2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

PathFormer: A Transformer-Based Framework for Vision-Centric Autonomous Navigation in Off-Road Environments

Bilal Hassan[1,2], Nadya Abdel Madjid[1,2], Fatima Kashwani[1,2], Mohamad Alansari[2], Majid Khonji[1,2], and Jorge Dias[1,2]

***Abstract*— The efficient navigation of autonomous vehicles across rugged and unstructured terrains remains a significant challenge. Most existing research in this area emphasizes the need for complex mappings or intricate multi-step methodologies. However, these traditional approaches often struggle to adapt to dynamic changes in environmental conditions. In this paper, we introduce PathFormer, an end-to-end framework designed specifically to address these challenges. PathFormer utilizes transformers to decode free-space semantics and configurations directly from camera images, enabling efficient path planning without the reliance on detailed, pre-existing maps. The performance of PathFormer was rigorously evaluated across diverse datasets, where it demonstrated superior capabilities, outperforming other state-of-the-art methods by 3.68% in precisely segmenting free-space regions and showing a 13.65% improvement in correctly predicting traversable paths.**

I. Introduction

Navigating autonomous vehicles through unstructured and remote terrains is a key challenge in the growing research area of autonomous driving [1], [2], [3]. Off-road environments, lacking structured paths, cover a wide range of natural settings such as forests, deserts, and mountain areas. These terrains are unpredictable and constantly changing, presenting challenges like uneven surfaces, vegetation, and the absence of clear road markings. Traditional navigation methods, which rely on pre-existing maps and complex processing stages, are not well-suited to the variable and uncertain nature of off-road environments [4]. Although effective in structured urban areas, they face difficulties with the rapidly changing conditions of off-road environments, especially when accurate maps are not available or are outdated [5], [6], [7].

In the wider context of autonomous navigation, a variety of strategies have been investigated, from rule-based algorithms to complex, multi-step systems and the latest techniques enabled by deep learning [7], [8], [9]. The inflexibility of rule-based algorithms limits their adaptability to new terrains, while multi-step systems, involving separate mapping, localization, and planning stages, may lack the necessary quick response, particularly in unstructured settings [10], [11], [12].

*This publication is based upon work supported by the Khalifa University under Award No. RC1-2018-KUCARS-8474000136.

[1]These authors are affiliated with the Khalifa University Center for Autonomous Robotic Systems (KUCARS), Abu Dhabi, 127788, UAE.

[2]These authors are affiliated with the Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, 127788, UAE. {bilal.hassan, 100049370, 100045434, 100061914, majid.khonji, jorge.dias}@ku.ac.ae

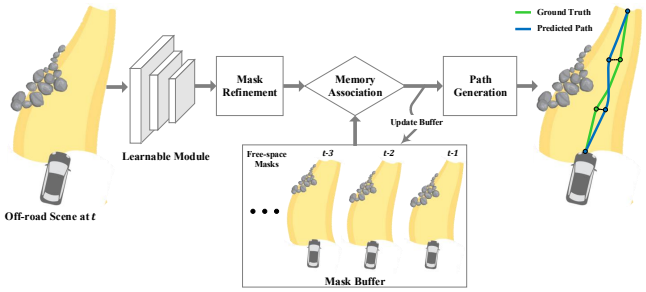

Fig. 1. Overview of the PathFormer framework, demonstrating the progression from off-road scene input to path prediction.

Deep learning has catalyzed significant advancements across numerous fields, resulting in the development of innovative and efficient methodologies [13], [14], [15], [16]. Within the realm of autonomous driving, it has emerged as a pivotal technology for fostering adaptive and efficient navigation [17]. Among the various deep learning approaches, convolutional neural networks (CNNs) have shown exceptional efficacy in on-road navigation by leveraging extensive datasets for improvement. However, their performance tends to diminish in off-road conditions, where detailed maps are frequently unavailable [4], [7], [18].

While urban areas benefit from high-definition (HD) maps [19], OpenStreetMap (OSM) data [20], and sensors like lidar, ultrasonic, and radar for environmental perception [21], [22], [12], off-road settings present unique challenges. The unpredictable nature of these terrains, which can change significantly after events like rainfall, makes traditional urban-focused tools less effective [23], [24]. This emphasizes the need for innovative approaches that focus on vision-based navigation, to which our solution, "PathFormer," is directed. Our research evaluates this method across a broad spectrum of off-road datasets, including natural parks, rural areas, deserts, and simulated environments, encompassing the diverse range of off-road scenarios. The major contributions of this research are:

1) Introducing an end-to-end framework for off-road path planning that relies on vision-based sensing, eliminating the need for prior HD maps. The framework uses CNNs for robust feature extraction, enhanced by transformer encoders for increased effectiveness.
2) The framework allows joint detection and segmentation of paths across diverse off-road landscapes, improving autonomous navigation by dynamically predicting

paths with greater temporal consistency and accuracy through memory association and a mask buffer.

3) Validated through rigorous testing on four unique off-road datasets, the framework demonstrated superior capabilities in path prediction, showcasing strong performance even without extensive training data, highlighting its adaptability and potential for generalization.

PathFormer integrates a novel path planning technique with a learning module to enable adaptive, flexible navigation for autonomous vehicles across diverse terrains. Subsequent sections detail the methodology, experimental setup, and results achieved in the proposed research.

II. Proposed Methodology

In this research, we propose a transformer-based framework, dubbed *PathFormer*, for segmenting the free-space and predicting the traversable path for autonomous navigation in off-road scenarios. The proposed framework consists of four main modules, as shown in Fig. 1. The framework starts with a learnable module that produces initial segmentation masks. These masks are then refined and temporally validated through memory association module. Subsequently, mask buffer manages temporal mask data, assisting in the generation of predicted paths aligned with ground truth. Lastly, the path generation utilizes the temporally coherent free-space mask to delineate the navigable path. The detailed explanation of all modules is presented next.

A. Learnable Module

The proposed learnable module is a hybrid architecture designed for detecting and segmenting traversable paths. It is inspired from state-of-the-art models [25], [26], [27]. The learnable module employs a CNN backbone for initial feature extraction from images. These features are further processed by a transformer architecture with a deformable attention mechanism applied to both encoder and decoder, as illustrated in Fig. 2. This approach significantly improves the model's ability to concentrate on relevant spatial locations, enhancing the efficiency of detection and segmentation. The workings of these adaptations in the encoder and decoder, particularly the utilization of the deformable attention mechanism, are explained in the subsequent sections.

1) Backbone: The learnable module begins with a backbone structure, which operates on an input image of resolution $H \times W \times 3$. This image is subjected to a transformation through a CNN backbone. In this work, we employ ResNet-50 [28] as the backbone network for extracting a series of feature maps, specifically from latter three stages of the ResNet-50 model, termed S_3 to S_5. This approach effectively leverages the robust feature extraction capabilities of ResNet-50 for capturing detailed and varied features of path class across different scales. Next, we employ a feature transform block in the learnable module. It refines the feature maps (S_3 to S_5) by a 1×1 convolution, where each S_j corresponds to a resolution of 2^j lower than the original image. The coarsest resolution map is generated through a 3×3 stride 2 convolution on the last S_5 stage, denoted as S_6. All feature maps hold a similar channel depth of $z = 256$.

These multi-scale feature maps, denoted as $\{f^j\}_{j=1}^{J}$, with each f^j being a representation of the feature map at scale j and dimension $R^{H/j \times W/j \times z}$, are pivotal in capturing the diverse scales and complexities inherent in the image. Consequent to the extraction of these multi-scale feature maps from the backbone, we have generated embeddings for each level j. These embeddings are comprehensive, incorporating not only positional information $\{\rho^j\}_{j=1}^{J}$ but also the scale data $\{\delta^j\}_{j=1}^{J}$, thus rendering a holistic spatial representation at each feature map level. Unlike the positional embedding which is fixed, the scale-level embeddings, $\{\delta^j\}_{j=1}^{J}$, are initialized randomly and are fine-tuned in conjunction with the network during the learning process. The multi-scale feature maps along with the embeddings are passed on to the transformer encoder, as explained in the next section.

2) Transformer Encoder: Next, in the proposed learnable module, we incorporated transformer layers. Transformers have had a profound impact on computer vision, primarily owing to the inclusion of the multi-head attention mechanism. It facilitates the aggregation of key information based on attention weights, which gauge the compatibility of query-key pairs. Consequently, the model can selectively focus on diverse features originating from different representation subspaces and spatial positions. Specifically, for a query element q indexed by a set Q with representation feature $h_q \in R^{f_d}$, and a key element k indexed by a set K with representation feature $h_k \in R^{f_d}$, where f_d is the feature dimension. Then, the multi-head attention (MHA) feature is computed as follows:

$$\text{MHA}(Q, K) = \sum_{i=1}^{I} W_i^Q \left[\sum_{k \in K} \beta_{iqk} \cdot (W_i^K h_k) \right] \quad (1)$$

where I indexes the attention heads, and $W_i^Q, W_i^K \in \mathbb{R}^{f_d \times f'_d}$ are learnable weights, with $f'_d = \frac{f_d}{I}$. The attention weights β_{iqk} are calculated using a softmax function applied to the scaled dot-product of the query with all keys:

$$\beta_{iqk} = \frac{\exp(h_q^T U_i^{qT} V_i^{kT} h_k)}{\sum_{k' \in K} \exp(h_q^T U_i^{qT} V_i^{kT} h_{k'})} \quad (2)$$

where $U_i^q, V_i^k \in R^{f_d \times f'_d}$ are also learnable matrices. This multi-head attention aggregates information across different positional embeddings and feature subspaces, but it is computationally intensive and lacks the inherent capability to prioritize spatial features, which is especially detrimental in tasks involving high-dimensional data such as images. In this work, we employ multi-scale deformable attention (MSDA) as an alternative to the conventional MHA in both the encoder and decoder layers of the learnable module. This module builds upon the principles of MHA but incorporates deformable sampling points to selectively attend to key features, as shown in Fig. 2. It significantly reduces the computational burden and enables spatial feature prioritization, which is essential for handling high-resolution images and detecting objects at varying scales.

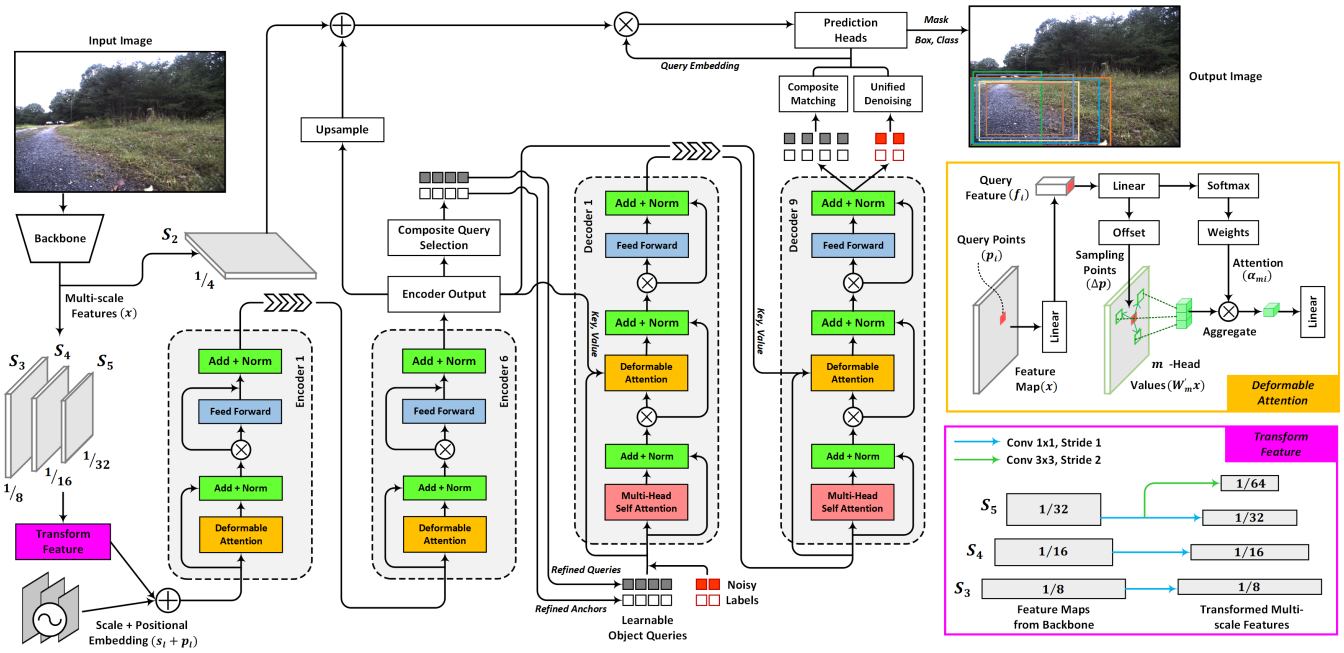

Fig. 2. Architecture of the PathFormer framework's hybrid learning module, combining CNNs and transformers for path detection and segmentation.

Given multi-scale feature maps $\{f^j\}_{j=1}^J$, where $f^j \in R^{H/j \times W/j \times z}$. Let a query element indexed by q with content feature h_q and a normalized 2D reference point coordinates $p_q \in [0,1]^2$, then MSDA is computed as:

$$\text{MSDA}(h_q, p_q, \{f^j\}_{j=1}^J) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \alpha_{ijqk}, \quad (3)$$

$$\alpha_{ijqk} = \left[W_i^Q \beta_{ijqk} \cdot W_i^K f^j\left(\phi(p_q) + \delta_{ijqk}\right)\right]. \quad (4)$$

Here, h_q represents the query feature, and p_q denotes the normalized coordinate of the reference point for the query. β_{ijqk} and δ_{ijqk} are the attention weights and sampling offsets, respectively, for the k-th point in the i-th head at the j-th scale level. ϕ represents the mapping of the normalized position coordinates. In contrast to the traditional vision transformer encoder, our encoder exhibits a targeted and strategic focus. Rather than processing all spatial locations indiscriminately, it selectively concentrates on specific sampling points. This methodological approach results in a notable reduction in computational demand and an enhancement in the precision and relevance of the extracted features.

3) Transformer Decoder: The decoder's role in the learnable module is to interpret the feature-rich data from the encoder and translate it into discernible detection and segmentation outputs. The inputs to the decoder comprise object queries and the multi-scale feature maps from the encoder. The decoder utilizes a variant of MSDA for its cross-attention mechanism, tailored to synergistically integrate information from both the encoder outputs and the object queries.

The final phase of the decoder involves the prediction heads. These components process the decoder's output and yield the final detection results, including bounding boxes of traversable paths and segmentation masks of the free-space regions. The design and functionality of these prediction heads are critical, as they convert the abstract features and attention weights into tangible, interpretable outputs.

B. Mask Refinement

The Mask Refinement process is initiated following the transformer-based learnable module, which outputs a sequence of instances, each comprising a bounding box and an associated segmentation mask, alongside a detection score. This score signifies the model's confidence in accurately classifying the instance as a path. In this stage, we aim to discern the most reliable mask for representing navigable spaces within off-road environments through a systematic refinement process.

Initially, we narrow down the selection of instances by focusing on the segmented area. This primary filtration step ensures that only masks depicting areas within a specific, predefined range are considered, effectively excluding instances that are either too small or excessively large to hold navigational relevance, as expressed:

$$M' = \{m_i \mid \Theta, \forall m_i \in M\}, \quad (5)$$

$$\Theta = (A_{\text{min}} < \text{Area}(m_i) < A_{\text{max}}). \quad (6)$$

Here, M denotes the set containing initial segmentation masks generated by the learnable module. A_{min} and A_{max} represent predefined minimum and maximum area thresholds, respectively. The function $\text{Area}(m_i)$ computes the area of the mask m_i, ensuring that only those masks representing areas within the specified range are included in M'. These thresholds are empirically specified after careful observation of the total area of navigable free-space paths across different datasets. The determination of A_{min} and A_{max} is crucial for

ensuring that the refined masks, M' , accurately represent significant navigable spaces. This empirical specification addresses instances of under-segmentation, where very small areas are incorrectly identified as free-space masks with high confidence scores, and cases of over-segmentation, which result in the erroneous designation of large image regions as navigable free-space paths. By setting these thresholds (A_{min} and A_{max}), the refinement process effectively mitigates these issues, ensuring that only masks of a practical size for navigation are carried forward for further refinement and eventual path generation.

Subsequently, from the reduced mask instances, M', we select the mask with the highest confidence score exceeding a designated threshold, C_{thresh}, as the optimal mask for additional refinement:

$$m_{\text{optimal}} = \underset{m_i \in M'}{\text{argmax}}\ C(m_i),\ \text{if } C(m_i) > C_{\text{thresh}} \tag{7}$$

If none of the masks surpass C_{thresh}, suggesting that the confidence levels fall short of the acceptable threshold, all masks for that frame are deemed unreliable. In such scenarios, if the temporal buffer, B, is populated, the system defaults to the last reliable mask from B. Otherwise, the framework continues to search for a reliable mask.

After identifying the optimal mask, m_{optimal}, from the subset of area and confidence-score filtered instances, the next critical step involves isolating the primary navigable space. This isolation is accomplished through a connected components analysis, which aims to identify the largest contiguous area within m_{optimal}. This step ensures focus on the most substantial navigable region for path prediction. The isolation of this key navigable space is formulated as:

$$m_{\text{final}} = \underset{k}{\text{argmax}}\ \text{Area}(C_k(m_{\text{optimal}})) \tag{8}$$

where $C_k(m_{\text{optimal}})$ represents the k-th connected component within the optimal mask m_{optimal}, and $\text{Area}(C_k(m_{\text{optimal}}))$ calculates the area of this component. This approach emphasizes the identification and isolation of the most significant navigable space, delineated by the largest area, to serve as the final free-space path mask, m_{final}, for subsequent stages.

The Mask Refinement process plays a crucial role, particularly in the context of initializing the temporal buffer, Bproviding a high-quality navigational cue integral for the temporal consistency and reliability of path predictions across challenging off-road terrains.

C. Memory Association

In memory association, we ensure the temporal consistency of navigational paths following the mask refinement process. Here, we utilize a historical buffer, B, consisting of masks that have been previously accepted, to assess each current refined mask, m_{current}. A foundational aspect of this phase is the initial condition of the buffer. When B is initially empty, it signifies the commencement of the navigation process, necessitating the addition of m_{optimal} directly to B, as explained earlier. This action establishes a reference point for subsequent evaluations of temporal consistency, ensuring B evolves dynamically to encapsulate the most recent and pertinent navigational information.

As the system progresses, m_{t} is evaluated for integration into B by examining its overlap with each mask in the buffer, $b_i \in B$, employing the Intersection over Union (IoU) metric. The mean IoU, $\overline{\text{IoU}}$, is calculated as follows:

$$\overline{\text{IoU}} = \frac{1}{|B|} \sum_{i=1}^{|B|} \text{IoU}(m_{\text{t}}, b_i) \tag{9}$$

where the IoU between the current mask and a buffered mask is quantified by:

$$\text{IoU}(m, b) = \frac{|m \cap b|}{|m \cup b|} \tag{10}$$

If $\overline{\text{IoU}}$ exceeds a predefined threshold, T_{IoU}, m_{t} is deemed temporally consistent and is added to B, enhancing the buffer with the latest navigational insight. This decision signifies the current mask's alignment with the historical trajectory, affirming its utility for subsequent path predictions.

Conversely, if $\overline{\text{IoU}}$ falls below T_{IoU}, indicating a lack of sufficient overlap with historical data, the system reverts to using the most recent mask from B for further processing. This approach ensures that navigation remains grounded in the most reliable and temporally consistent data available, thereby maintaining the integrity of path predictions amidst changing environmental conditions.

$$m_{\text{selected}} = \begin{cases} m_{\text{t}}, & \text{if } \overline{\text{IoU}} > T_{\text{IoU}} \\ b_{\text{t-1}}, & \text{otherwise} \end{cases} \tag{11}$$

To adapt to environmental changes and maintain the continuity of path prediction, B is dynamically managed, retaining only the most recent and relevant masks. This ensures the buffer remains an accurate reflection of the latest off-road navigational circumstances, critical for the system's adaptability and predictive accuracy.

D. Path Generation

In the final phase, Path Generation, the navigable path is extrapolated from the temporally coherent mask, m_{selected}. This phase centers on extracting a central trajectory from the refined mask, which is subsequently refined through a series of sophisticated smoothing operations.

The process initiates by identifying the central axis of the free-space mask, leading to the creation of a preliminary path, P_1. This path reflects the spatial distribution of navigable areas within the mask:

$$P_1 = \left\{ (x_i, y_i) \mid x_i = \frac{1}{N} \sum_{j=1}^{N} x_{ij},\ \forall y_i \right\} \tag{12}$$

where x_i denotes the average x-coordinate for each horizontal slice of the mask at a given y_i, and N is the number of points in each slice that are part of the navigable space.

This preliminary path P_1 undergoes refinement through a Savitzky-Golay filter, followed by additional smoothing techniques tailored to the specific characteristics of the path

and the off-road terrain. The dual-phase smoothing process is mathematically represented as:

$$P_{\text{smooth}}(x) = \frac{1}{2k+1} \sum_{n=-k}^{k} P_{\text{SG}}(x+n) \tag{13}$$

where P_{SG} is the output after applying the Savitzky-Golay filter to P_1 with window size w and polynomial order p, and P_{smooth} denotes an additional smoothing step applied directly to P_{SG}, enhancing its navigability by averaging each point x with its $2k$ neighboring points, where k determines the extent of the smoothing window. This approach, which balances fidelity to the detected navigable space with the need for a practical path, ensures P_{smooth} is optimized for navigation, making it responsive to the immediate environmental context and sufficiently smooth for realistic path following by autonomous vehicles.

The proposed approach to path generation, rooted in the immediate data from m_{final} and the historical context provided by the Memory Association phase, culminates in the derivation of a navigable path that is precise, adaptable, and suited for the challenges of off-road navigation. The balancing act between accurately representing traversable space and ensuring the path is practical for the challenges of off-road navigation.

III. Experimental Setup

A. Datasets

We utilized three distinct datasets for model training, each selected for its unique representation of terrain types, lighting conditions, and trail characteristics.

1) Off-Road Open Desert Trail Detection (O2DTD): The O2DTD dataset, containing 5045 images from desert environments, acquired to capture the nuanced light and shadow conditions specific to desert terrains at six different times of the day. It provides pixel-level annotations to differentiate between the sky, ground, and trails. For our purposes, trail annotations were specifically employed to define the *Path*, while all other categories were considered background.

2) Robot Unstructured Ground Driving (RUGD): The RUGD dataset [29] contains approximately 7436 images, originally classified into 24 categories including eight types of terrain. We streamlined these categories by focusing on the *Path* labels and reassigning the remaining as background to fit our analysis needs.

3) Off-Road Freespace Detection (ORFD): The ORFD dataset [30] features 12197 images with a broad spectrum of off-road conditions. The annotations include three main categories: traversable, non-traversable, and unreachable areas. We designated the traversable category as the *Path* and reclassified the other two as background for our study.

While the O2DTD dataset includes path trajectory labels, the RUGD and ORFD datasets do not contain them. To maintain uniformity across datasets for analysis, we generated these path labels for the test sets in RUGD and ORFD. Our methodology for generating ground truth paths involved an initial automated extraction of center-points from segmented paths, subsequently refined with precise manual corrections to ensure the accuracy and reliability of the ground truth paths. Additionally, we assessed the model's effectiveness using a blind test approach. This involved evaluating the model with one additional off-road datasets from real-world settings that was not previously encountered during its training. This approach rigorously tests the model's generalization capability and resilience. Detailed descriptions and specifications of each dataset and their respective subsets are mentioned in Table I.

TABLE I
DATASET DETAILS

Dataset	Training	Validation	Test
O2DTD	3235	504	1009
RUGD[a]	5235	280	1921
ORFD	8397	1245	2555
CAT[b]	-	-	544

[a] Randomly selected sequences for test ('trail-3', 'trail-7', 'trail-12', 'trail-15'). [b] Used only the test set to blind test model performance.

B. Training Configuration

1) Hyperparameters Selection: The training of the PathFormer framework employed the AdamW optimizer, setting the learning rate to 2.5×10^{-4}. The training process spanned over 2×10^5 iterations, with batches composed of 12 images each. This setup, executed in Python 3.8.17 and utilizing PyTorch 2.0.1, was strategically chosen to achieve an optimal balance between the speed of convergence and the accuracy of the model, ensuring efficient use of computational resources.

2) Composite Loss Function: For the task of simultaneous detection and segmentation within our framework, we adopted a composite loss function strategy akin to that described in [25]. This involves the integration of detection (L_{det}), box regression (L_{reg}), and mask segmentation (L_{seg}) losses. Focal loss (L_{focal})is utilized for detection accuracy, while box regression is refined using a combination of L1 (L_{L1}) and Generalized Intersection over Union (L_{GIoU}) losses. Segmentation accuracy is enhanced by merging cross-entropy (L_{cross}) with dice loss (L_{Dice}). The aggregate loss (L_{total}) is expressed as follows:

$$L_{\text{total}} = L_{\text{det}} + L_{\text{reg}} + L_{\text{seg}}, \tag{14}$$

$$L_{\text{det}} = \lambda_{\text{focal}} \cdot L_{\text{focal}}, \tag{15}$$

$$L_{\text{reg}} = \lambda_{\text{L1}} \cdot L_{\text{L1}} + \lambda_{\text{GIoU}} \cdot L_{\text{GIoU}}, \tag{16}$$

$$L_{\text{seg}} = \lambda_{\text{cross}} \cdot L_{\text{cross}} + \lambda_{\text{Dice}} \cdot L_{\text{Dice}}. \tag{17}$$

where the weighting factors, $\lambda_{\text{focal}} = 4$, $\lambda_{\text{L1}} = 5$, $\lambda_{\text{GIoU}} = 2$, $\lambda_{\text{cross}} = 5$, and $\lambda_{\text{Dice}} = 5$, are empirically set to ensure an effective balance across the model's learning objectives.

IV. Results

A. Qualitative Analysis

This section presents the qualitative analysis to assess the performance of the proposed PathFormer model.

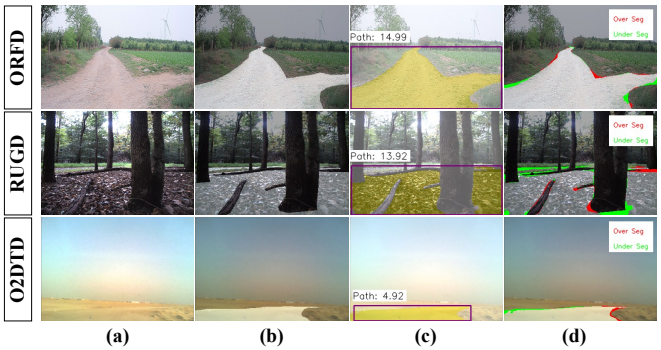

Fig. 3. Path segmentation results are shown on randomly selected images from all three test datasets. (a) Original images, (b) Ground-truth labels, (c) Segmented path predictions, and (d) Under-over segmentation are shown in red and blue color respectively.

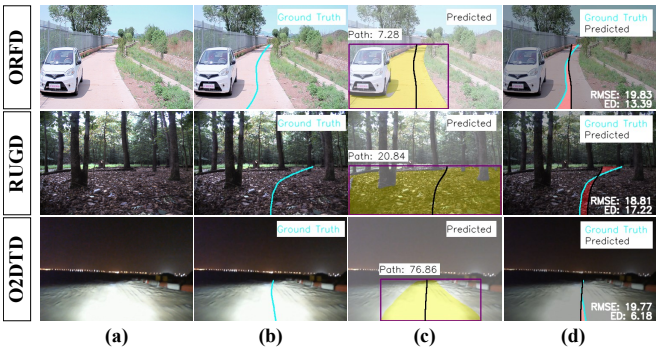

Fig. 4. Path prediction results are shown on randomly selected images from all three test datasets. (a) Original images, (b) Ground-truth labels, (c) Extracted paths, and (d) Overlayed ground-truth and predicted paths.

1) Segmentation Performance: Fig. 3 shows a detailed visual comparison encompassing the original off-road scenario, the ground truth, and the free-space segmentation masks by the model. Here, it can be observed that the PathFormer model adeptly identifies most features of the off-road environment, with its segmentation predictions aligning closely with the ground truth. This underscores its efficiency in recognizing and segmenting intricate path details crucial for navigation. Nonetheless, occurrences of under-segmentation—missing parts of the path—and over-segmentation—identifying non-path areas as path—indicate areas for future improvement. Rectifying these issues is imperative for improving the model's precision and ensuring safer off-road navigation.

2) Path Prediction Performance: Fig. 4 presents path prediction results using segmented free-space regions by the model. The results clearly illustrate the model's ability to accurately follow the free-space segments, showcasing its adeptness in navigating the landscape and identifying potential obstacles. The alignment of the model's path predictions algorithm with the segmented regions emphasizes its effectiveness for real-time navigation in off-road conditions. However, there are instances where the predicted paths diverge from the segmented terrains, indicating potential over-reliance on the segmentation phase or a failure to fully integrate critical environmental details. These discrepancies point to the necessity of further enhancements to the model, aiming to improve its reliability and ensure safe navigation through unstructured environments.

B. Quantitative Analysis

This section details the quantitative evaluation of the PathFormer framework.

1) Detection Performance: Table II shows the detection performance of the PathFormer framework using Average Precision (AP) at two IoU thresholds (AP50 and AP95). Additionally, we computed an aggregated AP score spanning an IoU range from 0.50 to 0.95, incrementing by 0.05. Notably, the PathFormer model was trained across three diverse datasets, encompassing various off-road terrain categories, weather conditions, and daylight times. The findings indicate superior performance on the O2DTD dataset, with a gradual decrease observed moving to the ORFD and subsequently the RUGD dataset. The lower performance on the RUGD dataset may highlight the intricate terrain characteristics and the broad diversity of vegetation classes present. Moreover, we can observe a notable reduction in AP values as the IoU threshold increases, particularly evident at the AP95 threshold. This pattern across the datasets underlines the model's adaptability to different environments but also highlights areas where enhancements are needed, particularly in dealing with the RUGD dataset's complexities. Overall, the aggregate AP scores offer a comprehensive view of the model's general performance across diverse scenarios, underscoring the necessity for further refinements to improve its versatility and effectiveness in navigating off-road conditions.

TABLE II
PATHFORMER PERFORMANCE ON TEST DATASETS

Dataset	Detection			Segmentation		Path	
	AP50	AP95	AP	DSC	IoU	RMSE	ED
O2DTD	0.9988	0.3412	0.5781	0.9674	0.9345	53.02	19.81
RUGD	0.7552	0.3002	0.4603	0.7831	0.6568	106.26	71.16
ORFD	0.8162	0.4194	0.5520	0.8643	0.8289	98.92	62.48

2) Segmentation Performance: This section examines the segmentation accuracy of the PathFormer framework on different test datasets, utilizing standard evaluation metrics: the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). The results are reported in Table II, where we can see that the model achieved DSC and IoU scores of 0.8643 and 0.8289, respectively, on the ORFD dataset. These scores indicate a reliable performance, though certain conditions or terrain types within the ORFD dataset may present challenges that require further investigation and optimization.

The RUGD dataset, however, displays a lower performance, with DSC and IoU scores of 0.7831 and 0.6568, respectively. This suggests that the RUGD dataset comprises a broader variety of terrain and environmental conditions not entirely captured in the model's training phase. Examining the specific attributes of the RUGD dataset's terrain could offer critical insights for enhancing the model's training strategy.

Conversely, the model showcases best performance on the O2DTD dataset, with DSC and IoU scores of 0.9674 and 0.9345, respectively. The minimal variability in these scores points to the dataset's homogeneity, predominantly featuring desert terrains at different times of the day. The specialized focus of the O2DTD dataset likely facilitates the model's adaptation and performance, as it only needs to learn and identify a limited variety of lighting conditions and terrain features.

3) Path Prediction Performance: We assessed the path prediction accuracy of the PathFormer framework by utilizing root mean square error (RMSE) and Euclidean distance (ED) metrics, as detailed in Table II. Due to the large image sizes employed in the comparisons, the metrics in Table II may seem relatively high. The ORFD results indicate a notable difference between the predicted and actual paths, with an RMSE of 98.92 and an ED of 62.48. For the RUGD dataset, the errors are slightly larger, with an RMSE of 106.26 and an ED of 71.16, suggesting more significant discrepancies.

In contrast, the framework performs relatively well on the O2DTD dataset, achieving a lower RMSE of 53.02 and an ED of just 19.81. This shows that the model's predictions are closely aligned with the actual paths in the O2DTD dataset. Overall, the variation in path prediction accuracy across different datasets highlights the influence of dataset-specific challenges and the resolution of the images. While the results on the O2DTD dataset demonstrate the model's strengths, the outcomes on ORFD and RUGD point to potential areas for enhancement.

C. Comparison with State-of-the-Art

In this section, we compare the results of PathFormer framework with other models in the literature [31], [32], [33], [34], [35], [36], [37], [7], as illustrated in Table III. Our evaluation focuses on three key dimensions: segmentation accuracy, path prediction performance, and inference time. The aggregated results across all three off-road datasets are summarized in Table III.

The results indicate that the PathFormer framework consistently outperforms other methods. Notably, PathFormer shows a 3.68% improvement in segmenting free-space regions over the next best method. In terms of path prediction accuracy, PathFormer reduces the RMSE by 13.65% compared to the second-best approach.

D. Generalization Evaluation

Achieving strong generalization is crucial when deploying deep models in real-world scenarios. Therefore, we assessed the PathFormer's ability to adapt to out-of-distribution data. To explore this, we conducted an experiment where, after training, the model was tested on the CAT dataset [10], which represents an unseen real-world environment. The results, summarized in Table IV, demonstrate that PathFormer not only generalizes well to unseen data but also shows the potential to quickly adapt to new terrains with minimal additional training. This adaptability highlights the model's suitability for applications that require incremental learning and rapid adjustment to changing environments. Furthermore, Fig. IV provides a visual representation of PathFormer's performance on the CAT dataset, illustrating its domain adaptation capabilities.

TABLE III

PATHFORMER PERFORMANCE COMPARISON WITH OTHER METHODS

Method	Mask (IoU)	Path (RMSE)	Time (msec)
clipseg [32]	0.6429	131.81	352
OCRNet [37]	0.6119	148.16	336
OneFormer [34]	0.6475	131.38	435
PSPNet [35]	0.7461	110.11	376
CGNet [33]	0.5541	155.18	268
Mask2Former [31]	0.7385	109.92	385
GroupViT [36]	0.6899	122.80	419
TerrainSense [7]	0.7781	99.67	219
Ours	0.8067	86.07	237

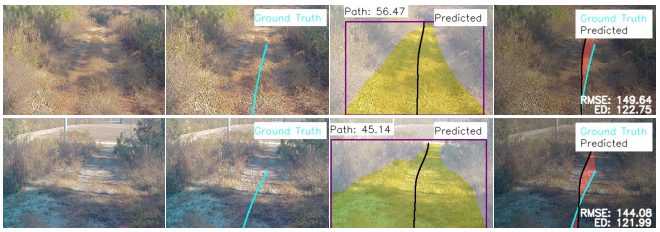

Fig. 5. Path segmentation and prediction results on unseen dataset. (a) Original images, (b) Ground-truth labels, (c) Segmented free space and extracted path, and (d) Overlayed ground-truth and predicted paths.

TABLE IV

PATHFORMER PERFORMANCE ON UNSEEN DATASET

Dataset	Detection			Segmentation		Path	
	AP50	AP95	AP	DSC	IoU	RMSE	ED
CAT [18]	0.5702	0.1378	0.1983	0.8165	0.6802	104.59	64.90

V. CONCLUSION

Navigating autonomous vehicles through unstructured and off-road terrains presents significant challenges, which impede the progression of autonomous driving technology. Traditional navigation methods, effective in the structured environments of urban settings, often falter amidst the variability and unpredictability of off-road landscapes. This research introduces the PathFormer framework, an innovative end-to-end approach leveraging vision-based, mapless strategies for traversing complex terrains. The effectiveness of PathFormer has been demonstrated through comprehensive analysis and evaluation, illustrating its ability to integrate the robustness of deep learning methodologies with the specific demands of off-road navigation. The future direction of this research aims to augment the framework by incorporating dynamic vehicle constraints, ensuring navigation solutions are grounded not only in visual data but also in alignment with the vehicle's physical characteristics and operational limits.

REFERENCES

[1] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 712–733, 2020.

[2] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 3, pp. 194–220, 2017.

[3] H. Ye, J. Mei, and Y. Hu, "M2f2-net: Multi-modal feature fusion for unstructured off-road freespace detection," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–7, IEEE, 2023.

[4] O. Mayuku, B. W. Surgenor, and J. A. Marshall, "A self-supervised near-to-far approach for terrain-adaptive off-road autonomous driving," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14054–14060, IEEE, 2021.

[5] D. W. Carruth, C. T. Walden, C. Goodin, and S. C. Fuller, "Challenges in low infrastructure and off-road automated driving," in *2022 Fifth International Conference on Connected and Autonomous Driving (MetroCAD)*, pp. 13–20, IEEE, 2022.

[6] X. Liang, T. Wang, L. Yang, and E. Xing, "Cirl: Controllable imitative reinforcement learning for vision-based self-driving," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 584–599, 2018.

[7] B. Hassan, A. Sharma, N. A. Madjid, M. Khonji, and J. Dias, "Terrainsense: Vision-driven mapless navigation for unstructured off-road environments," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 18229–18235, IEEE, 2024.

[8] C. Xu, W. Zhao, J. Liu, C. Wang, and C. Lv, "An integrated decision-making framework for highway autonomous driving using combined learning and rule-based algorithm," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 4, pp. 3621–3632, 2022.

[9] B. Gallazzi, P. Cudrano, M. Frosi, S. Mentasti, and M. Matteucci, "Clothoidal mapping of road line markings for autonomous driving high-definition maps," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1631–1638, IEEE, 2022.

[10] S. Sharma, L. Dabbiru, T. Hannis, G. Mason, D. W. Carruth, M. Doude, C. Goodin, C. Hudson, S. Ozier, J. E. Ball, *et al.*, "Cat: Cavs traversability dataset for off-road autonomous driving," *IEEE Access*, vol. 10, pp. 24759–24768, 2022.

[11] D. Maturana, P.-W. Chou, M. Uenoyama, and S. Scherer, "Real-time semantic mapping for autonomous off-road navigation," in *Field and Service Robotics: Results of the 11th International Conference*, pp. 335–350, Springer, 2018.

[12] J. S. Berrio, M. Shan, S. Worrall, and E. Nebot, "Camera-lidar integration: Probabilistic sensor fusion for semantic mapping," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7637–7652, 2021.

[13] R. Ahmed, A. Al Shehhi, B. Hassan, N. Werghi, and M. L. Seghier, "An appraisal of the performance of ai tools for chronic stroke lesion segmentation," *Computers in Biology and Medicine*, p. 107302, 2023.

[14] B. Hassan, S. Qin, T. Hassan, M. U. Akram, R. Ahmed, and N. Werghi, "Cdc-net: Cascaded decoupled convolutional network for lesion-assisted detection and grading of retinopathy using optical coherence tomography (oct) scans," *Biomedical Signal Processing and Control*, vol. 70, p. 103030, 2021.

[15] R. Ahmed, Y. Chen, B. Hassan, L. Du, T. Hassan, and J. Dias, "Hybrid machine-learning-based spectrum sensing and allocation with adaptive congestion-aware modeling in cr-assisted iov networks," *IEEE Internet of Things Journal*, vol. 9, no. 24, pp. 25100–25116, 2022.

[16] R. Ahmed, A. Al Shehhi, N. Werghi, and M. L. Seghier, "Segmentation of stroke lesions using transformers-augmented mri analysis," *Human Brain Mapping*, vol. 45, no. 11, p. e26803, 2024.

[17] X. Cai, M. Everett, J. Fink, and J. P. How, "Risk-aware off-road navigation via a learned speed distribution map," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2931–2937, IEEE, 2022.

[18] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, "Deep multispectral semantic scene understanding of forested environments using multimodal fusion," in *2016 International Symposium on Experimental Robotics*, pp. 465–477, Springer, 2017.

[19] K. Wong, Y. Gu, and S. Kamijo, "Mapping for autonomous driving: Opportunities and challenges," *IEEE Intelligent Transportation Systems Magazine*, vol. 13, no. 1, pp. 91–106, 2020.

[20] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive computing*, vol. 7, no. 4, pp. 12–18, 2008.

[21] L. Caltagirone, S. Scheidegger, L. Svensson, and M. Wahde, "Fast lidar-based road detection using fully convolutional neural networks," in *2017 ieee intelligent vehicles symposium (iv)*, pp. 1019–1024, IEEE, 2017.

[22] L. Sun, Z. Yan, A. Zaganidis, C. Zhao, and T. Duckett, "Recurrent-octomap: Learning state-based map refinement for long-term semantic mapping with 3-d-lidar data," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3749–3756, 2018.

[23] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "Rellis-3d dataset: Data, benchmarks and analysis," in *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 1110–1116, IEEE, 2021.

[24] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4644–4651, IEEE, 2017.

[25] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3041–3050, 2023.

[26] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4794–4803, 2022.

[27] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[29] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[30] C. Min, W. Jiang, D. Zhao, J. Xu, L. Xiao, Y. Nie, and B. Dai, "Orfd: A dataset and benchmark for off-road freespace detection," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2532–2538, IEEE, 2022.

[31] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

[32] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7086–7096, 2022.

[33] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "Cgnet: A light-weight context guided network for semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 1169–1179, 2020.

[34] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2989–2998, 2023.

[35] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

[36] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144, 2022.

[37] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 173–190, Springer, 2020.