PathFormer: SegFormer Based Lane Detection for Curved Lanes

Joseph Fedoronko
Ann Arbor, United States
https://github.com/fedoronj
fedjz88@gmail.com

***Abstract*—With the growing adoption of delivery robots, robotaxis, and personal autonomous vehicles, lane detection models must be tailored to the unique needs of each use case. These vehicles offer many advantages specifically in the form of safety and ease of operation. One such example of this is that elderly populations can struggle to drive and properly see the road, especially at night. In addition, truck drivers suffer from eye strain and fatigue during long drives, many times resulting in crashes or hazardous driving. Lane detection models are able to assist or take over the task of driving, creating new opportunities for people and businesses that would never have been possible without this technology. While the task of identifying lane markers on a road is not new, lane detection remains at the forefront of computer vision problems due to its specific requirements of speed, accuracy, and hardware limitations. When designing lane detection models the most important factor remains accuracy achieved by the model as when deployed, the model must not have malfunctions to ensure user safety. The proposed model aims to improve on existing lane detection models by offering overall improved accuracy.**

I. Related Works

With advances in modern integrated hardware, deep learning algorithms have emerged as the dominant approach to both lane detection and path planning problems. Prior to this shift, traditional computer vision techniques were mainly used for image segmentation and feature extraction, relying on handcrafted filters, edge detection, and color thresholding to identify lane markings.

A. Computer Vision

OpenCV, a widely used computer vision library, provides a range of functions that enable lane detection without the need for model training. Such classical approaches are demonstrated in prior work [2], where an input image is processed using edge detection and Hough transform operations to generate a mask that segments lane markings from the road surfaces. With these solutions, computation time is low O(N) with most of the time complexity determined by the Hough transform, which maps each edge pixel into a parameter space (ρ, θ) where ρ represents the perpendicular distance of a line from the image origin and θ denotes its orientation. As edge pixels vote, each pixel supports multiple potential line locations; for possible line positions, consistent straight patterns receive the most support, enabling the algorithm to detect lane markers in the image. This time complexity allows these methods to be seamlessly integrated into existing environments that use low-power hardware such as FPGAs. While the computation cost of these algorithms is low, the work required to achieve a perfectly fine-tuned solution can prove tedious and time-consuming. Functions such as Canny edge detection/Hough transforms require specific brightness and contrast values in order to correctly detect features. This work requires programmers to hand-tune filters and parameters in order to have the model work for a variety of environments. Even then, this solution can struggle with adverse conditions such as snow, rain, fog, and even darkness due to the variability of the library functions.

B. CNNs

With the introduction of CNN, a major deep learning breakthrough, this new work was also applied to the task of lane detection. With these mathematical functions called convolutions, a model was then able to down-sample an original image to only focus in on the most important elements, the lane pixels in this case. Initially, CNN systems excelled at local context but struggled to integrate global context. The difficulty was due to these models making use of a small, fixed window, or kernel, that did not overlap and could only generate relationships between nearby pixels of the image. Without understanding the context of the whole image, CNNs struggled to recognize image-wide patterns, leading to decreased continuity in resulting segmentation masks. Spatial CNNs such as SCNN [3] aim to address this problem via a technique called message passing. Using this concept, an image can be processed sequentially in which each pixel, or pixel block, passes its associated weight that is then added to the new pixel state to determine patterns and predict what the next pixel should be. This message passing is, however, only executed along spatial directions (top-to-bottom, left-to-right, etc.). Consequently, correlations between spatially distant but semantically related pixels (e.g., along curved or merging lanes) may be missed.

C. Transformers

With the emergence of transformer architectures in modern deep learning, many systems now consider transformers as an alternative or complement to traditional CNN-based designs. Transformers have reshaped how machine learning models interpret data, particularly in natural language processing, where they first demonstrated the power of global context modeling.

Google's Attention Is All You Need [4] showed that relating all elements of an input sequence to one another enables the model to capture long-range dependencies that earlier architectures struggled to represent. When applied to vision tasks, this same mechanism allows each pixel or patch to attend to every other, revealing global spatial patterns that improve predictive accuracy and structural understanding. Although self-attention O(N^2) multi-head attention enables these relationships to be learned across multiple representational subspaces in parallel, improving modeling capacity and training efficiency without reducing the underlying asymptotic cost. This combination of global context and parallel processing forms the foundation for modern transformer-based vision models.

These transformers were quickly adapted into vision tasks such as lane detection. In the LaneFormer paper proposed by [1] a transformer takes the place of the previously used CNNs in the feature extraction backbone. The merit of this approach lies in its ability to perform pattern prediction based on a global contextual understanding rather than purely local features. This enables the model to infer lane structures even under challenging conditions, such as when lane markers are partially occluded or poorly illuminated by shadows. LaneFormer implements a row-column-based attention mechanism to generate this global context. This means that attention calculations are done using 1D vectors, in either the x or y direction. The effect of this method is that computation time is greatly reduced, especially when integrating the multi-head attention mechanism that was discussed earlier. Following this calculation, similar 1D vectors that are seen are able to apply these patterns to correctly identify lane markers, even in adverse conditions. However, an area of concern is the curved lane markers. When the lane markers do not align perfectly with these created 1D vectors, LaneFormer can struggle to identify the entire marker. Despite this limitation, LaneFormer represents a significant step toward leveraging global attention mechanisms for structured visual understanding. Its success demonstrates the potential of transformer-based architectures to outperform traditional CNN backbones in complex perception tasks such as lane detection, while also highlighting opportunities for further refinement in handling non-linear or irregular lane geometries.

II. Methodology

Drawing on insights from recent transformer-based perception systems, most notably SegFormer for hierarchical feature encoding and SCNN-based approaches for structural continuity. The proposed PathFormer model integrates global contextual reasoning with efficient lane-specific decoding. SegFormer's MiX Transformer (MiT) backbone is used to extract multi-scale features with high-resolution spatial detail, while an SCNN-based lane detection head enforces continuity along the lane markings through directional message passing. The following subsections describe the model architecture, encoder design, lane detection head, and overall model flow.

A. Model Architecture

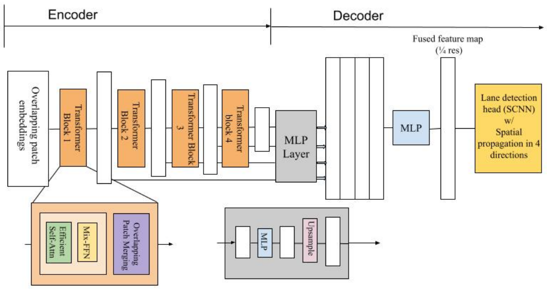

Fig. 1. Arch for PathFormer

PathFormer differs from lane detection architectures such as LaneFormer, which rely on row–column attention restricted to horizontal and vertical directions. Instead, PathFormer employs the SegFormer MiT-based encoder, which preserves the full self-attention over spatial tokens. This helps mitigate the directional bias inherent in row–column attention and improves the modeling of curved, nonlinear, or irregular lane structures. Multi-scale features from the encoder are fused in the decoder and subsequently processed by a spatially aware SCNN head that enhances lane continuity and structural consistency.

B. Encoder

PathFormer uses the SegFormer encoder to produce hierarchical feature maps at multiple scales. The MiX Transformer improves on standard Vision Transformers in two main ways, giving it stronger local context and broader global awareness while remaining more efficient and better at preserving spatial structure than traditional transformer backbones.

1) Overlapping Patch Embeddings: Instead of the non-overlapping P×P patches used in standard ViTs (where kernel size = stride = P), MiT applies convolutional embeddings with k>s, causing adjacent patches to overlap by (k - s) pixels. This preserves local spatial relationships and captures fine-grained texture information that non-overlapping patches typically lose.

2) Hierarchical Multi-Stage Structure: The encoder processes the image through four stages (with resolutions H/4, H/8, H/16, and H/32), improving representational efficiency without explicitly reducing computational complexity. This hierarchical design enables the model to encode both detailed and global semantic information, which is essential for detecting long continuous structures such as lane lines.

C. Lane Detection Head

The SegFormer model in its native state excels at object detection and semantic segmentation, where each pixel of an image is classified as a particular class with high levels of accuracy. The challenge faced when converting SegFormer into a lane detection model was the concern relating to the overall computation cost. Many modern lane detection models operate by classifying only the pixels related to lane markers. Some lane detection networks, such as Spatial As Deep: Spatial CNN for Traffic Scene Understanding (SCNN), operate as binary classifiers, distinguishing between lane and non-lane regions within down-sampled feature maps rather than at the pixel level of the original image. Adapting SegFormer to perform binary classification could improve computational efficiency; however, it would still struggle to match the performance of dedicated lane detection models unless the effective number of classified pixels were significantly reduced through spatial downsampling or attention-based focus mechanisms. Other methods still make use of a mask via supervised learning to overlay each lane marker [Cite-new].

In order to keep the segmentation power of SegFormer while also reducing the number of pixels classified, PathFormer implements a lane detection head as part of the decoder stage. Beginning with the decoder, multi-scale feature maps extracted from each MiT transformer stage are first passed through lightweight MLP layers to align their channel dimensions. These features are then up-sampled and fused to form a unified representation at one-quarter of the original image resolution, preserving both fine-grained spatial information and global semantic context. This fused feature map serves as the input to the proposed lane detection head, which employs a Spatial CNN (SCNN) mechanism to propagate information across rows and columns of the feature map. Through this directional message passing, the head reinforces lane continuity and connectivity, enabling robust detection of lane structures even in cases of partial occlusion or degradation. The SCNN head subsequently outputs a two-dimensional lane probability mask, which can be up-sampled to the original image size for visualization and evaluation.

This integration of transformer-based feature extraction with a spatially aware decoding mechanism provides several advantages for lane detection. The hierarchical structure of the SegFormer encoder captures the global road context and long-range dependencies, enabling the network to infer the presence and orientation of lanes even under challenging visual conditions such as occlusion, worn paint, construction zones, or varying illumination. Meanwhile, the SCNN lane detection head complements this global reasoning by enforcing local spatial coherence along the elongated and continuous structures characteristic of lane markings. Through directional message passing, the SCNN effectively connects fragmented lane segments and suppresses spurious activations from road artifacts or shadows. Together, these components allow PathFormer to balance the contextual understanding of the transformer with the structural consistency of the SCNN, resulting in a model that maintains segmentation accuracy while efficiently focusing the computation on regions relevant to the lane.

D. Combined Advantages

Integrating a transformer-based encoder with a spatially structured SCNN decoder provides several clear benefits. The MiT encoder supplies strong global context, helping the model understand overall road geometry, lane layout, and environmental structure. At the same time, the SCNN decoder reinforces local continuity, allowing the network to maintain coherent lane boundaries and reduce fragmentation that would otherwise occur when lanes curve, overlap, or fade. The architecture hierarchically fuses features, allowing both fine-grained spatial edges and high-level semantic information to significantly contribute to the final lane predictions. This also reduces the burden of pixel-level classification, enabling the system to remain efficient while preserving SegFormer's strong segmentation capabilities. Altogether, these components work together to balance large-scale contextual reasoning with detailed structural consistency, resulting in accurate and stable lane detections even in challenging real-world driving scenarios. This innovative approach not only enhances the precision of lane detection but also adapts seamlessly to varying road conditions and environments. As a result, it significantly improves the overall safety and reliability of autonomous driving systems.

III. Results

This section evaluates the proposed PathFormer architecture through qualitative comparison with existing lane detection models, particularly LaneFormer, and outlines the expected performance gains to be validated through future training and experimentation. LaneFormer, a similar model to the proposed SegFormer, performs with a 77.1 F1 score on the CULane dataset, meaning that the model overall shows a good balance between precision and recall on a variety of road images. PathFormer aims to produce an improved F1 score of 79-81 on the same dataset. These improvements in F1 score are expected to result from improved modeling of curved and non-linear lane structures.

The next steps to realize these improvements would be to first train the new PathFormer model on a dataset such as CULane or TUSimple. Both datasets contain images of lane markers in perfect conditions, as well as at night and in adverse conditions, giving an accurate depiction of the comprehensive driving environment.

This work aims to improve existing lane detection algorithms by leveraging features of powerful segmentation models. By changing the way that input data is encoded and maintained, PathFormer creates the opportunity for more accurate lane identification, specifically on curvy roads. Creating a more comprehensive attention calculation method allows for the creation of a more accurate mask, overlaying the original image. This work will benefit companies looking to improve their lane detection algorithms, specifically if there are areas that current models fail to predict.

References

[1] Jianhua Han, Xiajun Deng, Xinyue Cai, Zhen Yang, Hang Xu, Chunjing Xu, and Xiaodan Liang. Laneformer: Object-aware row-column transformers for lane detection, 2022.

[2] Astika Istiningrum, Umi Salamah, and Nurcahya Pradana Taufik Prakisya. Lane detection with conditions of rain and night illumination using hough transform. In *2022 5th International Conference on Information and Communications Technology (ICOIACT)*, pages 429–434, 2022.

[3] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding, 2017.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.