PVALane: Prior-Guided 3D Lane Detection with View-Agnostic Feature Alignment

Zewen Zheng[1,2*], Xuemin Zhang[1], Yongqiang Mou[1†], Xiang Gao[1,2], Chengxin Li[1,3], Guoheng Huang[2], Chi-Man Pun[4], Xiaochen Yuan[5]

[1]X Lab, GAC R&D CENTER, Guangdong, China
[2]Guangdong University of Technology, Guangdong, China
[3]South China Normal University, Guangdong, China
[4]University of Macau, Macau, China
[5]Macao Polytechnic University, Macao, China
yongqiang.mou@gmail.com

Abstract

Monocular 3D lane detection is essential for a reliable autonomous driving system and has recently been rapidly developing. Existing popular methods mainly employ a predefined 3D anchor for lane detection based on front-viewed (FV) space, aiming to mitigate the effects of view transformations. However, the perspective geometric distortion between FV and 3D space in this FV-based approach introduces extremely dense anchor designs, which ultimately leads to confusing lane representations. In this paper, we introduce a novel prior-guided perspective on lane detection and propose an end-to-end framework named PVALane, which utilizes 2D prior knowledge to achieve precise and efficient 3D lane detection. Since 2D lane predictions can provide strong priors for lane existence, PVALane exploits FV features to generate sparse prior anchors with potential lanes in 2D space. These dynamic prior anchors help PVALane to achieve distinct lane representations and effectively improve the precision of PVALane due to the reduced lane search space. Additionally, by leveraging these prior anchors and representing lanes in both FV and bird-eye-viewed (BEV) spaces, we effectively align and merge semantic and geometric information from FV and BEV features. Extensive experiments conducted on the OpenLane and ONCE-3DLanes datasets demonstrate the superior performance of our method compared to existing state-of-the-art approaches and exhibit excellent robustness.

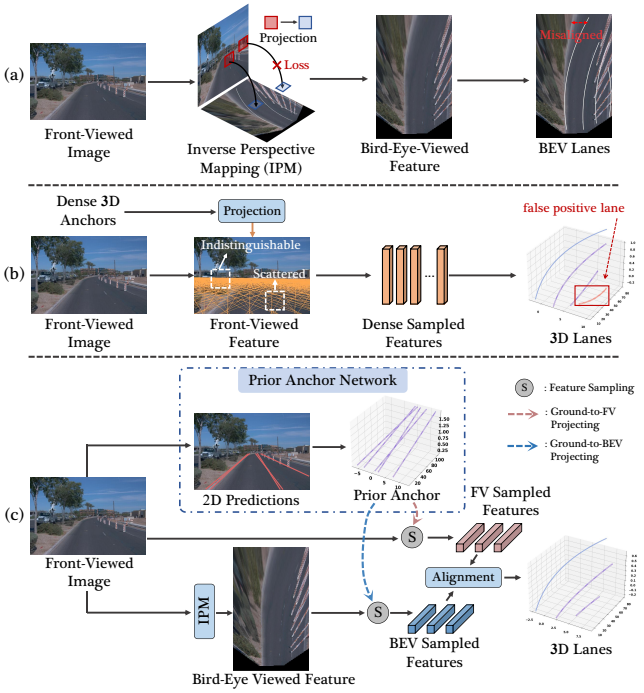


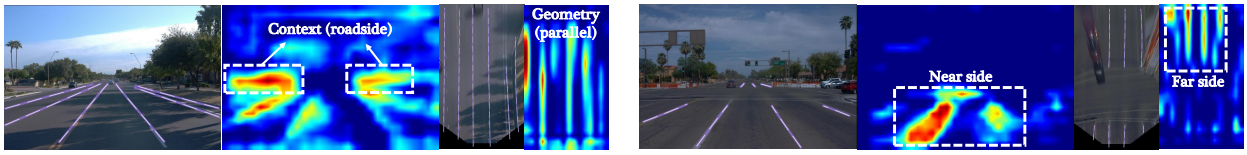
Figure 1: Illustration of the difference among (a) the BEV-based methods, (b) the FV-based methods, and (c) our PVALane.

Introduction

As a fundamental module in autonomous driving systems, robust lane detection has received a lot of research attention and made unprecedented progress. However, the front-viewed (FV) lane detection models that can only provide prediction results in the 2D image space are not directly applicable to complex real-world scenarios (Neven et al. 2018; Pan et al. 2018; Liu et al. 2021b). As a promising direction, 3D lane detection is proposed to tackle the above problem. It aims to construct a model that extracts lane features from a monocular 2D image and then detects the lanes in the ground coordinate system.

Fueled by the success of 3D object detection (Liang et al. 2018; Li et al. 2022), current 3D lane detection models (Garnett et al. 2019; Guo et al. 2020; Chen et al. 2022) often detect 3D lanes by transforming FV features to bird-eye-viewed (BEV) space using inverse perspective mapping (IPM). Due to the similar appearance and geometry of different lanes on the top-view plane, the BEV-based method exhibits geometry translational invariance. Although the representation in BEV space provides better geometric infor-

*Work done during the internship at GAC R&D CENTER.
†Corresponding author.



(a) **FV:** ***Context,*** **BEV:** ***Geometric*** (b) FV: ***Near side,*** **BEV:** ***Far side***

Figure 2: Visualization of activation maps for two image samples under FV and BEV. We can observe that BEV tends to perceive geometric properties (e.g., parallel) and lanes on the far side, while FV focuses more on capturing near and contextual information (e.g., roadside) for lane detection. (Best viewed in color).

mation, the dependence on IPM leads to several unexpected problems. As illustrated in Figure 1(a), the flat ground assumption of IPM makes the BEV and 3D ground truth (GT) spaces misaligned in uphill or downhill cases, and thus the method may not be generalizable to rough ground scenes with varying visual appearances. Secondly, IPM inevitably results in the loss of the original semantic and contextual information within the FV features.

To address these challenges, recent approaches (Yan et al. 2022; Huang et al. 2023) have shifted their focus towards 3D lane prediction using semantic information directly from FV. Specifically, this method acquires lane representations by projecting predefined 3D anchors onto corresponding locations in the FV image for sampling and then performs regression based on these anchors. While this approach eliminates the effects of view transformations, it introduces a perspective geometric distortion due to the ground-to-image projection. As shown in Figure 1(b), the anchor projections on the FV image are indistinguishable in the distance but significantly scattered in the nearby areas, thus necessitating a dense anchor design to mitigate these geometric variances. However, the introduction of these dense anchors might result in feature overlap and false detections in dynamic scenarios, which, in turn, limits its effectiveness for geometric-oriented tasks like 3D lane detection.

In this paper, we introduce PVALane, a prior-guided 3D lane detection model that utilizes 2D prior knowledge to accurately estimate the 3D lane locations, as illustrated in Figure 1(c). Instead of utilizing empirical dense anchors for lane detection directly (Huang et al. 2023), we establish Prior Anchors, which are obtained almost cost-free from the 2D prediction and applied for downstream 3D lane detection. Specifically, at the top of the backbone features, we construct a Prior Anchor Network (PAN) which projects predefined 3D anchors into 2D space to calculate their objectness probabilities, thereby filtering out anchors with no potential lanes and using them as prior anchors. This prior anchor explicitly provides a strong prior indicating lane localization and ensures that only high-quality prior anchors are used for 3D lane detection. Therefore, it can effectively improve the precision of PVALane due to the reduced lane search space. Furthermore, the PAN requires only a few additional fully-connected layers and proceeds directly on the FV feature, so it can be integrated as an easy-to-deploy module and trained in an end-to-end manner.

Based on the intuitive insight that various view features tend to utilize different view properties (e.g., contextual and geometric) and region information for lane detection (shown in Figure 2), we further propose a Prior-Guided View-agnostic Feature Alignment Module (PVFA). Specifically, PVFA projects the prior anchors into the FV and BEV spaces, acquires their corresponding features through sampling, and subsequently fuses them together. Since the prior anchors are defined in 3D space and are extremely sparse, they can effectively narrow the association between the FV and BEV sampling features. In addition, this shared sampled feature space decouples the downstream lane detection from the view space, making PVALane inherently view-agnostic and extendable to multi-view/cross-view scenarios.

The contributions of this work are summarized as follows:

- We introduce a prior perspective on lane detection and propose an end-to-end PVALane framework, which utilizes 2D prior knowledge to achieve precise and efficient 3D lane detection.
- We propose a novel prior anchor, which is obtained almost cost-free from accurate 2D predictions and explicitly provides a strong prior indicating lane localization.
- We develop a view-agnostic feature alignment method that leverages the prior anchor to effectively align and merge both geometric and semantic information across different views.
- Experiments show that PVALane achieves new state-of-the-art performance on two popular 3D lane detection benchmarks and exhibits excellent robustness.

Related Work

Monocular 3D Lane Detection

Monocular 3D lane detection (Garnett et al. 2019; Efrat et al. 2020; Chen et al. 2022; Yan et al. 2022; Luo et al. 2023) is a challenging task that has attracted the interest of the computer vision community in recent years. 3D-LaneNet (Garnett et al. 2019) is the pioneering work in this domain, which transforms the front-viewed (FV) features to bird-eye-viewed (BEV) for lane detection. Persformer (Chen et al. 2022) proposes a spatial feature transformation based on deformable attention (Zhu et al. 2021) for robust BEV features. BEV-LaneDet (Wang et al. 2023) proposes a virtual camera that guarantees the consistency of spatial relations among cameras. While representing lanes in BEV space offers better geometric properties, the flat ground assumption of IPM introduces several challenges. Anchor3DLane (Huang et al. 2023) directly detects 3D

lanes from FV images without IPM projection. However, this BEV-free approach requires a dense anchor design to mitigate the perspective geometric distortion, which leads to confusing lane representations. Therefore, our PVALane utilizes sparse prior anchors obtained from FV images to achieve distinct lane representations and thus eliminate the impact of redundant anchors. Furthermore, PVALane leverages both semantic and structural information from FV and BEV to achieve accurate lane detection.

Prior-guided Lane Detection

Prior knowledge is ubiquitous in lane detection as most methods (Neven et al. 2018; Pan et al. 2018; Liu et al. 2021a) aims to enhance model performance by leveraging existing information. However, those prior-guided approaches often lead to an inevitable increase in model complexity due to the introduction of additional modules or supervised loss. CondLaneNet (Liu et al. 2021a) uses the pre-extracted instance origin to guide the underlying visual features to describe the shape prediction of lane instances accurately. Gen-LaneNet (Guo et al. 2020) used a segmentation subnetwork to generate lane segmentation from 2D images and to support the regression of 3D lanes subsequently. CLGo (Liu et al. 2022) leverages pre-estimated camera pitches and heights to transform raw images into BEV images, enabling precise 3D lane detection. Different from the above methods, our approach introduces a nearly cost-free prior anchor to reduce the lane search space, thereby significantly reducing the complexity of the downstream lane detection model.

Methodology

Lane Representation

Similar to (Garnett et al. 2019; Chen et al. 2022), we define 3D Lanes as a series of 3D points with N_p pre-defined fixed y-coordinates. Specifically, given a 3D lane set $\mathbf{L}_{3D} = \left\{l^i\right\}_{i=1}^{N_l}$ containing N_l lanes, we formulate the i-th lane as:

$$l^i = \left\{\left(x^{(i,k)}, y^k, z^{(i,k)}, \mathrm{vis}^{(i,k)}\right)\right\}_{k=1}^{N_p}, \tag{1}$$

where $x^{(i,k)}, y^k, z^{(i,k)}$ and binary $\mathrm{vis}^{(i,k)}$ denote the ground coordinates and visibility of k-th point of the current lane.

Architecture

The overall architecture of our PVALane is illustrated in Figure 3. Given a 2D front-viewed (FV) image as input, our model first extracts FV features with a ResNet (He et al. 2016) backbone. These features are then passed through the prior anchor network, which generates lane probabilities indicating whether the corresponding 3D anchor contains the lane target. To filter redundant 3D anchors, we further select the anchors that score above a predefined threshold as the prior anchor. Subsequently, the FV and bird's-eye-view (BEV) features are encoded by two specially designed view encoders and aligned in a shared sampled feature space by projecting the prior anchor to the corresponding view. Finally, the aligned features are passed into the prediction head to obtain the 3D lane predictions.

Prior Anchor Network

Although 2D features are not directly suitable for geometric-oriented tasks due to perspective geometric distortion, we consider using them to enhance 3D lane detection with a prior-guided perspective. On one hand, this initial FV feature contains rich semantic and contextual information. On the other hand, it does not require a view transformation, enabling the quick generation of prior information. As demonstrated in the majority of two-stage object detection literature (Ren et al. 2015), a lightweight end-to-end prior network can quickly generate region prior, thereby significantly reducing the complexity of the object detection task. A visual explanation of the Prior Anchor Network is shown in Figure 4.

3D Lane Anchors Inspired by (Huang et al. 2023), we define lane anchors as 3D anchors in the ground coordinate system to better adapt to 3D lane shapes. Specifically, given a set of fixed y-positions $\mathbf{y} = \left\{y^k\right\}_{k=1}^{N_p}$, the j-th 3D anchor $\mathbf{A}^j = \left\{\mathbf{q}^{(j,k)}\right\}_{k=1}^{N_p}$ defines a 3D lane line using two vectors $\left(\mathbf{x}^j, \mathbf{z}^j\right)$, where $\mathbf{x}^j, \mathbf{z}^j \in \mathbb{R}^{N_p}$ are horizontal and vertical offsets relative to the positions of the N_p predefined points.

Anchor Projecting To obtain the lane probability of the 3D anchors, we first project them into the 2D plane of FV feature $\mathbf{F}_{fv} \in \mathbb{R}^{H_{fv} \times W_{fv} \times C}$ as 2D anchors to extract their corresponding features. Specifically, given the j-th 3D anchor $\mathbf{A}^j = \left(\mathbf{x}^j, \mathbf{y}, \mathbf{z}^j\right)$ as an example, we define the projection as:

$$\tilde{\mathbf{z}}^j \begin{bmatrix} \mathbf{u}^j \\ \mathbf{v}^j \\ 1 \end{bmatrix} = \mathbf{K} \cdot \mathbf{T}_{g \to c} \begin{bmatrix} \mathbf{x}^j \\ \mathbf{y} \\ \mathbf{z}^j \\ 1 \end{bmatrix}, \tag{2}$$

where $\mathbf{K} \in \mathbb{R}^{3\times3}$ is the intrinsic matrix, $\mathbf{T}_{g\to c} \in \mathbb{R}^{3\times4}$ denotes the transformation matrix from ground coordinates to camera coordinates, $\tilde{\mathbf{z}}^j$ denotes depth to the camera plane. Given the above projection denoted as $\mathcal{P}_{g2fv}(\cdot)$, we obtain the anchor feature $\mathbf{F}_a^j$ by sampling the FV feature $\mathbf{F}_{fv}$:

$$\mathbf{F}_a^j = \mathbf{F}_{fv}(\mathcal{P}_{g2fv}(\mathbf{A}^j)) \in \mathbb{R}^{N_p \times C}. \tag{3}$$

Prior Anchor Generation Since the anchor feature $\mathbf{F}_a^j$ contains the semantic feature of the corresponding 3D anchor, we further apply a primary classification head to $\mathbf{F}_a^j$ to obtain lane classification scores $\mathbf{p}_{pri}^j \in \mathbb{R}^{1+N_c}$, where N_c represents the number of lane categories. Then, the $\tilde{\mathbf{p}}_{pri}^j$ is calculated according to the classification score to supply potential lane probability:

$$\tilde{\mathbf{p}}_{pri}^j = 1 - \mathbb{1}\left(c_n = 0\right) \mathcal{S}(\mathbf{p}_{pri}^j) \in \mathbb{R}^1, \tag{4}$$

where $c_n \in \{0, 1, \ldots, N_c\}$, 0 represents non-lane category, $\mathbb{1}\left(\cdot\right)$ denotes an indicator function and $\mathcal{S}(\cdot)$ is the softmax function. Each value in the $\left\{\tilde{\mathbf{p}}_{pri}^j\right\}_{j=1}^{N_a}$ indicates the probability that the corresponding anchor contains lane. Therefore, to select high-quality 3D anchors as prior anchors, we simply filter out the low-probability anchors based on a threshold τ:

$$\mathbf{A}_{pri} = \left\{\mathbf{A}^k\right\}_{k \in \{\Psi_j(\tilde{\mathbf{p}}_{pri}^j, \tau)\}}, \tag{5}$$

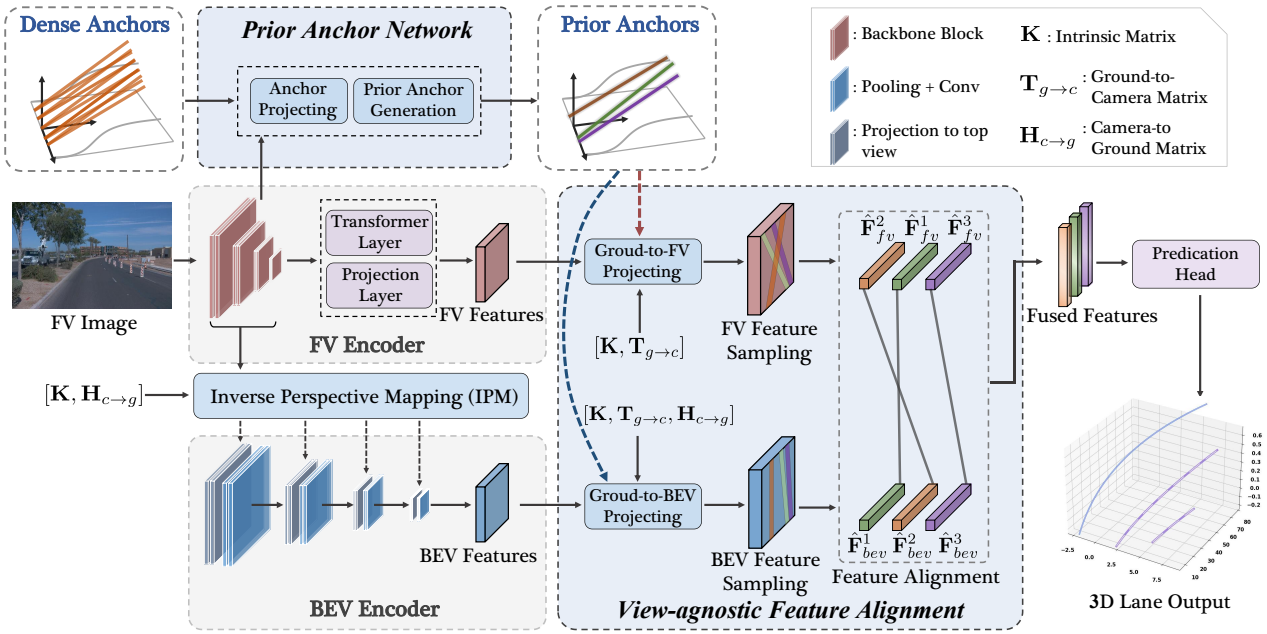


Figure 3: The overall architecture of PVALane. The prior anchor network narrows the lane search space by generating high-quality and sparse prior anchors. Afterward, a prior-guided view-agnostic feature alignment module is applied to align and merge geometric and semantic information from different view features.

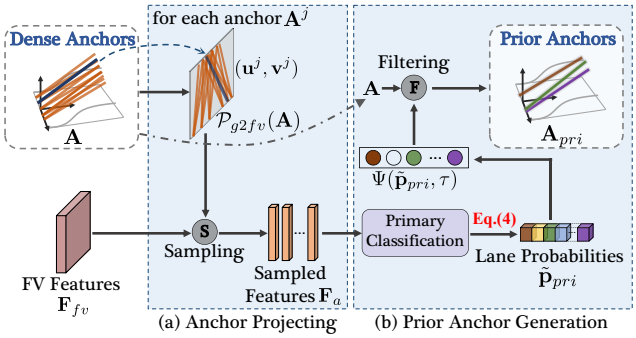


Figure 4: Illustration of Prior Anchor Network.

where $\Psi_j(\cdot)$ denotes an operator that returns the value of j satisfying $\tilde{\mathbf{p}}^j_{pri} > \tau$. By incorporating the prior provided by this efficient 2D prediction head, the complexity of 3D lane detection can be significantly reduced. This enables the model to prioritize the more challenging task of regression.

Loss Function To eliminate the effect of positive and negative sample imbalance due to perspective geometric distortion, we adopt focal loss (Lin et al. 2017b) for training classification:

$$\mathcal{L}_{pri} = -\sum_{j=1}^{N_a} \alpha \left(1 - \mathbf{p}^j_t\right)^{\gamma} \log \mathbf{p}^j_t, \tag{6}$$

where $\mathbf{p}^j_t$ is the predicted probability of the current category, α and γ are the hyperparameters for focal loss, which are set to 0.25 and 2 in our experiments, respectively.

View-specific Feature Encoding

To leverage semantic and geometric information present in FV and BEV, PVALane simultaneously learns features from both views. Given the distinct view representation of FV and BEV features, the model incorporates two specialized encoders to capture the specific information from each view independently.

FV Context-aware Encoder For each FV features $\mathbf{F}_{fv} \in \mathbb{R}^{H_{fv} \times W_{fv} \times C}$, we introduce a transformer encoder (Vaswani et al. 2017) with a projection layer to capture global semantic and contextual information:

$$\tilde{\mathbf{F}}_{fv} = \mathcal{P}\left(\mathcal{E}\left(\mathbf{F}_{fv}\right)\right) \in \mathbb{R}^{H_{fv} \times W_{fv} \times C}, \tag{7}$$

where $\mathcal{E}(\cdot)$ denotes transformer encoder layer and $\mathcal{P}(\cdot)$ is a linear projection. Such an encoder enables the model to incorporate a larger contextual field and improve its overall scene understanding by leveraging high-level semantic information.

BEV Geometric-aware Encoder To fully leverage the geometric properties (i.e., translational invariance) of the top-view plane, we propose a geometry-aware BEV encoder in BEV space. Specifically, given a point p_{fv} with coordinates (u_f, v_f) in the multi-scale FV feature $\left\{\mathbf{F}^l_{fv}\right\}_{l=1}^{N_f}$, IPM maps the point p_{fv} to the corresponding point p_{bev} with coordinates (x_b, y_b) in BEV space:

$$\begin{bmatrix} x_b \\ y_b \\ 0 \end{bmatrix} = \mathbf{S}_{f \to b} \cdot \mathbf{H}_{c \to g} \cdot K^{-1} \cdot \begin{bmatrix} u_f \\ v_f \\ 1 \end{bmatrix}, \tag{8}$$

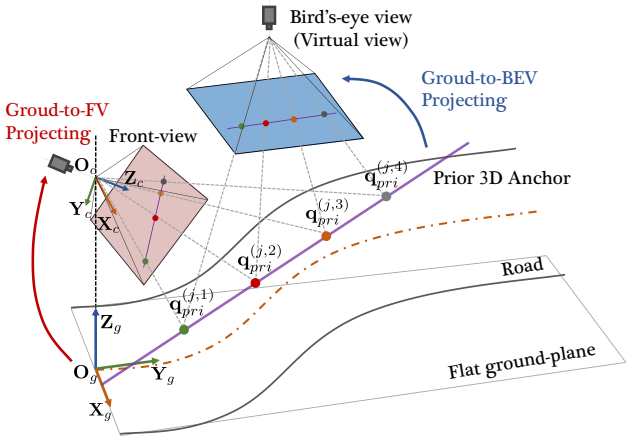


Figure 5: Illustration of 3D prior anchor in FV and BEV projections.

where $\mathbf{S}_{f\to b}$ is the scale matrix between front-view and BEV, $\mathbf{H}_{c\to g} \in \mathbb{R}^{3\times 3}$ denotes the homography matrix from camera coordinates to ground coordinates. Similar to FPN (Lin et al. 2017a) structure, two BEV features from adjacent pyramidal layers, $\mathbf{F}_{bev}^{l-1}$ and $\mathbf{F}_{bev}^{l}$, are merged after applying the downsampling layer $\mathcal{R}_l$ to the spatial dimension of the previous layer. The convolution block $\mathcal{C}_l(\cdot)$ then processes this mixture to propagate geometric information to higher layers in a coarse-to-fine manner. The process is defined as:

$$\tilde{\mathbf{F}}_{bev}^{l} = \mathcal{C}_l\left(\mathbf{F}_{bev}^{l} \oplus \mathcal{R}_l(\tilde{\mathbf{F}}_{bev}^{l-1})\right) \in \mathbb{R}^{H_{bev}^{l}\times W_{bev}^{l}\times C^{l}}, \quad (9)$$

where $\oplus$ is the concatenating operation. In the top view, sharing knowledge between these scales has the potential to enhance the model's robustness in handling complex scenes.

Prior-Guided View-agnostic Feature Alignment

By leveraging the prior anchors provided by PAN, we further introduce the Prior-Guided View-agnostic Feature Alignment Module (PVFA) to effectively align and merge the rich information from both views in a shared sampled space.

Prior-Guided Projection and Sampling Given the j-th prior anchor $\mathbf{A}_{pri}^{j} = \left\{\mathbf{q}_{pri}^{(j,k)}\right\}_{k=1}^{N_p}$ as an example, we project it into the FV and BEV spaces separately as shown in Figure 5. Similar to PAN, we use $\mathcal{P}_{g2fv}(\mathbf{A}_{pri}^{j})$ to denote the projection of prior anchor $\mathbf{A}_{pri}^{j}$ in the FV, which is described in detail in Eq.(2). Then, we employ bilinear interpolation to sample FV anchor features from the output feature $\tilde{\mathbf{F}}_{g2fv}$ of the FV encoder at the projection points $\mathcal{P}_{g2fv}(\mathbf{A}_{pri}^{j})$:

$$\hat{\mathbf{F}}_{fv}^{j} = \tilde{\mathbf{F}}_{fv}(\mathcal{P}_{g2fv}(\mathbf{A}_{pri}^{j})) \in \mathbb{R}^{N_p\times C_{pri}}. \quad (10)$$

By defining the projection in the BEV as $\mathcal{P}_{g2bev}(\cdot) = \mathcal{P}_{fv2bev}(\mathcal{P}_{g2fv}(\cdot))$, we further define the sampling process for BEV space features as follows:

$$\hat{\mathbf{F}}_{bev}^{j} = \tilde{\mathbf{F}}_{bev}^{N_f}(\mathcal{P}_{g2bev}(\mathbf{A}_{pri}^{j})) \in \mathbb{R}^{N_p\times C_{pri}}. \quad (11)$$

View-agnostic Feature Alignment Since $\hat{\mathbf{F}}_{fv}^{j}$ and $\hat{\mathbf{F}}_{bev}^{j}$ are sampled using a uniform prior anchor projection, they can be regarded as view-agnostic, thus easily aligned with the guidance of the prior anchor. Specifically, we first transform the N_p points in $\hat{\mathbf{F}}_{fv}^{j}$ and $\hat{\mathbf{F}}_{bev}^{j}$ into the channel dimension and merge the FV and BEV anchor features using a fusion module $\Phi_{fus}(\cdot)$ given by:

$$\mathbf{F}_{fus}^{j} = \Phi_{fus}(\mathcal{F}(\hat{\mathbf{F}}_{fv}^{j}), \mathcal{F}(\hat{\mathbf{F}}_{bev}^{j})) \in \mathbb{R}^{N_p C_{pri}}, \quad (12)$$

where $\mathcal{F}(\cdot)$ is the flatten operation. This enhanced feature contains different view information, which enables the network to infer the 3D structure in a road scene. Furthermore, the alignment of FV and BEV features using a sparse prior anchor significantly reduces the association space between them, thus improving the model efficiency.

Prediction Loss

Given the fused features corresponding to the j-th prior anchor, we utilize a classification head and a regression head to predict its lane probability $\mathbf{p}^{j} \in \mathbb{R}^{1+N_c}$, x-axis and z-axis offsets $\Delta\mathbf{x}^{j} \in \mathbb{R}^{N_p}$, $\Delta\mathbf{z}^{j} \in \mathbb{R}^{N_p}$, and visibility of each point $\mathbf{vis}^{j} \in \mathbb{R}^{N_p}$ respectively. Consequently, we define our 3D lane proposals based on the prior anchor $\mathbf{A}_{pri}^{j} = \left(\mathbf{x}^{j}, \mathbf{y}, \mathbf{z}^{j}\right)$ as $\mathbf{P^j} = (\mathbf{p}^{j}, \mathbf{x}^{j} + \Delta\mathbf{x}^{j}, \mathbf{y}, \mathbf{z}^{j} + \Delta\mathbf{z}^{j}, \mathbf{vis}^{j})$.

Given N_{pos} pairs of positive proposals $\{\mathbf{P}^{i}\}_{i=1}^{N_{pos}}$ and corresponding ground-truth $\{\mathbf{G}^{i}\}_{i=1}^{N_{pos}}$ with $\mathbf{G}^{i} = \left(\tilde{\mathbf{x}}^{i}, \tilde{\mathbf{z}}^{i}, \tilde{\mathbf{vis}}^{i}\right)$, the loss function can be written as:

$$\begin{aligned}
\mathcal{L}_{3D} = &-\sum_{j=1}^{N_{pri}} \alpha\left(1-\mathbf{p}_t^{j}\right)^{\gamma}\log\mathbf{p}_t^{j} \\
&+\sum_{i=1}^{N_{pos}}\sum_{k=1}^{N_p}\left\|\tilde{\mathbf{vis}}^{(i,k)}\cdot\left(\mathbf{x}^{(i,k)}+\Delta\mathbf{x}^{(i,k)}-\tilde{\mathbf{x}}^{(i,k)}\right)\right\|_1 \\
&+\sum_{i=1}^{N_{pos}}\sum_{k=1}^{N_p}\left\|\tilde{\mathbf{vis}}^{(i,k)}\cdot\left(\mathbf{z}^{(i,k)}+\Delta\mathbf{z}^{(i,k)}-\tilde{\mathbf{z}}^{(i,k)}\right)\right\|_1 \\
&+\sum_{i=1}^{N_{pos}}\sum_{k=1}^{N_p}\left\|\tilde{\mathbf{vis}}^{(i,k)}-\mathbf{vis}^{(i,k)}\right\|_1 .
\end{aligned} \quad (13)$$

To further enhance the representation of the FV and BEV, we introduce a segmentation loss $\mathcal{L}_{seg}$, inspired by LaneNet (Neven et al. 2018). The total loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{3D} + \lambda_{pri}\mathcal{L}_{pri} + \lambda_{seg}\mathcal{L}_{seg}, \quad (14)$$

Where λ_{pri} and λ_{set} are set to 1.0 and 0.1 in our experiments, respectively.

Experiment

Datasets and Evaluation Metrics

The experiments are conducted on two popular benchmark datasets of 3D lane detection: OpenLane (Chen et al. 2022) and Once-3DLanes (Yan et al. 2022). In our experiments, we applied the maximum F1 score, close (0-40m) and far (40-100m) X/Z errors to evaluate the performance of the model.

Methods	Feature Extraction		F1(%)	Cate Acc(%)	X err(m)	Z err(m)	FPS
	Backbone	Space Trans					
3D-LaneNet (Garnett et al. 2019)	VGG-16	IPM	44.1	-	0.479/0.572	0.367/0.443	-
GenLaneNet (Guo et al. 2020)	ERFNet	IPM	32.3	-	0.591/0.684	0.411/0.521	-
PersFormer (Chen et al. 2022)	EfficientNet	Transformer	50.5	92.3	0.485/0.553	0.364/0.431	-
Anchor3DLane (Huang et al. 2023)	ResNet-18	-	53.7	90.9	0.276/0.311	0.107/0.138	-
BEV-LaneDet (Wang et al. 2023)	ResNet-34	MLP	58.4	-	0.309/0.659	0.244/0.631	102
LATR (Luo et al. 2023)	ResNet-50	Transformer	61.9	92.0	**0.219**/0.259	**0.075/0.104**	11
PVALane (Ours)	ResNet-18	IPM	61.2	93.0	0.249/0.263	0.094/0.122	**108**
PVALane (Ours)	ResNet-50	IPM	62.7	93.4	0.232/0.259	0.092/0.118	53
PVALane (Ours)	Swin-B	IPM	**63.4**	**93.5**	0.226/**0.257**	0.093/0.119	31

Table 1: Comparison with state-of-the-art methods on OpenLane validation set. "Space Trans" denotes space transform. "Cate Acc" means category accuracy. Our PVALane achieves state-of-the-art performance on F1-score and category accuracy.

Methods	Backbone	Mean	Up& Down	Curve	Extreme Weather	Night	Intersection	Merge& Split
3D-LaneNet (Garnett et al. 2019)	VGG-16	41.7	40.8	46.5	47.5	41.5	32.1	41.7
GenLaneNet (Guo et al. 2020)	ERFNet	26.4	25.4	33.5	28.1	18.7	21.4	31.0
PersFormer (Chen et al. 2022)	EfficientNet	47.3	42.4	55.6	48.6	46.6	40.0	50.7
Anchor3DLane (Huang et al. 2023)	ResNet-18	50.1	46.7	57.2	52.5	47.8	45.4	51.2
BEV-LaneDet (Wang et al. 2023)	ResNet-34	53.8	48.7	63.1	53.4	53.4	50.3	53.7
LATR (Luo et al. 2023)	ResNet-50	58.3	55.2	**68.2**	57.1	55.4	52.3	**61.5**
PVALane (Ours)	ResNet-18	57.5	52.6	65.7	59.5	56.5	52.2	58.7
PVALane (Ours)	ResNet-50	59.0	54.1	67.3	62.0	57.2	53.4	60.0
PVALane (Ours)	Swin-B	**60.1**	**56.1**	67.7	**64.0**	**58.6**	**53.6**	60.8

Table 2: Comparison with state-of-the-art methods under different scenarios. "Mean" denotes the average F1 score of all scenarios. Our PVALane achieves a significant improvement in extremely challenging scenarios (e.g., Extreme Weather and Night).

Method	F1(%)	R(%)	P(%)	CD(m)
3D-LaneNet	44.73	35.16	61.46	0.127
PersFormer	74.33	69.18	80.30	0.074
Anchor3DLane	74.87	69.71	80.85	0.064
PVALane (Ours)	**76.35**	**70.83**	**82.81**	**0.059**

Table 3: Comparison with state-of-the-art methods on ONCE-3DLanes validation set. "P", "R", and "CD" denote Precision, Recall, and CD Error respectively.

Implementation Details We adopt ResNet-50 (He et al. 2016) with ImageNet (Deng et al. 2009) pre-trained weights as the CNN backbones. Anchor filtering threshold τ in Eq.(5) is set to 0.2 and the maximum number of prior anchors is set to 1000. For ResNet, the features after block1 are extracted to construct 4 pyramidal layers of the BEV encoder, and the block4 feature is passed into the FV encoder to obtain the FV features. Four A100s are used to train the model and the batch size is set to 32. In addition, PVALane is trained in an end-to-end manner using the Adam optimization algorithm (Kingma and Ba 2017) with a learning rate of $2e^{-4}$. During training, λ_{pri} and λ_{seg} in Eq.(14) are set to 1.0 and 0.1, respectively.

Main Results

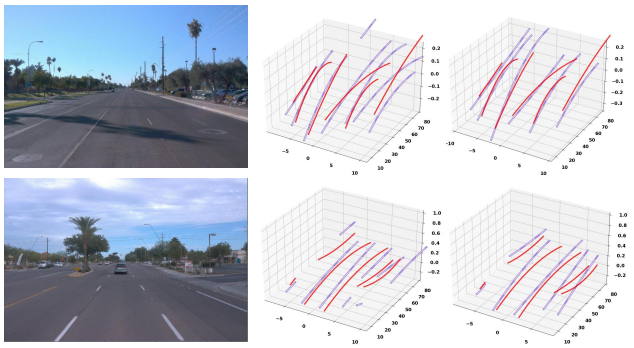
We compared our approach with five state-of-the-art methods: 3D-LaneNet (Garnett et al. 2019), GenLaneNet (Guo et al. 2020), PersFormer (Chen et al. 2022), Anchor3DLane (Huang et al. 2023) and BEV-LaneDet (Wang et al. 2023).

OpenLane We present results on the OpenLane validation set in Table 1, from which it can be seen that PVALane achieves state-of-the-art results on F1 score and category accuracy. Using ResNet-50 as the backbone, we outperform the BEV-LaneDet and LATR by 4.3% and 0.8% in F1 score. Furthermore, by utilizing Swin Transformer as the backbone, PVALane achieves a significant boost in performance. As shown in Table 2, taking the averaged F1 score of all scenarios as the metric, we outperform BEV-LaneDet and LATR by 5.2% and 0.7% respectively. In addition, our method achieves a significant improvement in extremely challenging scenarios (e.g., Extreme Weather and Night), demonstrating the robustness of our method.

To demonstrate the efficiency of PVALane, we conduct experiments on the inference speed of PVALane using various backbones, as shown in Table 1. Using the ResNet-18 as the backbone, PVALane achieves a high speed of 108 FPS, meeting the real-time requirements of autonomous driving.

ONCE-3DLanes In Table 3, we present results on the ONCE-3DLanes dataset. Specifically, PVALane outperforms state-of-the-art methods by 1.48% and achieves a significant improvement in precision. This indicates that PVALane is capable of filtering redundant anchors in a prior-guided manner, thereby reducing false detections as compared to the dense anchor-based approach.

Qualitative Results To better demonstrate our method, we visualized the detection results during the testing phase, as shown in Figure 6. It can be found that PVALane significantly reduces false positive lanes and shows more precise



(a) FV Image (b) Baseline (c) Our-3D

Figure 6: Qualitative results of the proposed PVALane and the baseline on the Openlane dataset. The **red** and **purple** lanes indicate ground truth and prediction, respectively.

	Model	F1(%)	Cate Acc(%)	X err(m)
(I)	Baseline	55.3	88.5	0.317/0.338
(II)	+ PAN	59.3	92.3	0.264/0.290
(III)	+ PVFA	60.1	92.2	0.266/0.301
(IV)	+ PVFA†	**60.5**	**92.5**	**0.258/0.296**

Table 4: Ablation study on OpenLane validation set. "PAN" denotes Prior Anchor Network. "PVFA" denotes Prior-Guided View-agnostic Feature Alignment. "PVFA†" denotes incorporating View-specific Feature Encoding.

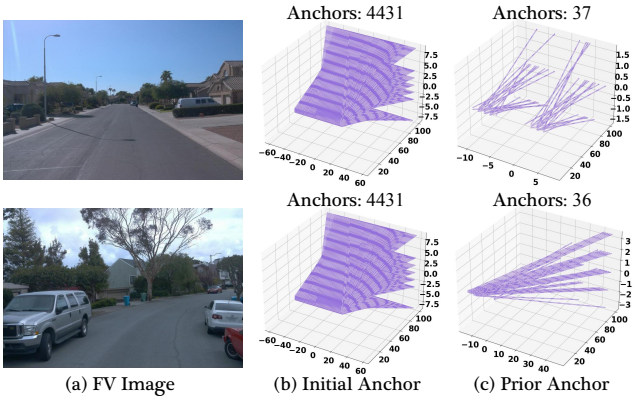
detection than the Anchor3DLane. In addition, we further visualize the generation process of the prior anchor, Figure 7. Based on the prior knowledge provided by the FV features, PVALane significantly reduces the number of anchors used for downstream lane detection.

Ablation Study

In this section, we show an ablation analysis to validate the effectiveness of the proposed modules and justify the parameter choices we made. All experiments were conducted using the ResNet-18 backbone with a batch size of 8.

Different components As we can see in Table 4, without PAN and PVFA, our baseline yields an F1 score of 55.4%. By simply introducing PAN to the baseline method, we achieve a significant boost in performance by 4.0%, improving the F1 score to 59.3%. What's more, the addition of PVFA was able to further improve the performance of our model to 60.1%, but enhancing information such as semantic or structural information from different views will make it better. We can observe that the view-specific encoding of FV and BEV features increases the F1 score to 60.5%.

Score threshold of the prior anchor To find out the influences of prior anchors with different numbers, we conducted a series of experiments with different score thresholds on prior anchor generation. As shown in Table 4, results above the threshold of 0.2 fail to fit the lanes since less than 50 anchors are selected as the prior anchor, which leads to a lower F1 score. When the threshold is higher than 0.2, some extra interference may be introduced into the prior anchor, and



(a) FV Image (b) Initial Anchor (c) Prior Anchor

Figure 7: Visualization of the prior anchor generation.

τ	Anchors	F1(%)	Cate Acc(%)	X err(m)
-	1000	60.0	91.6	0.274/0.314
0.2	108	**60.5**	**92.5**	**0.258/0.296**
0.4	49	60.2	92.1	0.265/0.300
0.6	25	58.5	91.1	0.284/0.314

Table 5: Ablation study on the threshold τ of prior anchor generation.

Methods	F1(%)	Cate Acc(%)	X err(m)
PVFA w/o FV	59.8	91.4	0.278/0.308
PVFA w/o BEV	59.1	92.1	0.272/0.300
PVFA	**60.5**	**92.5**	**0.258/0.296**

Table 6: Ablation study on Prior-Guided View-agnostic Feature Alignment Module (PVFA).

therefore our model performance gets slightly hurt. Therefore, PVALane equipped with about 100 prior anchors is desirable to guide downstream 3D lane detection. As such, we set τ=0.2 in all of our experiments.

Different view features Compared with simply using information from a single view (i.e., FV or BEV), the proposed PFVA achieves gains by 0.7% and 1.4% in F1 score (see Table 6). This demonstrates that FV and BEV often contain different information (i.e., semantic and geometric) in the feature space. By guiding the merge process through prior anchors, the PVFA can effectively utilize the information from both views to enhance the lane representation.

Conclusions

In this work, we propose PVALane, a simple yet accurate prior-guided framework tailored for 3D lane detection. By utilizing the strong prior provided by 2D predictions, a nearly cost-free prior anchor is generated to reduce the lane search space and thus achieve efficient 3D lane detection. Additionally, we further represent the lanes in different view spaces and align the semantic and geometric information from FV and BEV features under the guidance of the prior anchor. Extensive experiments demonstrate the superior performance of our method compared to existing state-of-the-art approaches.

Acknowledgments

This work was supported in part by the Key Areas Research and Development Program of Guangzhou Grant 2023B01J0029, Science and technology research in key areas in Foshan under Grant 2020001006832, the Science and technology projects of Guangzhou under Grant 202007040006, the Guangdong Provincial Key Laboratory of Cyber-Physical System under Grant 2020B1212060069, the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515012534, the National Statistical Science Research Project of China (No. 2022LY096), the Science and Technology Development Fund, Macau SAR, under Grant 0087/2020/A2 and Grant 0141/2023/RIA2.

References

Chen, L.; Sima, C.; Li, Y.; Zheng, Z.; Xu, J.; Geng, X.; Li, H.; He, C.; Shi, J.; Qiao, Y.; et al. 2022. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *Proceedings of the European Conference on Computer Vision*, 550–567. Springer.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Efrat, N.; Bluvstein, M.; Oron, S.; Levi, D.; Garnett, N.; and Shlomo, B. E. 2020. 3D-LaneNet+: Anchor Free Lane Detection using a Semi-Local Representation. arXiv:2011.01535.

Garnett, N.; Cohen, R.; Pe'er, T.; Lahav, R.; and Levi, D. 2019. 3d-lanenet: end-to-end 3d multiple lane detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2921–2930.

Guo, Y.; Chen, G.; Zhao, P.; Zhang, W.; Miao, J.; Wang, J.; and Choe, T. E. 2020. Gen-lanenet: A generalized and scalable approach for 3d lane detection. In *Proceedings of the European Conference on Computer Vision*, 666–681. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, S.; Shen, Z.; Huang, Z.; Ding, Z.-h.; Dai, J.; Han, J.; Wang, N.; and Liu, S. 2023. Anchor3dlane: Learning to regress 3d anchors for monocular 3d lane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17451–17460.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision*, 1–18. Springer.

Liang, M.; Yang, B.; Wang, S.; and Urtasun, R. 2018. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European Conference on Computer Vision*, 641–656.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017b. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2980–2988.

Liu, L.; Chen, X.; Zhu, S.; and Tan, P. 2021a. Condlanenet: a top-to-down lane detection framework based on conditional convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3773–3782.

Liu, R.; Chen, D.; Liu, T.; Xiong, Z.; and Yuan, Z. 2022. Learning to predict 3d lane shape and camera pose from a single image via geometry constraints. In *Proceedings of the AAAI Conference on Artifcial Intelligence*, volume 36, 1765–1772.

Liu, R.; Yuan, Z.; Liu, T.; and Xiong, Z. 2021b. End-to-end lane shape prediction with transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3694–3702.

Luo, Y.; Zheng, C.; Yan, X.; Kun, T.; Zheng, C.; Cui, S.; and Li, Z. 2023. LATR: 3D Lane Detection from Monocular Images with Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7941–7952.

Neven, D.; De Brabandere, B.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2018. Towards end-to-end lane detection: an instance segmentation approach. In *2018 IEEE intelligent vehicles symposium (IV)*, 286–291. IEEE.

Pan, X.; Shi, J.; Luo, P.; Wang, X.; and Tang, X. 2018. Spatial as deep: Spatial cnn for traffic scene understanding. In *Proceedings of the AAAI Conference on Artifcial Intelligence*, volume 32.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, R.; Qin, J.; Li, K.; Li, Y.; Cao, D.; and Xu, J. 2023. BEV-LaneDet: An Efficient 3D Lane Detection Based on Virtual Camera via Key-Points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1002–1011.

Yan, F.; Nie, M.; Cai, X.; Han, J.; Xu, H.; Yang, Z.; Ye, C.; Fu, Y.; Mi, M. B.; and Zhang, L. 2022. Once-3dlanes: Building monocular 3d lane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17143–17152.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. arXiv:2010.04159.