recognition tasks into sequence generation tasks. They transformed image coordinate information into an embedded matrix allows for direct text generation based on image features. This model presents greater possibilities for image processing tasks. This framework exhibits high adaptability for lane marking recognition tasks, which are primarily aimed at extracting coordinate points. Consequently, the Lane2Seq model [22] was proposed, establishing lane marking recognition as a sequence generation task for the first time. This model also establishes a unified paradigm for novel lane recognition approaches and demonstrates robust capabilities in the lane recognition task. This paradigm allows for the easy segmentation of multiple lane markings using a simple transformer layer, without introducing additional computational overhead or bias, thereby avoiding the cumbersome modeling processes associated with complex post-processing tasks for lane lines. The decoder framework employed in this paper also utilizes this architecture.

3 Method

As elaborated in Section 1, the recognition of lane markings has seen significant enhancements due to progress in the field of computer vision. Nevertheless, traffic conditions frequently exhibit increased complexity, as one must not only navigate through various obstructions on the roadway but also contend with barriers, traffic signs, walls, and other objects that may mimic lane markings. In the feature extraction phase, we establish a structure for feature extraction that integrates attention layers and convolutional layers in parallel. This framework addresses the long-range dependencies associated with lane markings while clearly distinguishing lane markings from similar interfering objects, consequently minimizing the chances of erroneously classifying other traffic signs as lane markings. Following the outcomes of feature extraction, we produce a sequence of textual representations pertaining to the coordinates of lane markings. The proposed method demonstrates strong capabilities in recognizing lane markings, even within intricate traffic scenarios.

3.1 Overall Network Architecture

This paper introduces a lane marking recognition model referred to as ACP2seq, as depicted in Fig. 1. The model is fundamentally composed of two integral components: the encoder and the decoder. The

encoder component incorporates a parallel spatial axis attention module alongside a multi-scale convolution module, where features are concatenated via self-attention mechanisms. The encoder is primarily used for extracting image features. In the decoder, the model performs a text generation task based on the results of feature extraction and the sequence of lane line coordinates, enabling the recognition and segmentation of multiple lanes.

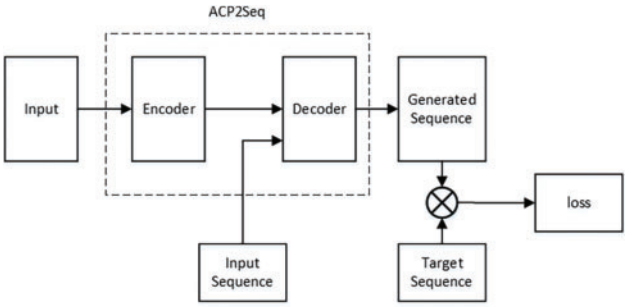

Figure 1: ACP2Seq model framework

3.2 Encoder

Attention mechanisms and convolutional structures are currently among the most frequently employed techniques in the field of image processing. Attention mechanisms are capable of emphasizing the overarching characteristics of images and proficiently managing the long-range dependencies present within them. Conversely, convolutional structures focus on local features, facilitating the enhanced extraction of intricate details within images.

The advent of transformers facilitates the effective integration of information from diverse sources. The decoder architecture enables the integration of information across various modalities. In the decoder architecture, each input element engages in unrestricted interactions with other elements, facilitating the interrelation of data from different sources and modalities to generate the final output. This unique decoder architecture allows the encoder to produce more comprehensive and varied data outputs. Ultimately, this integration enables the attention mechanism and convolutional architecture to collaboratively capture image feature information within the ACP2Seq model.

This study introduces an innovative architecture within the encoder section that combines spatial axis attention with parallel multi-scale convolution. The axis attention module effectively captures global information from the image and establishes long-range dependencies, thereby enhancing the recognition of lane markings in a smooth and coherent manner. The multi-scale convolution highlights the intricate details of the image, facilitating the distinction between lane markings and analogous traffic signs. The architecture of the model is depicted in the Fig. 2.

3.2.1 Spatial Axis Attention

In the conventional ViT architecture, every patch engages in attention calculations with all other patches within the image. This process, however, leads to a considerable computational load. The enhanced Swim architecture utilizes a sliding window method to calculate local attention. Nonetheless, for lane markings, which exhibit elongated linear features, axis attention proves to be more effective in addressing such linear structures.

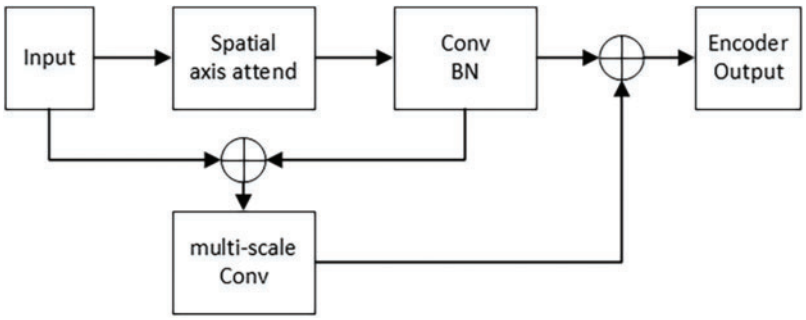

Figure 2: Encoder framework

Axis attention mitigates complexity through the decomposition of input data into various axes, enabling the calculation of attention in a distinct manner for each column. It significantly enhances the feasibility of processing extensive datasets. In axis attention operations, the information acquired by each patch is constrained to the patches located in the corresponding row and column. This approach significantly reduces the computational load of the model.

The axial attention model reduces computational complexity. Consider an input feature map x characterized by dimensions $h \times w \times d_m$, where h denotes the height, w represents the width, and d_m indicates the number of channels. In contrast to the global conventional attention mechanism, characterized by a complexity of $(O(h^2w^2))$, the computational complexity associated with both the spatial axis attention in this article is $(O(hwm))$, where m denotes the length of the axis. The axis attention mechanism significantly diminishes the computational complexity, and enhances the practicality of handling extensive input images. Axis attention, as a variant of local attention operations, is capable of emphasizing long-distance dependencies of lane lines while simultaneously minimizing the computational burden of the model.

The spatial axial attention model developed in this research is depicted in Fig. 3. In contrast to conventional axis attention frameworks, the spatial axis attention architecture introduced in this research sequentially processes the feature map after executing axial slicing. The outcome of the attention computation for each slice is processed through a gating unit and subsequently incorporated into the original data of the following slice, thereby serving as the input for the attention computation of the next slice. The transfer of spatial information facilitates an expanded receptive field for distinct patches. Moreover, regarding the distinct linear structure of lane markings, this sequential transfer of information between axes allows the model to concentrate more on the elongated shape characteristics of the image. This feature demonstrates a significant level of adaptability for lane marking recognition.

This paper additionally presents a gating unit structure designed to regulate the output results of each slice. The gating unit is characterized as a learnable variable that governs the output results via attenuation control. This guarantees that for any specified patch, the impact of the axis slices in proximity to it is more significant. This gating unit structure facilitates the continuous attenuation of spatial information throughout transmission, consequently diminishing the influence of extraneous information. The receptive field of the spatial axial attention structure developed in this study is depicted in Fig. 4.

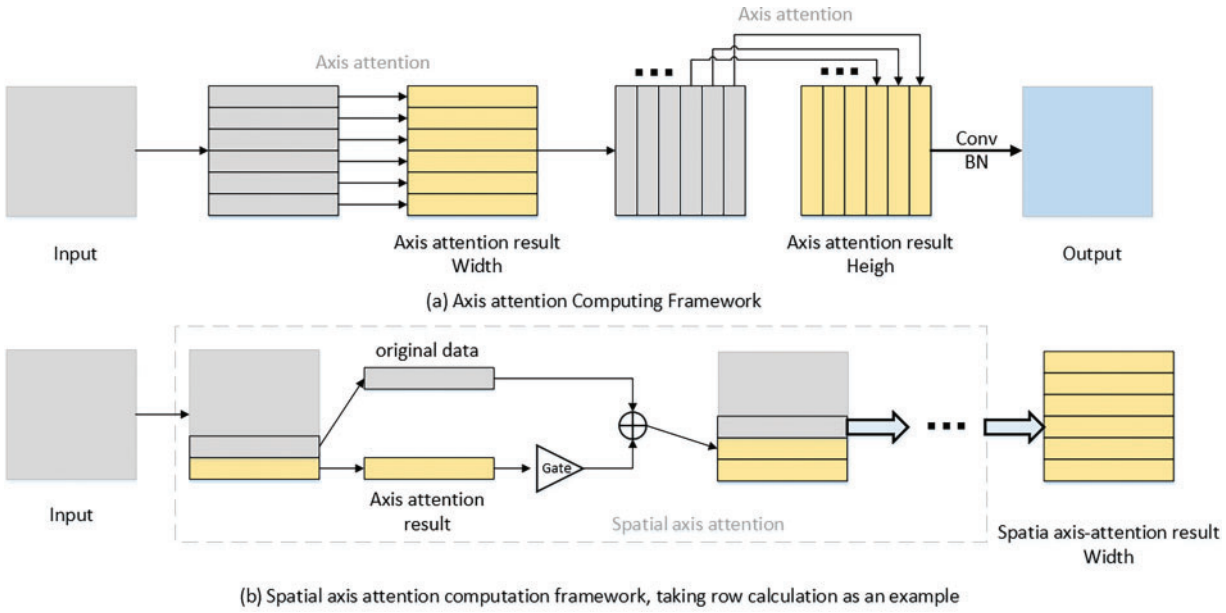

Figure 3: Differences in the structure of axis attention and spatial axis attention models

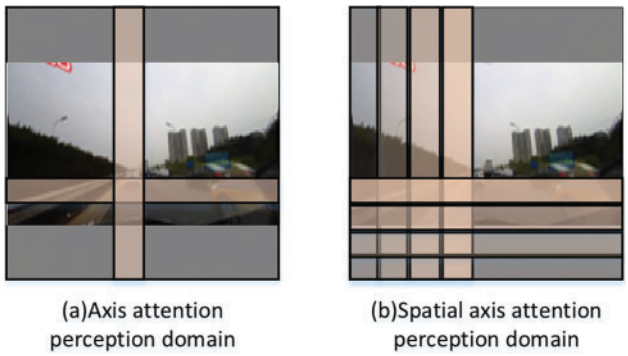

Figure 4: Difference between axis attention and spatial axis attention perception domain

Given an input feature map $x \in R^{h \times w \times d_m}$, with height h, width w, and channels d_m, the output at position $O = (i, j)$, the formula for calculating the output at any position y_o in the attention mechanism is presented in Eq. (1).

$$y_o = \sum_{p \in N} \text{softmax} x_p \left({q_o}^T k_p \right) v_p \tag{1}$$

where N represent the whole location lattice, x_p is the input at position $O = (i, j)$, and queries $qo = W_Q x_o$, keys $k_o = W_K x_o$, values $vo = W_V x_o$ are all linear projections of the input x_o. W_Q, W_K and W_V are all learnable matrices. The softmax function is used to compute the probability distribution of all inputs x_p.

This paper uses position-sensitive attention as a means of enhancement [10]. The position-sensitive attention introduces the incorporation of positional and significantly improves the model's capacity to

comprehend both the relative and absolute positions of elements within the input data. The output at y_o is computed by pooling over the projected input as:

$$y_{o1} = \sum_{p \in N_{m\times 1}(O)} \text{softmax} x_p \left(q_o^T k_p + q_o^T r_{p-o} + k_p^T r_{p-o}\right)\left(v_p + r_{p-o}\right) \tag{2}$$

where N represents the operational area of size $m \times l$, the learnable vector r_{p-o} is the added relative positional encoding. The inner product $q_o^T r_{p-o}$ and $k_p^T r_{p-o}$ denote the dependencies from location p to location $O(i, j)$, which are contingent upon the value and the key. The variable of $v_p + r_{p-o}$ represents the relative positions. Both vectors do not add a significant number of parameters because they are utilized across attention heads within a layer, and the count of local pixels O is usually small.

Upon the introduction of spatial positional information, the computed data for the initial row and column of the image remains constant, whereas the attention mechanisms for the remaining positions incorporate the output from a preceding sequence into the input of the original image, while also integrating gating units to mitigate the output. The spatial attention formula is articulated in Eq. (3).

$$y_o = \sum_{p \in N_{m\times 1}(O)} \text{softmax}(x_p + \alpha y_{p-1})\left(q_o^T k_p + q_o^T r_{p-o} + k_p^T r_{p-o}\right)\left(v_p + r_{p-o}^k\right) \tag{3}$$

where the α represents the gate unit parameters. It is a learnable variable. The y_{p-1} is the axial attention output of the corresponding position of $O(i-1, j)$, which is processed by the gate unit and added to x_p to serve as the input for y_o.

The computation flowchart of position-sensitive attention is shown in Fig. 5. It effectively captures the positional information of each element within the input sequence, thereby enhancing its ability to comprehend the semantic relationships in images, particularly in cases where the sequences are relatively lengthy and the order relationships are more pronounced.

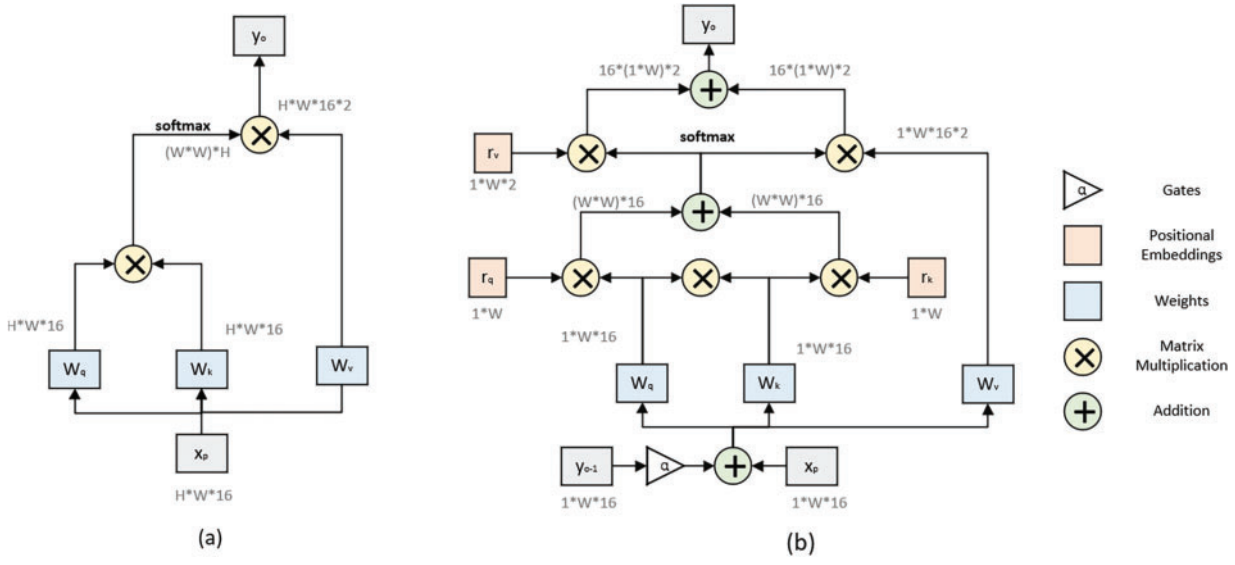

Figure 5: A non-local attention (a) vs. position-sensitive axial-attention applied along the width-axis (b)

3.2.2 Multi-Scale Convolution

Multi-scale convolution generally utilizes a range of convolutional kernels with different dimensions to facilitate feature extraction. The convolutional kernels of varying dimensions are capable of capturing features at multiple scales within the image. This enhances the model's comprehension and processing of input data characterized by diversity and complexity.

This study presents a feedforward component that employs feature layers derived from spatial axial attention, functioning as an additional channel. A multi-scale convolution layer is structured as depicted in Table 1, consisting of four convolutional layers with kernel sizes of 7×7 and 3×3, aimed at facilitating feature extraction. The extracted features are directly input into the encoder. This ensures that the model maintains its ability to effectively handle the complexities of the feature maps.

Table 1: Multi-scale convolution layer

Methods	Kernel size	Padding
conv1	7	3
conv2	3	1
conv3	3	1

3.3 Decoder

The decoder predominantly performs functions associated with text generation. The target sequence of the decoder component consists of distinct coordinate points that denote the lane markings. By considering these coordinate points as textual data, the task of lane recognition can be redefined as a text generation task. The design of the encoder component is depicted in Fig. 6.

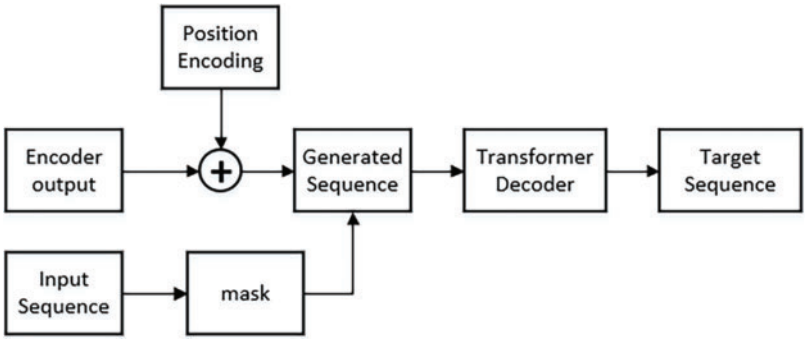

Figure 6: Decoder framework

The decoder component of ACP2Seq performs a text generation task, while the encoder component, as detailed in Section 3.2, extracts image features to form the input sequence. The coordinate points of lane markings, along with the position markers, form the target sequence.

The target sequence is composed of a series of sequential coordinate points, which are represented in the format [start, x0, y0, x1, y1,..., xn, yn, end, start, x0, y0,...]. In this context, "start" refers to the initial point of a lane, whereas "end" indicates the concluding point of that particular lane. The application of this flag information facilitates the integration of various lane markings into a unified collection of target sequence data. The dimensions of the embedded matrix corresponding to the target sequence are consistent with the

dimensions of the image, and there exists a correlation between the sequence length, the number of lanes, and the length of the lane point set. In order to address the issue of sequence length, this research utilizes uniform sampling on the coordinate points and implements polynomial fitting to effectively replicate the lane markings.

This study utilizes the cross-entropy function as the objective function to evaluate the similarity between the target sequence and the generated sequence. The formulation of the cross-entropy function is delineated in Eq. (4); its objective is to optimize the likelihood of the produced tokens in conjunction with the subsequent target tokens.

$$Loss_{obj} = -\sum_{j=1}^{N} \omega_j \log P(\hat{m}_j | I, m_{1:j-1}) \tag{4}$$

where m and $\hat{m}$ respectively denote the input and target sequences. N represents the length of the sequences. ω_j represents the weight for the j-th token, a value of 0 is assigned to the format transcription token, while a value of 1 is assigned to other tokens to ensure the model is trained to predict the desired tokens but not the format transcription tokens.

4 Experiments

4.1 Datasets

We conduct experiments on two classic lane markings detection datasets: CULane [14] and Tusimple [23].

CULane is a widely utilized large-scale dataset for lane detection. It encompasses numerous challenging scenarios, including crowded roads, night scenes, and adverse weather conditions. The CULane dataset comprises 88.9 K images for training, 9.7 K images in the validation set, and 34.7 K images for testing. All images have a resolution of 1640 × 590 pixels.

Tusimple is a real-world highway dataset. It contains authentic driving scene images captured from U.S. highways, along with corresponding lane line annotation information. The Tusimple dataset comprises 3626 images for training, 358 images in the validation set and 2782 images for testing. All images have a resolution of 1280 × 720 pixels.

4.2 Evaluation Metrics

For CULane dataset, we adopt the F1 score to measure the performance. F1 score is the harmonic mean of precision and recall.: $F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$, where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$. Where TP (True Positive) denotes the count of positive samples that the model accurately classifies as positive. FP (False Positive) denotes the number of negative samples that the model erroneously classifies as positive. FN (False Negative) denotes the number of positive samples that the model mistakenly classifies as negative.

For Tusimple dataset, we use the F1 score, accuracy, FP, and FN to evaluate the model performance. Accuracy is defined as $Accuracy = \frac{\sum_{clip} C_{clip}}{\sum_{clip} S_{clip}}$, where C_{clip} represents the number of accurately predicted lane polynomial parameters and S_{clip} denotes the total number of lane polynomial parameters. A lane polynomial parameter is considered as correct if its absolute error is smaller than the given threshold $t_{pc} = \frac{20}{\cos a_{yl}}$, where $\cos a_{yl}$ denotes the angle of the corresponding ground truth [22].

4.3 Implementation Details

Model: The patch size is established as 16 × 16 for the attention operations. The dimensions of the word embeddings for the vocabulary and nbins is 512 and the maximum input length is 300.

Training: All images are resized to 512 × 512 with the addition of gray bars to prevent distortion, and data augmentation is performed through random horizontal flips and affine transformations. The model utilizes the AdamW activation function [24], with a weight decay parameter set to $1e{-}4$ and a dropout rate of 0.1. Additionally, the learning rate follows a linear decay schedule [25], starting with an initial value of $1e{-}4$ with an initial value set to $1e{-}4$. The model is configured to utilize a batch size of 4 and is trained for a total of 50 epochs.

Experimental environment: Windows 10 operating system, Intel Core i7-8700 CPU at 3.19 GHz, 32 GB RAM, and NVIDIA GTX 2070 Ti.

4.4 Comparison with the State-of-the-Art Methods

Evaluation of performance on CULane. This study presents a comparison of the ACP2Seq model against other leading methodologies utilizing the CULane dataset, with the findings detailed in Table 2. The ACP2Seq model demonstrated a commendable F1 score across the four scenarios of the CULane dataset: Crowded, Dazzle, Arrow, and Night. When juxtaposed with the Lane2Seq model that employs ViT as the encoder, the proposed model utilizing ACP-Attend as the encoder enhanced the CULane score from 78.39% to 79.24%, thereby showcasing remarkable performance in lane markings recognition within the context of single-scale detection.

Table 2: Comparison of the F1 score performance of different models on the CULane testing set. For the crossroad scenario, only FP is presented. The optimal data is indicated in bold

Methods	Normal	Crowded	Dazzle	Shadow	No line	Arrow	Curve	Cross	Night	Total
R34-UFLD [26]	90.70	70.20	59.50	69.30	44.40	85.70	69.50	2037	66.70	72.30
R34-ADNet [27]	92.90	77.45	71.71	79.11	52.89	89.90	70.64	1499	74.78	78.94
R34-SCNN [14]	90.60	69.70	58.50	66.90	43.40	84.10	64.40	1990	66.10	71.60
R34-LaneATT [28]	92.14	75.03	66.47	78.15	49.39	88.38	67.72	1330	70.72	76.68
R122-LaneATT [28]	91.74	76.16	69.47	76.31	50.46	86.29	64.05	1264	70.81	77.02
CurveLanes-NAS-S [29]	88.30	68.60	63.20	68.00	47.90	82.50	66.00	2817	66.20	71.40
R50-Laneformer [30]	91.77	75.74	70.17	75.75	48.73	87.65	66.33	**19**	71.04	77.06
LaneAF [31]	91.80	75.61	71.78	79.12	51.38	86.88	72.70	1360	73.03	77.41
GANet-S [32]	93.24	77.16	71.24	77.88	53.59	89.62	**75.92**	1240	72.75	78.79
ViT-Lane2Seq [22]	**93.03**	76.42	72.17	78.32	**52.89**	89.67	72.67	1319	73.98	78.39
ACP2Seq (ours)	92.42	**77.97**	**75.03**	**78.35**	51.44	**91.77**	74.93	1469	**74.22**	**79.24**

Due to the reduction in computational complexity achieved by the spatial axis attention mechanism in the encoder and the simple structure of the decoder, which consists of only one layer of transformer, the ACP2Seq model proposed in this study can be trained at a rate of 23 FPS (Frames Per Second). This operational speed significantly exceeds that of models such as LaneNet and SCNN but remains lower than the image processing rate of over 50 FPS characteristic of the Enet model.

The results of lane detection are presented in Fig. 7. The ACP2Seq model exhibited strong capabilities in detecting lane markings across the CULane datasets, while sustaining effective inference performance in challenging environments, such as night time and congested conditions. Moreover, when evaluated against alternative models, the ACP2Seq model demonstrates a superior capability to differentiate between lane

markings and linear traffic objects, including guardrails and walls. Both LaneATT and Lane2Seq (ViT) have instances of misidentifying other markers as lane lines; however, the model proposed in this paper effectively addresses this issue.

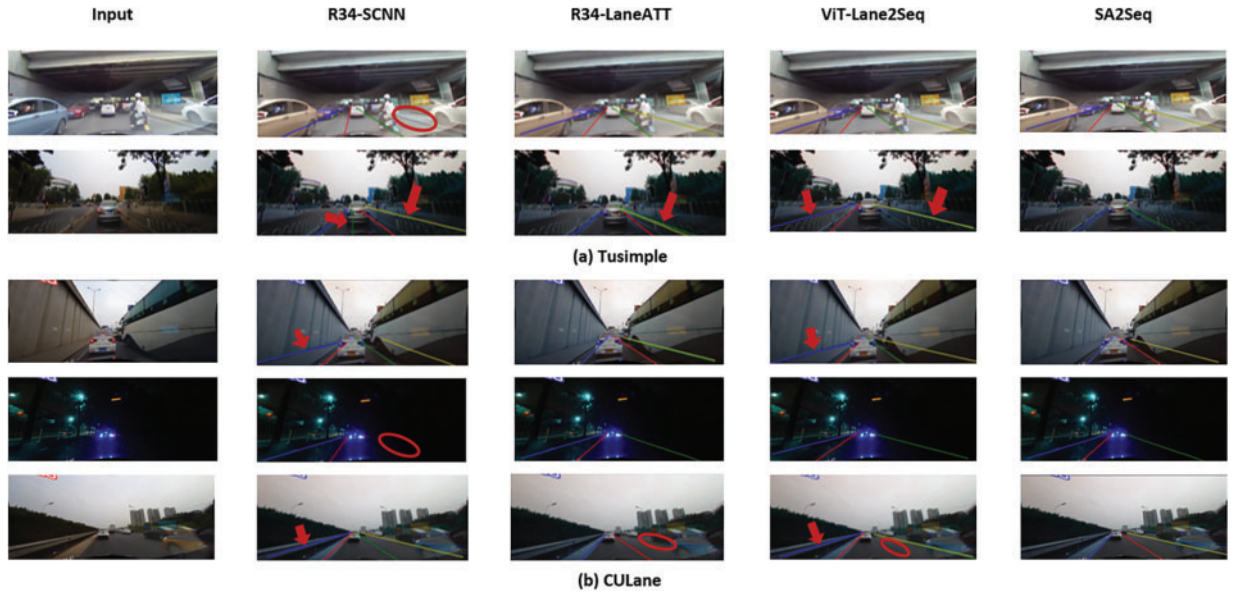

Figure 7: Visualization results of SCNN, LaneATT, Lane2Seq, ViT-Lane2Seq and ACP2Seq on Tusimple and CULane, where the red arrows indicate false positives and the red circles represent false negatives

Evaluation of performance on Tusimple. This study conducts a comparative analysis of the ACP2Seq model against alternative methodologies utilizing the TuSimple dataset, with the findings detailed in Table 3 and Fig. 7. On the Tusimple dataset, ACP2Seq also demonstrates the ability to effectively distinguish between lanes and guardrails, thereby facilitating precise lane marking recognition. The ACP2Seq model demonstrated an impressive accuracy of 97.25% alongside a minimal false positive rate of 1.86%. Nevertheless, owing to the limited size and relatively consistent characteristics of the TuSimple dataset, the enhancement of the model is not significant. It still validates the strong lane line recognition capability of the model proposed in this study.

Table 3: Comparison of the performance of different models on the Tusimple dataset. The optimal data is indicated in bold. Acc denotes accuracy

Methods	F1 (%)	Acc (%)	FP (%)	FN (%)
R34-UFLD [26]	88.02	95.86	18.91	3.75
R34-SCNN [14]	95.97	96.53	6.17	**1.80**
R34-LaneATT [28]	96.77	95.63	3.53	2.92
R122-LaneATT [28]	96.06	96.10	5.64	2.17
LaneAF [31]	96.71	–	3.24	2.82
GANet-S [32]	97.71	95.95	1.97	2.62
ViT-Lane2Seq [22]	**97.95**	96.85	2.01	2.03
ACP2Seq (ours)	97.82	**97.25**	**1.86**	2.09

4.5 Ablation Study

Ablation experiments are conducted on the CULane dataset to assess the effectiveness of each component. This research focuses on the standard subset of the CULane dataset, analyzing the recognition outcomes of models that incorporate convolution as a parallel architecture against those that do not, in order to enhance feature extraction within the encoder framework. Furthermore, it evaluates the recognition accuracy and F1 scores of various convolutional architectures, such as standard convolution, dilated convolution, and multi-scale convolution, as sources of parallel information. The findings are delineated in Table 4. The incorporation of convolutional structures clearly improves the model's capacity to extract features from images. Among the various convolutional architectures, multi-scale convolution demonstrates superior recognition performance. This suggests that multi-scale convolution effectively captures more intricate details.

Table 4: Results on the CULane dataset by changing convolution layer. The optimal data is indicated in bold. CNN: convolutional neural network, DC: dilated convolution, MSC: multi-scale convolution

ACP2Attend	CNN	DC	MSC	F1 (%)	ACC (%)
1	–	–	–	87.25	88.46
2	✓	–	–	91.03	91.55
3	–	✓	–	88.54	89.72
4	–	–	✓	**92.42**	**92.78**

This study examines the effects of employing position-sensitive attention and the information transmission architecture on the model, utilizing the standard subset of the CULane dataset. The findings are illustrated in Table 5. The transfer of inter-axis information markedly improves the model's recognition performance; nonetheless, a pure axis attention model continues to exhibit effective performance in lane line extraction. This indicates that a significant quantity of information is contained within the inter-axis data of the horizontal and vertical axes. The transfer of inter-axis information significantly enhances the accuracy of the model's recognition capabilities. Additionally, Table 5 presents the results of fusion ablation experiments incorporating multi-scale convolution (MSC), demonstrating that the combined effects of multi-scale convolution and axis attention facilitate the precise recognition of lane markings.

Table 5: Results on the CULane dataset by changing the attention calculation. The optimal data is indicated in bold. PS: position-sensitive self-attention, ITS: information transmission

ACP2Attend	PS	ITS	F1 (%)	ACC (%)	F1 (%) (MSC)	ACC (%) (MSC)
1	–	–	85.41	87.52	88.57	89.74
2	–	✓	87.62	87.95	89.93	90.03
3	✓	–	86.54	87.22	89.62	90.54
4	✓	✓	87.25	88.46	**92.42**	**93.78**

5 Conclusion

This study proposes a lane markings extraction model, ACP2Seq. In the encoder component, this study presents a parallel feature extraction module that integrates spatial axis attention alongside multi-scale convolution techniques. The spatial axis attention structure is designed to enhance the model's perceptual capabilities through the combination of global attention and inter-axis information transfer. This framework

systematically divides and analyzes images along both the horizontal and vertical dimensions, computing position-sensitive attention for each segment and transferring the information through the gating mechanism. Multi-scale convolution is used to supplement detailed information at different scales to enhance the model's recognition capabilities. In the decoder component, this paper considers lane marking coordinate points as target sequences in the framework of a sequence generation task, employing markers to distinguish between different lanes. In the process of achieving lane line detection, it concurrently accomplishes the segmentation of multiple lane markings.

This module illustrates robust segmentation inference capabilities for lane markings. The model demonstrates extremely strong recognition capabilities and high operating speed on both the Tusimple and CULane datasets. Compared to other lane detection models, ACP2Seq is more adept at capturing the morphological features of lane markings. ACP2Seq is better at avoiding confusion between lane markings and interfering objects such as roadside barriers, wall edges, and directional arrows, thereby reducing the occurrence of misidentifications. Due to the large receptive field provided by the attention mechanism, the model can effectively cope with disturbances in dark or complex environments, thereby avoiding both missed detections and false positives. Additionally, due to the use of a computationally efficient axial attention structure in the encoder and a simple single-layer transformer structure in the decoder, this model demonstrates a significantly higher frame rate compared to models such as LaneATT and SCNN. The model operates at a faster speed, ensuring better real-time performance. This robust discriminative capability significantly enhances the model's reliability, enabling it to distinguish and recognize lane lines with greater precision in complex environments.

However, in scenarios where the number of lane markings exceeds five or where the curvature of the lane markings is excessively high, the length of the generated text sequences can become significantly longer. This results in a slowdown of the model's processing speed, indicating a need for further optimization and modification. Additionally, for the model to be effectively applied within the Internet of Vehicles, it is essential to implement embedded migration and adaptation of the model to facilitate real-world applications.

Acknowledgement: Not applicable.

Funding Statement: Duo Cui reports financial support was provided by Ministry of Education Higher Education Industry University Research Innovation Fund Special Project. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions: Duo Cui: Conceptualization, methodology, software, etc. Qiusheng Wang: Formal analysis, resources, investigation, etc. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data openly available in a public repository. The data that support the findings of this study are openly available in Tusimple dataset at http://doi.org/10.1109/ICCV.2019.00110 and Culane dataset at http://doi.org/10.1109/ICCV.2019.00112.

Ethics Approval: Not applicable.

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

Abbreviations

CNN	Convolutional Neural Networks
ViT	Vision Transformers
ACP2Seq	Attention Multi-Scale Convolution Parallel to Sequence

TP	True Positive
FP	False Positive
FN	False Negative
MSC	Multi-Scale Convolution
DC	Dilated Convolution
PS	Position-Sensitive Self-Attention
ITS	Information Transmission

References

1. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 801–18.
2. Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: fully convolutional DenseNets for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops; 2017. p. 11–9.
3. Chen L, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587. 2017.
4. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell. 2018;40(4):834–48. doi:10.1109/TPAMI.2017.2699184.
5. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. 2017.
6. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
7. Neven D, De Brabandere B, Georgoulis S, Proesmans M, Van Gool L. Towards end-to-end lane detection: an instance segmentation approach. In: 2018 IEEE Intelligent Vehicles Symposium (IV); 2018; IEEE. p. 286–91.
8. Kim J, Lee M. Robust lane detection based on convolutional neural network and random sample consensus. In: Neural information processing. Cham: Springer International Publishing; 2014. p. 454–61, 489.
9. Gopalan R, Hong T, Shneier M, Chellappa R. A learning approach towards detection and tracking of lane markings. IEEE Trans Intell Transport Syst. 2012;13(3):1088–98. doi:10.1109/TITS.2012.2184756.
10. Ho J, Kalchbrenner N, Weissenborn D, Salimans T. Axial attention in multidimensional transformers. arXiv:1912.12180. 2019.
11. Chen T, Saxena S, Li L, Fleet DJ, Hinton GE. Pix2seq: a language modeling framework for object detection. arXiv:2109.10852. 2021.
12. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. arXiv:1505.04597. 2015.
13. Paszke A, Chaurasia A, Kim S, Culurciello E. ENet: a deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147. 2016.
14. Pan X, Shi J, Luo P, Wang X, Tang X. Spatial as deep: spatial CNN for traffic scene understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2018. Vol. 32.
15. Zheng T, Fang H, Zhang Y, Tang W, Yang Z, Liu H, et al. RESA: recurrent feature-shift aggregator for lane detection; 2021. arXiv:2008.13719. 2021.
16. Liu L, Chen X, Zhu S, Tan P. CondLaneNet: a top-to-down lane detection framework based on conditional convolution. arXiv:2105.05003. 2021.
17. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30:5998–6008.
18. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. arXiv:2103.14030. 2021.
19. Hassani A, Walton S, Li J, Li S, Shi H. Neighborhood attention transformer. arXiv:2204.07143. 2023.

20. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation. arXiv:2102.10662. 2021.
21. Wang H, Zhu Y, Green B, Adam H, Yuille AL, Chen L. Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation; 2020. arXiv:2003.07853. 2020.
22. Zhou K. Lane2Seq: towards unified lane detection via sequence generation. arXiv:2402.17172. 2024.
23. TuSimple. TuSimple benchmark. [cited 2025 Mar 30]. Available from: https://github.com/TuSimple/tusimple-benchmark.
24. Loshchilov I, Hutter F. Fixing weight decay regularization in adam. arXiv:1711.05101. 2017.
25. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv:1512.03385. 2015.
26. Qin Z, Wang H, Li X. Ultra fast structure-aware deep lane detection. 2020. arXiv:2004.11757.
27. Xiao L, Li X, Yang S, Yang W. ADNet: lane shape prediction via anchor decomposition. arXiv:2308.10481. 2023.
28. Tabelini L, Berriel R, Paixão TM, Badue C, Souza AFD, Oliveira-Santos T. Keep your eyes on the lane: real-time attention-guided lane detection. arXiv:2010.12035. 2020.
29. Xu H, Wang S, Cai X, Zhang W, Liang X, Li Z. CurveLane-NAS: unifying lane-sensitive architecture search and adaptive point blending. arXiv:2007.12147. 2020.
30. Han J, Deng X, Cai X, Yang Z, Xu H, Xu C, et al. Laneformer: object-aware row-column transformers for lane detection. arXiv:2203.09830. 2022.
31. Yang J, Zhang L, Lu H. Lane detection with versatile atrousformer and local semantic guidance. Pattern Recognit. 2023;133(3):109053. doi:10.1016/j.patcog.2022.109053.
32. Wang J, Ma Y, Huang S, Hui T, Wang F, Qian C, et al. A keypoint-based global association network for lane detection. arXiv:2204.07335. 2022.