Article

PortLaneNet: A Scene-Aware Model for Robust Lane Detection in Container Terminal Environments

Haixiong Ye 1, Zhichao Kang 1, Yue Zhou 1, Chenhe Zhang 2, Wei Wang 1 and Xiliang Zhang 3,*

1 College of Engineering Science and Technology, Shanghai Ocean University, Shanghai 201306, China; hxye@shou.edu.cn (H.Y.); m220851438@st.shou.edu.cn (W.W.)
2 Shanghai East Container Terminal Co., Ltd., Shanghai 200137, China
3 School of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, Shanghai 201209, China
* Correspondence: xlzhang@sspu.edu.cn

Abstract: In this paper, we introduce PortLaneNet, an optimized lane detection model specifically designed for the unique challenges of enclosed container terminal environments. Unlike conventional lane detection scenarios, this model addresses complexities such as intricate ground markings, tire crane lane lines, and various types of regional lines that significantly complicate detection tasks. Our approach includes the novel Scene Prior Perception Module, which leverages pre-training to provide essential prior information for more accurate lane detection. This module capitalizes on the enclosed nature of container terminals, where images from similar area scenes offer effective prior knowledge to enhance detection accuracy. Additionally, our model significantly improves understanding by integrating both high- and low-level image features through attention mechanisms, focusing on the critical components of lane detection. Through rigorous experimentation, PortLaneNet has demonstrated superior performance in port environments, outperforming traditional lane detection methods. The results confirm the effectiveness and superiority of our model in addressing the complex challenges of lane detection in such specific settings. Our work provides a valuable reference for solving lane detection issues in specialized environments and proposes new ideas and directions for future research.

Keywords: intelligent systems; container port; lane detection; deep learning; feature fusion

1. Introduction

In recent years, the synergy between the burgeoning automotive industry and artificial intelligence technology has led to significant advancements in computer vision technologies, which have become a pivotal part of autonomous driving and advanced driver-assistance systems (ADAS). Lane detection, serving as a fundamental requirement in the sensory modules of driving systems, is primarily executed through vision cameras mounted on vehicles, utilizing these advanced technologies for assistive positioning and navigation. Currently, lane detection technology utilizes the powerful iterative learning capabilities of deep learning and extensive data resources, achieving significant progress in common scenarios such as urban roads, characterized by high detection accuracy and low-latency real-time functionality.

The allure of automated container terminals has expanded due to their attributes of high efficiency, energy savings, reliability, and economic benefits, leading to an increasing number of ports globally adopting automation upgrades and transformations. The introduction of Automated Container Trucks (ACTs) has been a critical factor in the transition of traditional ports from being geographically confined comprehensive logistics hubs to dynamic catalysts for regional growth [1]. The automation of container terminals relies not only on efficient logistics management systems but also on precise and reliable autonomous driving technologies, including advanced lane detection, obstacle recognition, and vehicle
positioning capabilities. These technologies ensure that Automated Container Trucks can operate safely and effectively in the complex environment of container terminals, thereby enhancing the operational efficiency and safety of the entire terminal. Therefore, optimizing lane detection technology for port environments to meet their unique operational conditions is not only a necessity for technological development but also a crucial step towards advancing port automation and enhancing regional logistics efficiency. This transformation is heavily reliant on the continuous innovation in autonomous driving technologies and assistance systems, which are essential for the efficient and safe operation of ACTs within the complex environment of container terminals. The continuous innovation in autonomous driving technologies and assistance systems plays a key role in this transition.

However, challenges persist in special environments like enclosed terminal areas. On one hand, the frequent intense operations of horizontal transport vehicles and heavy machinery in terminals can lead to temporary obstructions and wear on lane markings, as shown in Figure 1a,b, presenting non-standardized lane layouts. This poses higher demands on the adaptability and error tolerance of the model. Additionally, the terminal environment’s susceptibility to complex and variable weather conditions like dense fog and rain or snow affects the visibility of lane markings, further complicating detection efforts. Therefore, facing the dual challenges of lane markings themselves and variable weather, lane detection models need to possess high robustness and anti-interference capabilities to ensure effective vehicle operation in various special environments, which is vital for ensuring the operational safety and improving the efficiency of terminal operations. On the other hand, terminal environments feature unique characteristics, such as the complex ground markings shown in Figure 1c–e, which may visually resemble lane lines but serve entirely different functions. This urgently necessitates the development of new lane detection methods capable of accurately identifying lanes and effectively distinguishing these different types of ground markings. Traditional lane detection methods, often based on handcrafted features, may perform well in some environments but face limitations in the complex conditions of terminals.

![Figure 1. Challenges encountered in lane detection at container terminals.](image-url)
In light of these considerations, this paper innovatively proposes a lane detection method named PortLaneNet, optimized for the special environment of enclosed terminals, breaking through the limitations of traditional lane detection models. At the core of PortLaneNet is the introduction of a Scene Prior Perception Module, which provides crucial a priori information for lane detection. Furthermore, by integrating an attention mechanism, the model can focus more effectively on key areas crucial for lane detection.

Our contributions are summarized as follows:

Novel Lane Detection Method: Addressing the special requirements of complex environments at container ports, this paper builds upon the LaneATT model to design the PortLaneNet model’s architecture for ports’ diverse and complex environmental challenges, enhancing the model’s adaptability and accuracy in complex scenarios.

Scene Prior Perception Module: By incorporating feature fusion technology, the model is jointly trained with scene information and the lane detection task, enabling it to fully utilize high-level semantic information in the images, thereby further improving detection accuracy and adaptability.

Superior Performance: Through rigorous experiments, PortLaneNet has been demonstrated to outperform existing lane detection models in enclosed terminal environments, effectively addressing the detection challenges presented by such specialized settings.

The remainder of this paper will detail PortLaneNet’s research methods and experimental results. First, we review recent advancements in the field of lane detection. Then, we delve into the architectural design and methodology of the proposed model. Following that, we present a series of experiments conducted in enclosed container terminal environments and their outcomes, validating our model’s superior performance over existing methods in such specialized scenes. Finally, we discuss the research findings and propose potential directions for future work.

2. Related Works

2.1. Lane Detection Methods

Lane line detection is considered a pivotal technology in autonomous driving systems. According to surveys [2], over 60% of traffic accidents are associated with vehicles veering off their lanes, making the accurate detection of lane lines crucial for vehicular safety. Based on the representation of lane lines, existing deep learning-based lane detection methods can be categorized as follows:

(1) Image-segmentation-based methods [2–7] model lane detection as a pixel-level classification problem, categorizing each pixel as either a part of the lane region or the background. These methods can be trained end-to-end, but they typically require prediction for every pixel in the image, which is highly computationally intensive. For example, the lane detection network CurveLane_NAS [3] demands 5000 GPU hours for a single dataset. Representative methods based on image segmentation include SCNN [2], ENet-SAD [4], and Bi-Lanenet [5]. To reduce computational burden, Hou et al. use knowledge distillation to compress a large network into a lightweight one in their ENet-SAD model. A recurrent shifting module is employed by Zheng et al. in their RESA method to aggregate contextual information. Concurrently, low-cost local prior knowledge is introduced by Qiu et al. [6] to enhance the segmentation performance of the PriorLane model. Despite these promising strategies employed by various models to reduce computational demands, tangible challenges persist, especially when deploying these models in on-board vehicle systems. The efforts to minimize computational resources, while notable, still often fall short in providing a pragmatic balance between computational efficiency and detection accuracy, particularly in real-time scenarios experienced in vehicular applications. Even with models like ENet-SAD and RESA, which have endeavored to lower computational costs, the requisite resources for real-time lane detection on vehicle-mounted hardware continue to be a significant barrier. The inherent complexity of image-segmentation methods, which necessitate high computational power to process pixel-level predictions across
entire images, could potentially limit their applicability in resource-constrained environments, such as those encountered in autonomous vehicles operating in bustling container terminals. Thus, achieving an optimal trade-off between computational efficiency and robust lane detection performance, especially in environments like container terminals with diverse and complex visual elements, remains a critical challenge for image-segmentation-based methods.

(2) Key point-based methods [8–11] predict and associate key points of lane lines, offering a more flexible representation of the lane line shape. Ko et al. [8] introduce PINet, which transforms the problem of clustering predicted key points into an instance segmentation problem by training multiple hourglass models simultaneously. Additionally, GANet, developed by Wang et al. [10], globally predicts the offset of key points from the starting point of the lane to realize global association. While these methods demonstrate flexibility and are adept in various scenarios, they might face challenges in container terminals, where complex ground markings and lane wear could hinder accurate key point prediction and association, especially in conditions of poor visibility or degraded lane markings.

(3) Parameter curve-based methods [12–15] fit lane lines with polynomial functions, with the detection network directly predicting these parameters. Since only a few parameters need to be predicted, these methods have fewer parameters and faster inference speeds, but they are more sensitive to parameter prediction. Representative methods include PolyLaneNet [13] and BezierLaneNet [12]. While these methods exhibit commendable performance in curve lane detection owing to their need to predict only a minimal number of parameters, thus ensuring fewer parameters and accelerated inference speeds, their sensitivity to parameter prediction poses a limitation. Particularly in dock environments, where lane lines are predominantly straight and often subject to wear and obstruction due to the frequent movement of heavy vehicles, coupled with the inherent sensitivity of these methods to parameter prediction, there is potential for compromised lane prediction accuracy in complex scenarios. This limits their reliability in ensuring the safe navigation of autonomous driving vehicles through container terminals.

(4) Anchor-based methods are primarily categorized into two types: line anchor-based detection [16–19] and row anchor-based detection [20–22]. Line anchor-based methods utilize the prior shape of lane lines and pre-set anchor points at the image borders for classification and regression, offering high efficiency. Notable methods in this category include Line-CNN [16], LaneATT [17], and CLRNet [18]. A dynamic head and self-attention module were designed in DILane [19], utilizing cross-layer features and global context information. A cross-layer feature optimization strategy was proposed in CLRNet. Row anchor-based detection methods use the prior position of the lane lines, making row-by-row predictions for the lane lines. This approach is extremely fast, but it may sacrifice a degree of accuracy. UFLD [20] is a representative method in this category, treating the lane detection process as a row-based selection problem using global features, greatly reducing the model’s computational cost.

Anchor-based approaches, especially those employing line anchors, utilize a top-down design principle, capitalizing on the inherent prior knowledge of lane lines. This design not only boosts real-time detection capabilities but also ensures continuity in lane line detection even under challenging conditions such as significant occlusions. While the predefined anchor shapes might seem to compromise detection flexibility, it is essential to recognize that our primary focus lies in the container terminal environments. In these settings, the lane structures are relatively consistent and less complex. Consequently, the flexibility of the model is not a paramount concern. Instead, the anchor-based design provides an efficient and stable lane detection solution tailored specifically for these container terminal scenarios.
2.2. Semantic Understanding for Lane Detection

Although lane detection technology has made significant progress in recent years, it still faces formidable challenges when dealing with complex scenes. Enhancing the capability for profound semantic understanding can markedly boost the algorithm’s detection accuracy under challenging conditions such as occlusions, varying illumination, and intricate lane markings. Especially in dock environments, the frequent movement of heavy vehicles, designated parking areas, and the presence of other ground markings underline the critical importance of a deep capability for semantic understanding of the scene. Numerous studies have sought to amplify this capability within networks:

SCNN [2] proposes a message-passing mechanism to exchange information between different feature layers, thereby enhancing semantics. CurveLanes [3] uses an attention module to aggregate global context features. CondLaneNet [22] introduces conditional convolution, which selectively enhances features according to prior knowledge. Point2Lane [15] presents a spatial context-aware feature flip fusion (SCFF) module, further refining the capability for nuanced semantic interpretation in complex driving scenarios. This advancement aids the network in better understanding lane positions even in the absence of direct visual clues, enhancing lane detection accuracy under challenging conditions. Meanwhile, PriorLane [6] designs a knowledge-encoding module that encodes prior knowledge into cues inputted into the network, thereby enhancing the network’s semantic understanding of complex scenes. Existing research reveals that enhancing a network’s deep semantic understanding of the environment plays a crucial role in improving the robustness of lane detection. Faced with variable actual conditions, a deeply understanding network can more accurately identify and predict lane lines, thus enhancing the precision and stability of detection. This paper also starts from the perspective of enhancing the network’s understanding of scene semantics, aiming at the enclosed dock environment. It is hoped that through the Scene Prior Perception Module, the network can establish a better semantic cognition, so as to better solve the problem of lane detection in such complex scenarios.

2.3. Attention Mechanisms

Attention mechanisms are widely applied in computer vision tasks to aggregate long-range context information. In lane detection tasks, attention mechanisms such as transformers [23] have been extensively explored. LSTR [24] is an early work that introduced attention mechanisms, building a network through transformers to learn richer structures and context information. LaneATT applies attention mechanisms on anchor box representations to aggregate local features and generate global features. Fastdraw [9] designed an attention module based on the start point and achieved good detection results. Recently, PriorLane [6] proposed a mixed transformer structure that integrates the knowledge encoder layer and fusion encoder layer, providing explicit prior knowledge for lane detection. Additionally, CurveFormer [25] offers a single-stage transformer-based approach that directly calculates 3D lane parameters, bypassing the challenging view transformation step typically required in other methods.

In dock environments, the scene is intricately dynamic due to the presence of various activities and objects, making lane delineation increasingly challenging. Factors such as frequent cargo handling and the movement of large machinery can cause sporadic visual occlusions. In this context, the attention mechanism proves invaluable by enabling the model to focus on essential lane line features while effectively sidelining potentially confusing, unrelated elements. Moreover, given the docks’ susceptibility to uneven lighting and varying weather conditions, attention mechanisms ensure the model zeroes in on pivotal features under diverse illumination scenarios, bolstering detection robustness. These insights underscore the pivotal role of attention mechanisms in enhancing lane feature representation and detection accuracy amidst the unique challenges presented by docks. In alignment with these findings, this paper integrates attention modules to capture holistic information, specifically tailored to address the intricacies of dock environments.
In summary, current methodologies provide exemplary frameworks for enhancing semantic understanding and applying attention mechanisms. This paper builds on these insights and refines the structure of existing lane detection networks. We design our modules with reference to these experiences, with the objective of augmenting the model’s capacity for semantic understanding and its robustness. The intention is to ensure that our model is not only aware of the intricate nuances that define the lanes in different scenarios but is also resilient to the variances found in real-world applications, thus promising a reliable performance in diverse environments.

3. Methodology

In this chapter, we present PortLaneNet, a real-time lane detection model specifically designed for dock environments. PortLaneNet takes the RGB images \( I \in \mathbb{R}^{3 \times H \times W} \) captured by a monocular camera mounted on a dock container truck as input, and predicts the lanes \( L = \{l_1, l_2, \ldots, l_N\} \). The overall architecture of the algorithm is shown in Figure 2. At the input layer, we utilize RGB images captured by a camera. Feature extraction is performed by a ResNet-based backbone, and a pyramid structure is employed to extract and integrate features from high to low levels, denoted as \( L_0, L_1, \) and \( L_2, \) respectively. The Scene Prior Perception Module processes features from the \( L_0 \) layer to perform a scene classification task. Learnable anchors are used for anchor-based feature pooling, further enhancing the model’s performance. The spatial attention module is responsible for feature fusion, combining \( F_L^{0} \) and \( F_L^{2} \) feature maps through weighted summation and element-wise multiplication to generate the ultimate fused feature map \( F_{\text{final}} \). Finally, the model outputs for the lane detection classification task \( (L_{\text{cls}}) \) and regression task \( (L_{\text{reg}}) \) are produced.

Figure 2. Overview of the proposed PortLaneNet framework.

3.1. Lane Representation

We adopt a point set representation for lane lines, specifically, each lane line is composed of coordinates of \( N \) points: \( l = \{P_0, P_1, \ldots, P_N\} \), where each point \( P_i = (x_i, y_i) \) represents a key point on the lane line. This representation method directly regresses on points, avoiding bias in parameter fitting, and since each point is independent, it has stronger robustness to deformed lanes. Compared with other representation methods, such as parameter curve representation, it requires less computation and is easier to optimize. In the implementation process, we uniformly sample \( N \) points in the \( y \)-axis direction to give...
the point set a certain regularity. During the training process, we directly regress the offset of each point to make the predicted point set as close as possible to the true lane line.

3.2. Backbone

In this study, the backbone of our model relies on ResNet [26] for feature extraction, specifically employing variants such as ResNet-18 and ResNet-34. Known for their deep residual learning framework, these architectures incorporate residual blocks with skip connections that facilitate the flow of gradients, effectively addressing the challenge of vanishing gradients during the training of deep neural networks. The selection of ResNet-18 and ResNet-34 is motivated by their capability to offer robust feature extraction while maintaining a lower computational complexity, making them ideal for real-time image processing tasks.

Furthermore, we have designed a pyramid structure [27] to extract and integrate features of different levels. In this design, we have three levels of features from high level to low level, denoted as L0, L1, and L2, respectively. To enhance the model’s performance, we perform pooling on the features of the L0 and L2 layers and fuse these features, improving the model’s representational capability. Moreover, to augment the model’s understanding of the scene, we input the L0 layer’s features into a Scene Prior Perception Module for conducting a scene classification task. This approach allows the model to utilize low-level features to extract basic structural information of the image, and also understand the image’s more complex semantic information via high-level features, thereby improving the accuracy and robustness of our model.

3.3. Feature Pooling Based on Learnable Anchors

In our design, we have adopted a crucial characteristic from the DILane model, which involves the use of a limited number of learnable anchor points as opposed to the conventional practice of using a large number of predefined anchor points. As the distribution of lane positions exhibits certain statistical regularities, this strategy allows our model to adapt more flexibly to the characteristics of lane lines in a confined container terminal environment. These anchor points are dynamically adjusted during the training process through the backpropagation algorithm, in order to align more accurately with the lane lines. We believe that this approach not only facilitates more accurate lane localization by our model and avoids misidentification of other ground markings as lane lines, but also enhances the computational efficiency of the model.

In the task of lane line detection, we define each anchor point as a four-dimensional vector, \( \text{reg}_i = \{S_{ix}, S_{iy}, \theta_i, \text{len}_i\} \), where \( S_{ix}, S_{iy} \) are the normalized \( x \) and \( y \) coordinates of the starting point, \( \theta_i \) represents direction, and \( \text{len}_i \) represents length. For each fixed \( y^j_k = k \cdot \frac{H_i}{K-1} \), each predicted \( x \)-coordinate can be calculated through the following formula:

\[
\hat{x}^j_k = \frac{1}{\tan \theta_i} \cdot \left( y^j_k - S_{iy} \right) + S_{ix} + \Delta x^j_k
\]

where \( \Delta x^j_k \) is the predicted offset relative to the anchor point. This parameter allows our model to dynamically adjust the position of each anchor point corresponding to the lane line, thereby capturing the actual trajectory of the lane line more accurately. These learnable anchor point parameters, \( \text{reg}_i \), are dynamically updated during the training process via the backpropagation algorithm. We have chosen a straightforward initialization strategy to allow for quicker algorithm convergence. As for the pooling process, we have referred to the pooling method used in the LaneATT, which realizes a single-stage detector by pooling with the anchor points themselves. This design enables our model to make more effective use of the global information in the feature map, not just the boundary information, thus further improving the model’s performance.
3.4. Attention Module

In the architecture of PortLaneNet, we have integrated a feature fusion strategy predicated on a spatial attention mechanism [28]. This strategy significantly enhances the model’s expressive power and its ability to discern lane line features. The spatial attention mechanism plays a pivotal role in the phase of feature fusion, where it can autonomously learn and allocate weights to different input feature maps in the final fusion outcome. The underlying premise of this approach is that high-level features carry a wealth of semantic information, while low-level features encapsulate a myriad of detail-oriented information. By fusing these two types of features, we can obtain a feature representation that is both semantically rich and detail-preserving, which is of paramount importance for the task of lane detection. The input to our spatial attention module consists of two sets of feature maps pooled from both the upsampled results from L0 and L2. These feature maps undergo an initial transformation through a fully connected (FC) layer, subsequently reshaped to achieve spatial congruence. This process leads to the formation of an initial fused feature map, \( F_{\text{init}} \), by performing a weighted summation of \( F_{L0} \) and \( F_{L2} \) feature maps:

\[
F_{\text{init}} = F_{L0} \oplus F_{L2}
\]

where \( \oplus \) signifies the weighted summation operation, integrating these distinct sets of feature maps into a unified representation.

To determine the significance of each pixel within this fused map, pixel-level self-attention scores, \( S \), are computed, highlighting areas of interest:

\[
S = \text{softmax}(W_s \cdot F_{\text{init}} + b_s)
\]

In this formula, \( W_s \) and \( b_s \) are the weight matrix and bias term of the spatial attention module, respectively. The softmax function is applied to these self-attention scores, normalizing them into a probability distribution that ranges between 0 and 1. Each pixel’s resulting attention weight thereby indicates its relative importance in the context of feature fusion.

The culmination of this process is the application of the spatial attention weight map, \( S \), to the initial fused feature maps through element-wise multiplication, leading to the generation of the final fused feature map \( F_{\text{final}} \):

\[
F_{\text{final}} = S \odot (F_{L0} \oplus F_{L2})
\]

The operation \( \odot \) represents the Hadamard product, or element-wise multiplication, ensuring that the contribution of each pixel is adjusted according to its derived importance, as indicated by the spatial attention weight map.

Through the implementation of this spatial attention module, PortLaneNet achieves a dynamic and adaptive feature fusion process. This not only elevates the model’s performance by enabling a more discerning perception of lane lines in varied and complex scenarios but also significantly improves the model’s interpretability. Researchers and practitioners can thus gain deeper insights into the model’s decision-making process, particularly how it discerns and prioritizes different features during lane detection tasks.

3.5. Scene Prior Perception Module

The Scene Prior Perception Module stands as a pivotal component within our PortLaneNet framework, offering crucial high-level contextual semantic insights crucial for refining lane detection. This module is adept at discerning broader scene contexts within the input image, identifying specific scenarios such as “container areas”, “main roads”, among others. Recognizing these distinct scene categories is essential for lane detection, given that different environments possess unique lane layouts and characteristics. This contextual comprehension is invaluable, aiding the model in distinguishing actual lane lines from other ground markings, especially since lane patterns may vary across different scenes.
A significant innovation in our model is the fusion of scene-specific features derived from the Scene Prior Perception Module with attention-driven features from the primary lane detection pathway. These feature sets operate within distinct representational spaces: scene-specific features provide a global scene understanding, acting as a form of prior knowledge, while attention-driven features capture more localized, detailed aspects crucial for lane detection. To harness both types of information effectively, we adopt a fusion strategy akin to multi-modal fusion by concatenating these features. This fusion ensures that the lane detection model benefits from a comprehensive scene context alongside the critical granular details essential for precise lane delineation.

The mathematical foundation for explaining how the Scene Prior Perception Module enhances lane detection accuracy through additional information can be elucidated as follows [29,30]:

Initially, let \( x \sim N(m_x, P_{22}) \) represent a set of features (additional information) from a convolution block, where \( N(m_x, P_{22}) \) denotes a normal distribution with mean \( m_x \) and covariance \( P_{22} \). Let \( y \sim N(m_y, P_{11}) \) be the output set, with \( e \sim N(0, \sigma^2 I) \) representing Gaussian white noise, where \( I \) is the identity matrix. Here, \( m_y = E[y] \) is the mean of \( y \), with \( E \) denoting the expectation.

In the absence of additional information, the expectation of the output is \( \hat{x}_k \), assuming \( P_{22} > 0 \) (positive definite) and that \( x \) can be expressed through a complex nonlinear function \( f \), i.e., \( x = f(y) + e \).

Given \( x, y, e \) each follow a Gaussian distribution, their joint distribution is given by:

\[
\begin{bmatrix} y \\ x \end{bmatrix} \sim N \left( \begin{bmatrix} m_y \\ m_x \end{bmatrix}, \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \right) \tag{5}
\]

Considering a new random variable \( y \mid x \) as a new measurement given additional information \( x \), based on Bayesian principles, marginal distribution, and properties of normal distributions, it too follows a Gaussian distribution:

\[
P(y \mid x) = \frac{p(y, x)}{p(x)} \sim N \left( m_{y \mid x}, P \right) \tag{6}
\]

where \( m_{y \mid x} = E[y \mid x] = \hat{x}_{k+1} \) defines the new expectation of the output given additional information.

By least-squares estimation, this simplifies to:

\[
\hat{x}_{k+1} = \hat{x}_k + P_{12}P_{22}^{-1}(x - m_x) \tag{7}
\]

The error is defined as: \( y' = (y \mid x) - \hat{x}_{k+1} \). The expectation of this error is:

\[
E[y'] = E[y \mid x] - E[\hat{x}_{k+1}] = \hat{x}_{k+1} - \hat{x}_{k+1} = 0 \tag{8}
\]

indicating that the estimation is unbiased.

Moreover, the discrepancy between the new measurement given additional information and the original output expectation is:

\[
E[y \mid x] - E[\hat{x}_k] = \hat{x}_{k+1} - \hat{x}_k = P_{12}P_{22}^{-1}(x - m_x) \neq 0 \tag{9}
\]

demonstrating that, with additional information, the original output expectation is biased. This framework can also be interpreted within a convolutional neural network, where one convolution block acts as the additional information, and the classification block acts as the new measurement given this information. In this manner, the integration of the Scene Prior Perception Module provides a deeper understanding of complex scene contexts for the lane detection model, improving detection accuracy by incorporating additional information.

Specifically, the Scene Prior Perception Module’s original output, \( C \), is mapped into a richer representational space using a fully connected layer, denoted as FC(C). This enriched
scene representation encapsulates a deeper understanding of the scene context. This enhanced representation is then concatenated with the primary output of the lane detection model, producing two augmented outputs: $L_{cls}$ for the classification task and $L_{reg}$ for the regression task. Incorporating this module endows PortLaneNet with a comprehensive understanding of the scene, bolstering its capability to differentiate and accurately detect lanes in complex and varied scenarios typical of port environments.

3.6. Loss Function

Each anchor’s classification output, $L_{cls}$, is a $(k + 1)$-dimensional vector, where $k$ is the number of lane types, and the additional dimension represents “background” or invalid proposals. Each anchor’s regression output $L_{reg}$ is an $(N_{pts} + 1)$-dimensional vector, where $N_{pts}$ is the number of horizontal offsets, and the additional dimension represents the proposed length.

Our model employs a multi-task loss function that optimizes both the classification and regression tasks simultaneously. Specifically, Focal Loss [31] is used to calculate the classification loss $L_{cls}$ for each anchor, and Smooth L1 Loss is used to calculate the regression loss $L_{reg}$ for each anchor, defined as follows:

$$L = \lambda \sum L_{cls}(p^i, p^*_i) + \sum L_{reg}(r_i, r^*_i)$$  (10)

where $L_{cls}(p^i, p^*_i)$ is the Focal Loss for the $i$-th anchor, and $L_{reg}(r_i, r^*_i)$ is the Smooth L1 Loss for the $i$-th anchor. $p^i$ and $r^*_i$ are the true classification label and regression target for the $i$-th anchor, respectively. $\lambda$ is a weight factor used to balance the classification loss and regression loss. Focal Loss is primarily used to solve the imbalance problem between positive and negative samples. In our model, it is applied to calculate the classification $L_{cls}$. Focal Loss is designed to reduce the contribution of easily classified samples (mainly negative samples) to the total loss, enabling the model to focus more on those hard-to-classify samples (mainly positive samples) during the training process. This is crucial for lane line detection tasks because, in most cases, lane line samples (positive samples) are much fewer compared to background samples (negative samples). Smooth L1 Loss, on the other hand, is used to calculate the regression loss $L_{reg}$ for positive samples. This loss function measures the discrepancy between the lane line parameters (including horizontal offset and length) predicted by the model and the actual values. Smooth L1 Loss has a large penalty for large errors but a small penalty for small errors. This allows the model to predict the parameters of the lane lines more accurately while ensuring stability.

4. Experiments

In this section, we first introduce the production process of our proprietary dataset and its efficient evaluation metrics. Following this, we provide detailed information about the experimental setup and specific details of the experiment. Finally, we present the experimental results of our proposed method and other similar methods on this dataset.

4.1. Dataset and Evaluation Metrics

To collect the PortLane dataset, we installed a camera with a resolution of $1080 \times 1920$ on a tractor truck within the operation area and conducted continuous image collection for a week. The dataset not only covers various areas of daily operations, including main roads, crane areas, container areas, and parking lots, but also includes typical images under various climatic types such as sunny, rainy, and cloudy days. As shown in Table 1, we selected 1504 images for data annotation, of which 1204 images were used as the training set and 300 images as the test set. We referred to the mainstream dataset Tusimple [32] to establish the annotation format and evaluation metrics of the dataset. In addition to annotating lane lines, we also annotated scene labels for each image to pre-train the Scene Prior Perception Module. Figure 3 shows in detail the compositional proportion of each type of scene in the dataset.
Table 1. Lane detection dataset description.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Test</th>
<th>Visualize</th>
<th>Resolution</th>
<th>Lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>PortLane</td>
<td>1204</td>
<td>300</td>
<td>2849</td>
<td>1080 × 1920</td>
<td>≤4</td>
</tr>
</tbody>
</table>

![Figure 3. Examples and distribution of the dataset classified by scene.](image)

We adopted accuracy as the main evaluation metric for PortLane datasets, which is defined as follows:

\[
\text{accuracy} = \frac{\sum_{\text{clip}} C_{\text{clip}}}{\sum_{\text{clip}} S_{\text{clip}}}
\]  

(11)

In this formula, \( C_{\text{clip}} \) represents the count of lane points accurately predicted within the clip, while \( S_{\text{clip}} \) signifies the total points count within the clip. A point prediction is deemed accurate if it falls within a 20-pixel range of the actual ground truth. A lane prediction is classified as a true positive, significant for FDR and FNR metrics, only when its correctly predicted points exceed 85%. Additionally, we present the rate of false positive (FP) and false negative (FN), where \( FP = \frac{F_{\text{pred}}}{N_{\text{pred}}} \) and \( FN = \frac{M_{\text{pred}}}{N_{\text{gt}}} \).

To further enhance the comprehensiveness of our evaluation, we introduced the \( F1 \) score as an additional metric. The \( F1 \) score is calculated as follows:

\[
F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]

(12)

where Precision is the ratio of correctly predicted positive observations to the total predicted positives, and Recall is the ratio of correctly predicted positive observations to all observations in the actual class. The \( F1 \) score serves as a harmonic mean of Precision and Recall, providing a balance between the Precision and the Recall of our model. By incorporating the \( F1 \) score, we aim to offer a more balanced and nuanced understanding of the model’s performance, especially in scenarios where the class distribution is imbalanced.

4.2. Experiments Settings

The experiments in this study were conducted on a high-performance laptop equipped with an AMD Ryzen 7 5800H processor and NVIDIA GeForce RTX 3060 graphics card, running the Windows 11 operating system. This hardware configuration not only ensures efficient model training and testing but also meets the computational demands of deep learning models.

In terms of model selection, we employed the pre-trained ResNet18 as our backbone network. The choice of ResNet18 over deeper networks like ResNet34 was a deliberate balance between accuracy and computational cost. While deeper networks could potentially offer slightly higher accuracy, they significantly increase computational and memory requirements. This is particularly critical in resource-constrained onboard systems, such as the automated driving systems of container trucks. ResNet18 maintains a high level of accuracy while offering lower computational complexity and memory demand, which is key for applications operating in resource-limited environments.
All input images were uniformly resized to 320 × 800 pixels, balancing the capture of image details with computational efficiency. To optimize the learning process, the Adam optimizer was utilized with an initial learning rate set at $2 \times 10^{-3}$, and the model was trained over 150 epochs. Moreover, we applied a cosine annealing strategy for learning rate adjustment, aiding in finer model tuning during the later stages of training.

Regarding data processing, we applied a series of data augmentation techniques to the PortLane dataset, including random affine transformations and horizontal flipping. These techniques enhance the diversity of the data and improve the model’s adaptability to varying environmental conditions. The Intersection over Union (IoU) threshold for Non-Maximum Suppression (NMS) was set at 0.5, effectively removing duplicate detections while retaining critical lane line information.

Collectively, these settings were designed to ensure the model’s efficiency and accuracy in lane detection tasks, while also considering the repeatability of the experiment and practical applicability in real-world applications.

4.3. Algorithm Process

The specific training process of the PortLaneNet model is illustrated in Figure 4:

Figure 4. Training process diagram of the PortLaneNet model.
(1) Data Collection and Preprocessing

The core framework of the PortLaneNet model is initialized based on the pre-trained ResNet-18, which is the backbone network part shown in the diagram. Data collection involves acquiring images of various scenes within container port environments. Data preprocessing includes data cleaning, annotation, and other necessary preprocessing steps to ensure the quality and consistency of the input data.

(2) Feature Extraction and Scene Understanding

Starting from the initialized backbone network, the model forward-propagates layer by layer, enhancing the understanding of complex scene semantics through the Scene Prior Perception Module and attention mechanisms. These modules work together to improve the model’s accuracy in detecting lane line features.

(3) Optimization and Evaluation

The network weights are updated using the Adam optimizer, dynamically adjusting the learning rate based on the gradient of the loss function and the estimates of the first and second moments. Figure 5 illustrates the change in the total loss function during the training process of the PortLaneNet model, compared with the baseline method LaneATT. As depicted, the loss curve of PortLaneNet drops rapidly in the initial training phase, indicating that the model is quick to learn useful features and adapt to the characteristics of lane lines in container terminal environments. In contrast, while LaneATT also shows a swift decline in loss early on, its rate of decrease slows down in later training, possibly due to its model structure and pre-trained features not being as targeted as those of PortLaneNet. Notably, in the mid-to-late stages of training, the loss curve of PortLaneNet presents a more stable downward trend compared to LaneATT, suggesting that PortLaneNet possesses superior fitting abilities and generalization performance when faced with complex samples from the training set. This advantage is attributable to its integrated design that incorporates the Scene Prior Perception Module. This design ensures that PortLaneNet can more accurately capture and utilize key information affecting lane detection, enabling more precise lane line detection in the complex and variable environment of container terminals.

![Loss Curve](image_url)

**Figure 5.** Loss curve comparison between the PortLaneNet model and the LaneATT baseline during training.
(4) Model Performance Tuning

Based on the model’s performance on the validation set, adjustments were made to model performance, including learning rate adjustments, changes to data augmentation strategies during training, and fine-tuning of network parameters. Through continuous iteration, the model gradually achieved optimal performance.

4.4. Result

In this study, we selected LaneATT as our baseline method and compared our proposed approach with other methods based on both ResNet-18 and ResNet-34 backbones. As demonstrated in Table 2, our model shows a significant improvement in accuracy primarily due to the integration of the Scene Prior Perception Module, which considerably enhances the recognition and interpretation of high-level semantics in complex ground markings. However, the incorporation of this module results in a decreased detection speed compared to the baseline methods. With ResNet-18 as the backbone, the model’s speed decreased from 121 frames per second (FPS) to 103 FPS. In the operational context of port areas, this detection speed level is more than capable of meeting the operational requirements. Given that vehicles in port environments do not demand extremely high processing speeds, prioritizing accuracy is crucial for ensuring safety and efficiency in such scenarios. Our model was compared with four recent mainstream efficient algorithms, and the results, achieved under the same experimental conditions, show that the model integrating the Scene Prior Perception Module outperforms LaneATT by approximately 2.6% in accuracy. Therefore, this method can better understand the semantic information of complex scenes and improve recognition accuracy.

Table 2. Comparison with mainstream algorithms on our dataset.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy (%)</th>
<th>FP (%)</th>
<th>FN (%)</th>
<th>Frame Rate (/s)</th>
<th>F1 Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LaneATT-Resnet 18</td>
<td>77.6</td>
<td>5.0</td>
<td>23.4</td>
<td>177</td>
<td>81.6</td>
</tr>
<tr>
<td>LaneATT-Resnet 34</td>
<td>78.4</td>
<td>4.9</td>
<td>21.6</td>
<td>98</td>
<td>82.1</td>
</tr>
<tr>
<td>DILane-Resnet 18</td>
<td>78.5</td>
<td>6.2</td>
<td>22.5</td>
<td>121</td>
<td>82.4</td>
</tr>
<tr>
<td>DILane-Resnet 34</td>
<td>80.1</td>
<td>6.5</td>
<td>22.3</td>
<td>58</td>
<td>82.7</td>
</tr>
<tr>
<td>GANet-Resnet 18</td>
<td>79.5</td>
<td>5.6</td>
<td>24.9</td>
<td>127</td>
<td>82.5</td>
</tr>
<tr>
<td>GANet-Resnet 34</td>
<td>80.4</td>
<td>5.3</td>
<td>22.4</td>
<td>54</td>
<td>82.9</td>
</tr>
<tr>
<td>CLRNet-Resnet 18</td>
<td>79.2</td>
<td>7.5</td>
<td>25.9</td>
<td>141</td>
<td>82.4</td>
</tr>
<tr>
<td>CLRNet-Resnet 34</td>
<td>81.0</td>
<td>6.8</td>
<td>24.7</td>
<td>72</td>
<td>83.7</td>
</tr>
<tr>
<td>PortLaneNet-Resnet 18</td>
<td>80.2</td>
<td>4.1</td>
<td>21.1</td>
<td>103</td>
<td>83.1</td>
</tr>
<tr>
<td>PortLaneNet-Resnet 34</td>
<td>81.5</td>
<td>4.3</td>
<td>21.7</td>
<td>50</td>
<td>83.5</td>
</tr>
</tbody>
</table>

All measures were computed using the official source code.

The results showcased in Figure 6 emphatically substantiate our model’s adeptness at navigating complex environments, as reflected by the precision and F1 scores in Table 2. Our model remains unfazed in the face of scenarios that typically challenge visual perception systems: intense sunlight glare (Dazzle), reflective wet surfaces (Wet Lane), partially obscured views (Occluded Lane), and even where the lane integrity is compromised due to breakage (Breakage Lane). Despite these adverse conditions, PortLaneNet accurately discerns and traces the correct lane lines. Moreover, the model excels in contexts with unconventional lane indicators, such as mixed ground markings (Mixed Markings), open spaces with no clear lane demarcations (No Lane), and roads with directional arrows (Arrow). It differentiates between actual lane lines and other similar-looking patterns, maintaining a high detection accuracy. Collectively, these examples in Figure 6 highlight PortLaneNet’s robust feature extraction and contextual comprehension, confirming its potential for real-world deployment where diverse and unpredictable elements are commonplace.
work may consider incorporating features with greater illumination robustness. In addition, the model's performance still declines under extreme lighting conditions. Future research can build on the foundation of this paper, continuing to explore how to establish a model's deep semantic understanding ability of complex scenes to adapt to the needs of different application scenarios. It is believed that with the continuous enhancement of the model's robustness and reliability, more advanced feature extraction techniques and network structures can be introduced to improve the adaptability of the model.

This paper also presents areas that warrant further improvements. First, the accuracy of the Scene Prior Perception Module directly affects the final lane detection results, thus further optimizing the performance of the Scene Prior Perception Module is crucial. Secondly, the model’s performance still declines under extreme lighting conditions. Future work may consider incorporating features with greater illumination robustness. In addition, for heavily occluded lane lines, more advanced feature extraction techniques and network structures can be introduced to improve the adaptability of the model.

Overall, this paper provides an effective methodological reference for the application of lane detection tasks in enclosed terminal environments. Future research can build on the foundation of this paper, continuing to explore how to establish a model’s deep semantic understanding of complex and dynamic scenes, greatly improving its robustness and reliability. This innovation, coupled with the integration of an attention mechanism, enables PortLaneNet to adeptly navigate and interpret the unique ground markings and environmental conditions characteristic of port settings, setting it apart from existing methodologies. Experiments conducted on our proprietary PortLane dataset demonstrate that our method shows significant advantages compared to other lane detection algorithms, especially in dealing with ground markings unique to port environments. This research provides valuable insights into applying lane detection tasks in enclosed port settings. The Scene Prior Perception Module offers an important priori contextual information for lane detection, while the attention module enables aggregation of global contexts, both enhancing the model’s capability in deciphering complex scenes. The model exhibits strong robustness when encountering obscured, worn-out lane lines, more advanced feature extraction techniques and network structures can be introduced to improve the adaptability of the model.

5. Conclusions and Discussion

In this paper, we have proposed a real-time lane detection model, PortLaneNet, tailored for enclosed port environments. Our main contribution includes the introduction of a Scene Prior Perception Module, which incorporates scene classification tasks with the lane detection model through feature fusion. This approach significantly enhances the model’s capacity for semantic understanding of complex and dynamic scenes, greatly improving its robustness and reliability. This innovation, coupled with the integration of an attention mechanism, enables PortLaneNet to adeptly navigate and interpret the unique ground markings and environmental conditions characteristic of port settings, setting it apart from existing methodologies. Experiments conducted on our proprietary PortLane dataset demonstrate that our method shows significant advantages compared to other lane detection algorithms, especially in dealing with ground markings unique to port environments. This research provides valuable insights into applying lane detection tasks in enclosed port settings. The Scene Prior Perception Module offers an important priori contextual information for lane detection, while the attention module enables aggregation of global contexts, both enhancing the model’s capability in deciphering complex scenes. The model exhibits strong robustness when encountering obscured, worn-out lane lines or ground markings that bear semblance to lane lines.
semantic understanding ability of complex scenes to adapt to the needs of different application scenarios. It is believed that with the continuous enhancement of the model’s expressive capabilities, end-to-end lane detection systems will be widely applied, better serving fields such as autonomous driving, and promoting the development of the transportation industry.

Author Contributions: Methodology, H.Y., Z.K. and X.Z.; Software, Z.K.; Resources, C.Z.; Writing—original draft, Z.K.; Writing—review & editing, H.Y., Y.Z. and X.Z.; Visualization, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: Chenhe Zhang is an employee of Shanghai East Container Terminal Co., Ltd. The paper reflects the views of the scientists, and not the company.

References


30. Zhang, X.; Zhao, N.; Lv, Q.; Ma, Z.; Qin, Q.; Gan, W.; Bai, J.; Gan, L. Garbage Classification Based on a Cascade Neural Network. Neural Netw. World 2023, 2, 101–112. [CrossRef]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.