

Flexible 3D Lane Detection by Hierarchical Shape Matching

Zhihao Guan¹, Ruixin Liu¹, Zejian Yuan¹,
Ao Liu², Kun Tang², Tong Zhou², Erlong Li², Chao Zheng², Shuqi Mei²

¹Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China

²T Lab, Tencent Map, Tencent, China

{duduguan,sweetylrx}@stu.xjtu.edu.cn, yuan.ze.jian@xjtu.edu.cn,

{allliu, kyriezhou, kunntang, erlongli, chriszheng, shawnmei}@tencent.com

Abstract

As one of the basic while vital technologies for HD map construction, 3D lane detection is still an open problem due to varying visual conditions, complex typologies, and strict demands for precision. In this paper, an end-to-end flexible and hierarchical lane detector is proposed to precisely predict 3D lane lines from point clouds. Specifically, we design a hierarchical network predicting flexible representations of lane shapes at different levels, simultaneously collecting global instance semantics and avoiding local errors. In the global scope, we propose to regress parametric curves w.r.t adaptive axes that help to make more robust predictions towards complex scenes, while in the local vision the structure of lane segment is detected in each of the dynamic anchor cells sampled along the global predicted curves. Moreover, corresponding global and local shape matching losses and anchor cell generation strategies are designed. Experiments on two datasets show that we overwhelm current top methods under high precision standards, and full ablation studies also verify each part of our method. Our codes will be released at <https://github.com/Doo-do/FHLD>.

Introduction

While advanced driver-assistance system (ADAS) entering the lives of millions, one of its supporting tasks, centimeter-level lane detection for high-definition(HD) map construction is still limited by many **challenges**. First, diverse difficult visual conditions. The visual clue is not always clear in real driving scenes due to occlusions or severe illumination conditions. Second, complex typologies. The multifarious combination of varying numbers and directions of lanes, along with merges and splits make the discrimination of lane instances beset with difficulties. Moreover, more strict precision is needed. Widely used evaluation metrics like TuSimple's are too permissive to errors in local offset(Tabelini et al. 2021b), and thus methods with higher precision are still needed for accurate uses like HD map construction.

Over the past decades, most researchers carry out lane detection with front-view images as input. Limited by the physical characteristics of RGB optical sensors, it cannot deal with poor illumination well. These years, more and more attention has been paid to lane detection with point

cloud data. **Point cloud** is not only insensitive to illumination conditions, it can also preserve the accuracy of 3D information of lanes even in a very long distance. Moreover, its privacy-free feature is also a necessity for actual uses. Following all methods extracting lane lines from point cloud that we know, we structure the point cloud in a Bird's-eye-view (BEV) and regard encoded BEV feature maps as input.

Existing methods function in a segmentation manner which needs cumbersome clustering processes(Ma et al. 2020), or focuses on global information by predicting parametric lanes while ignoring local errors(Liu et al. 2021b), or some others based on preset anchors while losing the flexibility of predictions(Tabelini et al. 2021a). In this paper, we propose an end-to-end framework with hierarchical outputs, to fully combine both the advantages of top-down manner which can collect all the global clues indicating instance information, and the strength of bottom-up manner that focuses on modeling local geometry and predicting accurate positions, to predict each lane as a sequence of points, providing more flexibility for detection results.

Specifically, in the global parametric curve branch, a Transformer structure is used to get a 3D parametric description for each lane. Notably, we propose a new representation for each curve based on the local adaptive reference axis, which incorporates the position of lane terminals, and represents the curve w.r.t its own reference axis to help the network fit better on complex scenes like oblique and curved lanes and hard topologies. Given the shape and instance information predicted from the global branch, dynamic anchor cells are generated along the lane and the corresponding cropped features are fed to the local branch for more precise and flexible predictions.

In the local shape prediction branch, we detect line segment structure inside each anchor cell and use the center point of each as the final result, instead of directly detecting the specific lane center point inside every anchor. In addition, we optimize it by matching the ground truth Gaussian distribution along each lane segment, instead of regressing the segment parameters directly, introducing more explicit contextual information into account to shrink the searching space and optimize all the parameters simultaneously. Moreover, the positional relationships among predictions in different cells are also constrained, improving the smoothness of outputs. Notably, local predictions in dynamic anchor

cells generated from global results are of certain instances and orders, thus eliminating clustering post-processing.

We collect a point cloud dataset RoadBEV on both highways and country roads in different cities for realistic evaluation. The results of comparison experiments on it and also on a self-annotated public dataset sub-KCUD, under a relatively strict metric with distance thresholds of 10 cm and 30 cm, show that our method outperforms recent state-of-the-art works in accurate detection. And also, full ablation studies validate the effectiveness of every part of our method. We summarise our main contributions as follows:

- An end-to-end hierarchical framework is designed which fully fuses both the global and local level information for accurate and flexible 3D lane detection.
- We propose both a new global parametric representation for curves and a local shape description for line segments to enhance the robustness and flexibility of results.
- The dynamic anchor cells generation module and also hierarchical shape matching strategy are proposed to help networks detect line elements hierarchically better.

Related Works

Traditional methods(Sun, Tsai, and Chan 2006; Jung et al. 2019) exploit hand-crafted filters and adopt manual-designed processes like morphological filtering to extract and synthesize specialized features, resulting in poor robustness. Recently, data-driven trends as represented by CNN make it easier to distinguish more subtle and complex features for lane detection. These deep learning methods can be divided into several paradigms as illustrated below.

Segmentation-based methods. Following the bottom-up manner, SAD(Hou et al. 2019) and RESA(Zheng et al. 2021) predict a label for each pixel in the front-view camera images, and similar methods are also used for 3D point cloud data(Wen et al. 2019; Ma et al. 2020). These methods improve the detection performance by strengthening the priors of lanes or ameliorating networks for better feature extraction. However, cumbersome post-processing is still needed to cluster these isolated points into corresponding instances. Meanwhile, methods following this paradigm will output unnecessary points and be sensitive to local imperfections as well. By comparison, our method fully utilizes top-down information to guide the local shape prediction to generate flexible results and avoid post-processing.

Grid-based methods. This can be viewed as a combination of grid-level segmentation and accurate offset regression. PointLaneNet(Chen, Liu, and Lian 2019) directly predicts x-coordinates for fixed y values for each lane. UFast(Qin, Wang, and Li 2020) speeds up predictions by modeling it as a row-based selecting problem and handles the no-visual-clue problem by enlarging the receptive field. PINet(Ko et al. 2021) improves the instance clustering by contrastively learning the embedding features for each grid. CondlaneNet(Liu et al. 2021a) resolves instance-level discrimination with the help of the proposal head and conditional convolution. The cells in these methods are not flexible so the performance may be influenced under the situations when two adjacent lanes fall in the same grid, or one

lane line locates at the middle seam of two grids. To avoid these difficulties, our method predicts lane elements only within specific dynamically proposed anchor cells along the predicted global parametric curves.

Anchor-based methods. Conventional detection methods(Chen et al. 2020) use the bounding box as a coarse proposal for the whole lane, which is not suitable for the slender structure of lanes. Dagmapper(Homayounfar et al. 2019) and HRAN(Homayounfar et al. 2018) use iteratively proposed cells for local topology nodes searching from 3D point cloud data. However, they do not explicitly make full use of global semantic information, and thus local errors may be accumulated at the later stage of iterations. Some other methods like Lane-ATT(Tabelini et al. 2021a) and SGNet(Su et al. 2021) use dense line-shape anchors to provide better priors, where the anchors need to be pre-defined, and NMS is needed to remove redundant predictions. Following the idea of proposals but instead of setting dense anchors, we sample anchor cells along the predicted parametric lines to both provide a better global understanding and also enable end-to-end training.

Parametric-based methods. LSTR(Liu et al. 2021b), PolyLaneNet(Tabelini et al. 2021b) and BezierLaneNet(Feng et al. 2022) follow the top-down manner to model each lane as a parameterized curve and predict the global parameters directly from the network. However, these methods model the curve equation defined w.r.t fixed y-axis, which limits the flexibility for predicting diverse lanes. More importantly, local errors are ignored by these methods. In contrast, we propose a new 3D parametric representation for lane lines providing more flexibility, and also introduce extra local shape prediction to help get more precise localization.

Problem Definition

We take the aggregated point cloud data as raw input, and feed the encoded BEV maps to networks to predict each lane line as a sequence of sparse 3D points along it, since the point set is more flexible for the slender structure of lane lines than other representations.

Since lane lines have few complex structures along height direction, we follow the common practice to structure the point cloud in BEV. Specifically, we project N points (each with 4 attributes: coordinates x, y, z , and reflective intensity r that distinguishes road markings from surroundings) onto the ground plane, rasterize it to $W \times H$ grids, and obtain the encoded BEV maps $I \in \mathbb{R}^{W \times H \times 4}$, of which the 4 channels in turn represent the average intensity, density, height variance, and minimum height of points in each grid.

Methodology

The end-to-end Flexible and Hierarchical Lane Detector (FHLN) we proposed is shown in Fig. 1, where line elements are represented and predicted at different levels. For global understanding, each lane line is predicted in a parametric way based on its adaptive reference axis. For more flexible and accurate outputs, a local line segment shape prediction branch is also proposed, to finally produce a sequence of points in dynamic anchor cells generated along each glob-

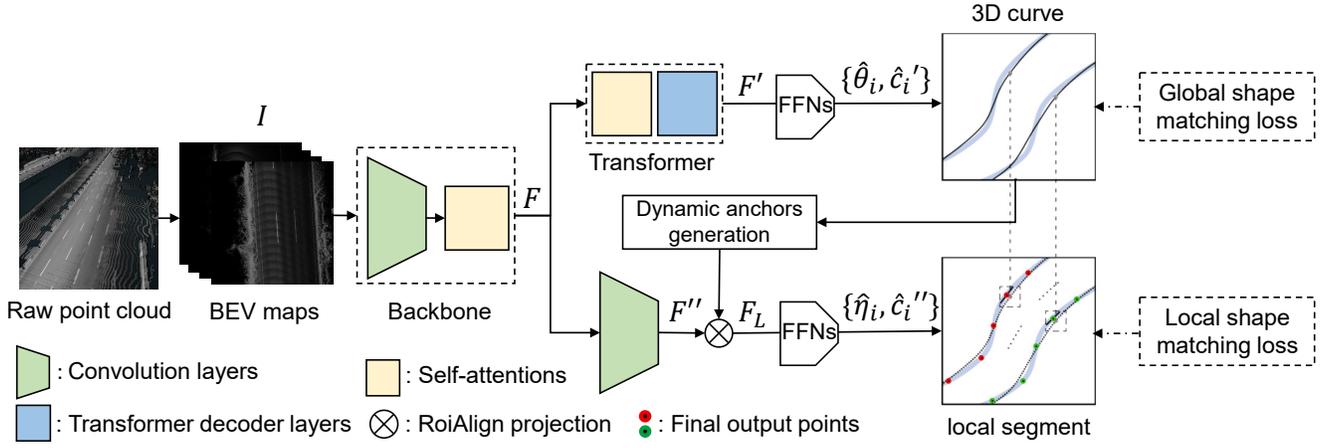


Figure 1: Framework structure. FHLD contains a shared backbone for feature extraction, two different branches separately detecting lane elements in different levels, and a dynamic anchors generation module to pass global prediction to local scope.

ally predicted curve. With accordingly designed hierarchical losses and training strategies, our method fully fuses high-level semantic information and local fine structure of lanes to both obtain global vision and avoid local errors.

Hierarchical Shape Representation

As shown in Fig.2, lane lines are represented at different levels, emphasizing global features of lane curves and local features of line segment shapes simultaneously to carry out hierarchical predictions.

Global parametric shape representation. To represent 3D lines in a more flexible way, we model each one as a cubic polynomial curve w.r.t. its own reference axis adaptively. Specifically, as shown in Fig.2(a), for each line we incorporate its starting position (lower terminal) \mathbf{p}_s and ending position (higher one) \mathbf{p}_e as a part of our parametric expression, which indicates the basic position of the lane. We set \mathbf{p}_s as the origin and $\overrightarrow{\mathbf{p}_s\mathbf{p}_e}$ as the positive direction of reference axis. Then, each line can be described with a set of curve parameters $\theta = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{p}_s, \mathbf{p}_e)$ which determines both the shape and position of this curve, written as:

$$\mathbf{p} = f(t; \theta) = \mathbf{A}t^3 + \mathbf{B}t^2 + \mathbf{C}t + \mathbf{D}, \quad (1)$$

$$t = \frac{\|\overrightarrow{\mathbf{p}_s\tilde{\mathbf{p}}}\|}{\|\overrightarrow{\mathbf{p}_s\mathbf{p}_e}\|}, \quad (2)$$

where $t \in [0, 1]$ denotes the relative position of any point $\tilde{\mathbf{p}} \in \mathbb{R}^3$ moving on the reference axis between \mathbf{p}_s and \mathbf{p}_e , $\mathbf{p} \in \mathbb{R}^3$ denotes the corresponding point on the curve, and $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D} \in \mathbb{R}^3$ the cubic polynomial coefficients, $\|\cdot\|$ the magnitude of vector.

Noted that when $t = 0$ or $t = 1$, corresponding reference point $\tilde{\mathbf{p}}$ overlaps on \mathbf{p}_s or \mathbf{p}_e respectively so the boundary conditions are shown below:

$$\begin{cases} \mathbf{p}_s = f(0; \theta) = \mathbf{D} \\ \mathbf{p}_e = f(1; \theta) = \mathbf{A} + \mathbf{B} + \mathbf{C} + \mathbf{D} \end{cases} \quad (3)$$

It can be seen that 12 parameters (4 for each dimension) are required to describe each 3D curve. Compared with existing descriptions that model the curve w.r.t. y-axis, our

formation gets rid of the binding with fixed reference axis, providing more flexibility and reasonable searching space, especially for inclined and complex lanes, as proven by experiments. Globally parametric prediction can fully excavate lane semantics in a top-down manner and provide abundant positional priors for local shape detection to further reduce local errors.

Local shape description. To further predict more flexible and accurate results in local vision, each lane line is modeled as a sequence of segments, with their center points regarded as final outputs. Specifically, within dynamically sampled anchor cells along each lane (illustrated in the following section), we model the local line structure in each cell as a thin bar with a fixed pre-defined width w_l , as shown in Fig.2(b). Its position and shape can be described by a set of segment parameters $\eta = (\mathbf{p}_o, l, \alpha)$, where \mathbf{p}_o denotes the 3D center point position of the bar, i.e. final outputs, l is the length of this line segment in this cell, and α is the heading angle of this line segment. Notably, since the size of anchor cells only corresponds to a very tiny distance in real scenes, each comprising very little height difference, and thus the tilt and roll angles of line segments are just ignored in practice.

Compared to directly modeling the local structure as the isolated center point, searching for a detailed bar structure helps to provide more context information in a smaller searching space, reducing the difficulty of regressing. Finally, with the curvature-aware dynamic anchoring strategy when inferencing, such local shape representation provides more flexible and locally accurate prediction results.

Network Structure with Hierarchical Outputs

FHLD contains a shared feature extraction backbone, two different branches separately searching lane elements at different levels, and a dynamic anchors generation module to pass global information for local segment shape searching. Specifically, input BEV maps I , our backbone composed of CNNs and a self-attention module fuses the global contextual information and outputs feature maps $F \in \mathbb{R}^{w \times h \times c}$.

Global parametric curve branch. A DETR-like structure

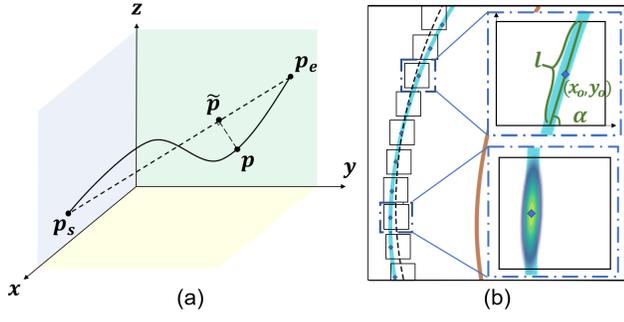


Figure 2: Hierarchical shape representation of lane lines. (a): Global parametric description. (b): Local segment shape representation & Visualization of Gauss distribution.

is adopted. To fully gather the context across entire feature maps to perceive the global shape, position, and instance information of lanes, extra self-attention blocks are adopted to encode F , and together with positional queries, they are then fed to Transformer decoder to generate F' , the attended feature sequences for each of N predictions. Finally, FFNs in the classification head and parameter head output result set $\{(\hat{c}'_i, \hat{\theta}_i) | i=1, \dots, N\}$, where \hat{c}'_i are the confidence scores for each class (lane or non-lane). In practice, the outputs from each layer of the decoder are supervised.

Local segments shape branch. After further encoding F with convolution layers to $F'' \in \mathbb{R}^{w \times h \times c}$, 2D positions of generated anchors are mapped to F'' using RoiAlign(He et al. 2017) to crop corresponding feature sequences F_L . For each of these cropped local features, a set of results $\{(\hat{c}'_i, \hat{\eta}_i) | i=1, 2, \dots, B\}$ is predicted, where \hat{c}'_i are predicted classification scores, B is the number of cells generated on this lane. The position of segment center \hat{p}_o is predicted in a relative form w.r.t corresponding anchor cell center.

Dynamic anchors generation module. Along each predicted parametric lane, a sequence of anchor cells is sampled following the strategy introduced in the following section. Every cell is projected onto F'' from the original resolution image and is quantified to F_L with the help of RoiAlign. This projection guarantees that the local segment branch concentrates on the explicitly specific region of interest along the lane and reduces the false positive predictions.

Training with Hierarchical Shape Matching Loss

To supervise hierarchical outputs in two branches, The loss function is also designed in two parts accordingly.

$$L = L_{GSM} + L_{LSM}. \quad (4)$$

Global Shape Matching Loss. Since the ground truth lanes are described in the global coordinate system while the predicted ones are defined on the local reference axis, to define the difference between them, a bridge is needed. Specifically, we represent the i -th ground truth lane Q_i as a sequence of densely sampled points. For each point $p \in Q_i$, its corresponding relative position t_p on the reference axis can be calculated by projecting it to $\vec{p_s p_e}$, written as:

$$t_p = \frac{\vec{p_s p_e} \cdot \vec{p_s p}}{\|\vec{p_s p_e}\|^2}, \quad (5)$$

where \cdot means dot product of two vectors, p_s and p_e are the ground truth terminals. At each relative position t_p , the corresponding prediction point $\hat{p} = f(t_p, \hat{\theta}_i)$ can be calculated. And then, a curve fitting loss L_f is used to describe the gap between each pair of points on i -th curve, written as:

$$L_f(\hat{\theta}_i) = \frac{1}{|Q_i|} \sum_{p \in Q_i} \|\hat{p} - p\|, \quad (6)$$

where $|\cdot|$ denotes the size of the set.

Moreover, a cross-entropy loss is used for the predicted logits as well. Overall, with $o(c_i)$ denoting the predicted probability for label $c_i \in \{0, 1\}$, with assigning $\hat{e}(i)$ -th predicted lane matched to Q_i after bipartite matching described in next section, and with ground truth set Q padded to the number of N with non-lanes, the global shape matching loss can be written as:

$$L_{GSM} = \sum_{i=1}^N (-\lambda_1 \log o_{\hat{e}(i)}(c_i) + \mathbb{1}(c_i \neq 0) L_f(\hat{\theta}_{\hat{e}(i)})), \quad (7)$$

where $\mathbb{1}(\cdot)$ is an indicator function, and λ_i is the weight to balance the effect of loss terms.

Local Shape Matching Loss with Curve Smoothing. Constraints on local branch are mainly composed of a shape matching loss L_{kl} locating precise line segments, and a curve smoothing loss L_{sm} aiming at constraining the smoothness of neighbored local predictions. With the binary classification loss written as L_{cls} and L_z denoting L_2 regression loss for the height of output points, it can be written as:

$$L_{LSM} = \lambda_2 L_{kl} + \lambda_3 L_{sm} + \lambda_1 L_{cls} + L_z. \quad (8)$$

For better matching and locating the shape of segment elements in local scope, we optimize the Gauss distribution g covered on each segment bar, determined by the parameter set η (as visualized in Fig.2(b)). The mean μ and covariance Σ of distribution g can be written as:

$$\begin{aligned} \Sigma^{1/2} &= \mathbf{R} \mathbf{A} \mathbf{R}^\top = \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix} \cdot \\ &\begin{pmatrix} w_l & 0 \\ 0 & \frac{l}{2} \end{pmatrix} \cdot \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix}, \\ \mu &= (x_o, y_o)^\top, \end{aligned} \quad (9)$$

where x_o and y_o are the two components of p_o , and w_l is doubled to cover more contextual message.

Instead of directly regressing the isolated parameters of segments, we constrain the predicted distribution from the ground truth ones using KL-divergence, denoted as L_{kl} :

$$L_{kl} = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^B (KL(g_{i,j} | \hat{g}_{i,j}) + KL(\hat{g}_{i,j} | g_{i,j})), \quad (11)$$

where $\hat{g}_{i,j}$ is the predicted distribution in the j -th anchor on i -th lane. Such design makes its optimization easier than detecting a point or a rotated line directly because the position, angle, and length information of the slender structure can be optimized simultaneously with adaptively adjustable strategy, providing more accurate gradient information (Yang et al. 2021), and abundant contextual information around the lane is explicitly taken into account.

Moreover, thanks to the structured results of global branch, the anchor cells on each lane are of explicit order. We also constrain the positional relationship among predicted results in different cells. Concretely, L_{sm} is used to constrain the second derivative of each curve, i.e., of consecutive output points, to guarantee the smoothness of predictions and avoid local imperfect predictions in unclear scenes like occlusion. With $\hat{\mathbf{p}}_o^{i,j}$ denoting the predicted absolute coordinate in the j -th anchor on i -th lane, it can be written as:

$$L_{sm} = \sum_{i=1}^N \sum_{j=1}^{B-2} \|(\hat{\mathbf{p}}_o^{i,j+2} - \hat{\mathbf{p}}_o^{i,j+1}) - (\hat{\mathbf{p}}_o^{i,j+1} - \hat{\mathbf{p}}_o^{i,j})\|. \quad (12)$$

Training Strategies

Dynamic anchors generation. The local structure of the lane line is searched under the guidance of top-down information, obtaining a better understanding of the entirety. Say concretely, a sequence of anchor cells is sampled dynamically along each parametric lane for further local fine prediction. Dynamic anchoring, instead of conventional fixed grid-based methods, can avoid both dense computations and failure in troublesome situations when the line lies at the middle seam of two grids, or two lanes are in the same grid.

When training, B points are sampled along each predicted parametric lane as the centers of dynamic anchor cells. According to empirical attempts and experimental results, the shape and size of cells here do not affect the results much so we set a fixed size for anchors at $r \times r$ pixels. Thanks to the global branch, here most of the sampled cells contain lane segments, i.e., are positive samples. For better distinguishing the local features, following the idea of contrast learning, negative anchor cells are also randomly sampled around each positive one at a set proportion. Notably, only classification loss L_{cls} is used to supervise these negative samples. Moreover, a little random offset is added as noise to the sampled positive positions to mimic the imperfection in global predictions and improve the generalization.

When inferencing, only positive anchors are sampled, and the sampling rate of each lane is set proportional to its curvature. That is to say, there will be more points predicted on the complex lane structure than on the simple straight one, keeping the compactness of outputs.

Bipartite matching for global predictions. The correspondence between N predictions and M (padded to N) ground truth lanes is built using Hungarian Algorithm (HA) following LSTR. To design the cost matrix $D \in \mathbb{R}^{N \times N}$, only two parts are covered in it, the classification cost and the position cost. Though a more precise comparison between two line instances can be done, under our adaptive parametric representation of curves it is sufficient to match the terminals due to their unambiguity and importance in all road scenes. With $d_{i,j}$ denoting the (i -th, j -th) element in D , and $o_j(c_i)$ the probability of target class label c_i predicted by j -th network output, it can be written as:

$$d_{i,j} = -\lambda_1 o_j(c_i) + \mathbb{1}(c_i \neq 0) L_1 \left((\hat{\mathbf{p}}_s^{(j)}, \hat{\mathbf{p}}_e^{(j)}), (\mathbf{p}_s^{(i)}, \mathbf{p}_e^{(i)}) \right). \quad (13)$$

With the help of HA, a one-to-one match can be guaranteed where the i -th ground truth lanes are matched with the

$\hat{\epsilon}(i)$ -th prediction with $\hat{\epsilon} = \arg \min_{\epsilon} \sum_{i=1}^N d_{i,\epsilon(i)}$.

Experiments

To show the superiority of our framework and also the effectiveness of each design, we have done abundant comparative experiments and ablation studies.

Dataset: Experiments are carried out on two point cloud dataset, a self-collected one named RoadBEV and sub-KCUD, a subset of public KAIST Complex Urban Dataset (Jeong et al. 2019) annotated by ourselves, since there are no readily available large-scale public datasets with 3D lane annotations for the HD map construction task. For each dataset, from the raw point cloud, regions of 25×25 m are cropped along the trajectory at a stride of 13 m. Then we project and rasterize the points to 800×800 pixel BEV maps encoded as mentioned before. To annotate each lane, 3D polylines crossing the center of lane lines are used. At most 8 and at least 0 lines are contained in each BEV map, and complex scenes like splits and merges are included. RoadBEV contains around 25k and sub-KCUD contains nearly 4k BEV images. For each dataset, we randomly set 80% sequences for training and 20% for testing.

Metrics: Following the testing paradigm of DagMapper, we set a more strict evaluation scheme compared with the ones defined by Tusimple or Culane, which is more necessary for high-precision HD maps. Firstly, a large number of points are densely sampled respectively from predicted and ground truth lanes, and then, recall and precision rates are calculated by searching the number of ground truth or predicted points within a distance threshold from the predicted or ground truth points. Under the 10 cm and 30 cm distance threshold, we report the precision(%), recall(%), F1 score(%), and also FPS of all methods. Notably, we care more about stricter **10 cm** standard. FPS is calculated with batch size 1 on Nvidia RTX3090 implemented by Pytorch.

Details: For backbone, we modify commonly used Resnet-18(He et al. 2015) by cutting the output channels of each block into [16, 32, 64, 128] to avoid overfitting. We augment each input BEV map by random shifting, rotating, flipping, scaling, and cropping, transforming it to a size of 640×640 . We train each model for 200k iterations with batch size 64 and initial learning rate 0.0001 using Adam optimizer. The loss coefficients λ_1 , λ_2 and λ_3 are set to 1, 1 and 0.5, anchors sampling ratio is 1:100, and $N=15$, $B=40$. If multiple segments appeared in one cell, the one closest to the cell center is chosen as the ground truth in this cell. For better performance, L_{LSM} is not optimized for first 2k iterations. All of the backbones and hyper-parameter settings are set to the same for fair comparison for all experiments if applicable. Three parallel experiments are carried out for each method.

Comparative Experiments

Baseline: Because there are few strong open-sourced baselines for point cloud lane detection, and due to its similarity to front-view detection tasks, we have to modify some strongest RGB methods to adapt to this task as comparisons. Our method is compared with adapted Lane-ATT, LSTR, PINet, UFast, CondLaneNet, SAD, PolyLaneNet,

Method	RoadBEV						sub-KCUD						FPS
	10 cm			30 cm			10 cm			30 cm			
	Precision	Recall	F1										
LaneATT	71.72	70.35	71.03	90.26	89.99	90.12	72.15	71.99	72.07	91.32	90.02	90.67	94
LSTR	71.70	72.24	71.97	91.51	90.23	90.87	72.01	72.40	72.20	91.45	90.80	91.12	116
PINet	70.92	66.32	68.54	92.33	90.27	91.29	71.58	67.85	69.67	90.39	88.99	89.68	5
UFast	63.52	68.34	65.84	87.55	93.81	90.57	69.92	68.16	69.03	89.02	93.12	91.02	133
CondLaneNet	74.13	73.02	73.57	93.09	92.97	93.03	72.25	72.30	72.27	90.30	92.79	91.53	75
SAD	71.33	70.25	70.79	89.77	93.42	91.56	68.72	67.99	68.35	89.71	90.89	90.30	35
PolyLaneNet	58.69	57.07	57.87	82.23	80.99	81.61	60.73	59.78	60.25	83.25	84.39	83.82	104
RESA	60.16	56.45	58.25	79.80	80.75	80.27	62.93	60.42	61.65	80.65	78.32	79.47	25
FHLD(Ours)	76.68	78.49	77.57	92.99	94.50	93.74	74.42	74.45	74.43	91.49	93.05	92.20	97

Table 1: Comparison experiments results on RoadBEV and sub-KCUD dataset.

Global description		Local matching		Training		Results (10 cm)			Results (30 cm)		
Adp.	Fixed	Shape	Point	L_{sm}	Noise	Precision	Recall	F1	Precision	Recall	F1
✓		✓		✓	✓	76.68	78.49	77.57	92.99	94.50	93.74
	✓	✓		✓	✓	74.89	76.93	75.90(-1.67)	92.47	94.01	93.23(-0.51)
✓			✓	✓	✓	74.77	76.48	75.62(-1.95)	92.36	93.84	93.09(-0.65)
✓						72.66	73.25	72.95(-4.62)	92.18	91.63	91.99(-1.75)
	✓					71.73	72.23	71.98(-5.59)	91.50	90.24	90.87(-2.87)
✓		✓			✓	75.58	77.79	76.67(-0.90)	92.49	94.22	93.35(-0.39)
✓		✓		✓		75.50	77.81	76.64(-0.93)	92.53	94.25	93.38(-0.36)

Table 2: Ablation studies results on RoadBEV dataset.

and RESA. Notably, since most of these methods are hard to expand to 3D prediction, we directly set their height predictions as the ground truth ones, while ours predicts 3D coordinates straight from the model outputs. That is to say, we put other methods we compare with in a better position.

It’s important to note that SAD, RESA, and UFast all need to pre-define the instance labels, and we assign each lane one of 10 labels according to the horizontal positions. LSTR’s modeling for curves is designed for the geometry of front-view images so we adapt it to polynomial curves in BEV. Methods beside LSTR need more or fewer post-processings and lane-ATT needs pre-designed anchor settings. In contrast, our method trains and inferences in a totally end-to-end manner, regardless of the number, position and direction of the lanes in the image. Besides, our method can flexibly output point set at different sparsities depending on the curvature, or following other customized requirements. In this section the qualitative results of output point set are shown equidistantly for clear comparison with other methods. More flexible output options demonstrated in 3D point clouds, as well as more hard samples of complex topologies with both intermediate and final results are visualized in the **Supp.**

The testing results on two datasets can be found in Tab.1. Clearly, although our method is tested under a harder situation, it still overwhelms others, especially under the more accurate 10 cm standard since our flexible output can greatly improve local errors. Specifically, on RoadBEV dataset, FHLD respectively obtains 5.4% and 3.0% higher F1 score than the second best-performed method under 10 cm distance threshold, and obtains 0.8% and 0.7% higher under 30

cm distance threshold. Our methods successfully improve the accurate prediction at the centimeter level. Compared with the parametric-based and cell-based methods most similar to ours, LSTR and PINet, we achieve 7.8% and 13.2% higher scores under 10 cm standard on RoadBEV. Besides, we can notice that thanks to the anchor generation module, the recall of FHLD is also improved much, especially when compared to segmentation methods. In addition, The corresponding qualitative results can be seen in Fig.3, where FHLD performs accurately thanks to flexible and hierarchical outputs, while others either ignore local offsets(LSTR), lack constraints on relative relationships(CondLaneNet), or just miss the whole line due to preset anchors(Lane-ATT).

Ablation Studies

To show the contributions of each part in our method, we carry out abundant ablation studies, including replacing our global curve representation w.r.t adaptive axis(**Adp.**) with traditional description w.r.t fixed axis(**Fixed**), replacing local shape Gauss matching(**Shape**) with direct lane center point regression(**Point**), with or without L_{sm} and **noise** when generating anchors. The results of each combination are shown in Tab.2, where we can see that the local prediction branch contributes much to the final result (+4.62) and the new parametric representation of the global curve (+1.67) as well as the local shape matching strategy (+1.95) also helps a lot, especially under 10 cm threshold.

Effectiveness of global curve representation. To better show the superiority of our global representation, we further test our method with adaptive(**Adp.**) or traditional(**Fixed**)

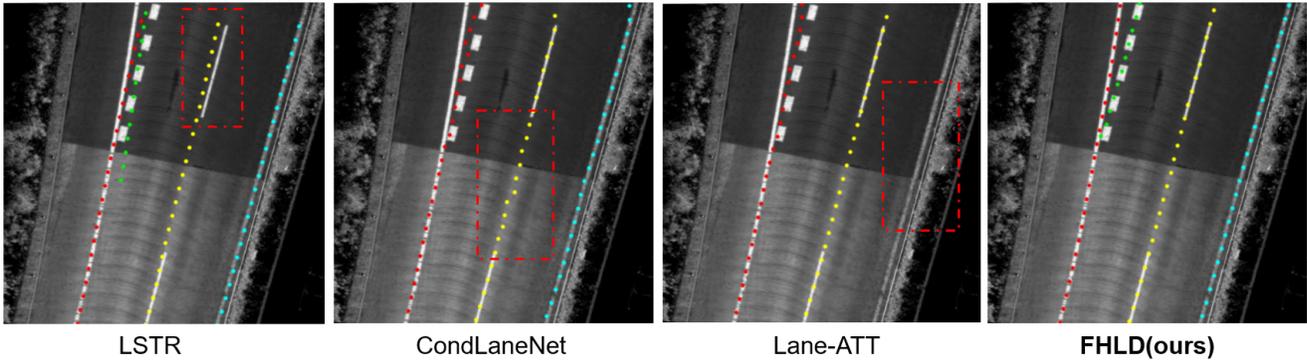


Figure 3: Visualization of prediction results on RoadBEV (Best viewed in colors).

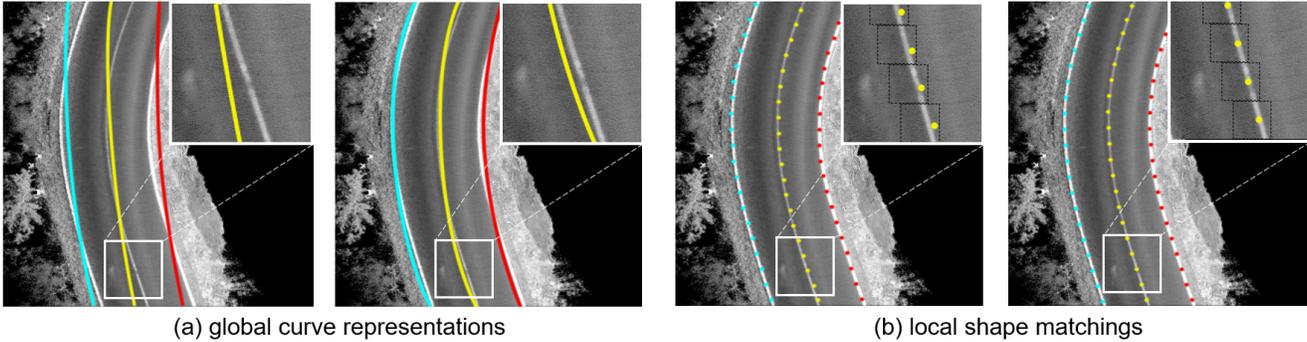


Figure 4: Prediction results visualization of ablation studies (Best viewed in colors). (a): Direct global output with fixed(left) and adaptive(right) representations. (b): Final point output of local prediction by point regressing(left) or shape matching(right).

	Parametric output			Point set output		
	Precision	Recall	F1	Precision	Recall	F1
Adp.	71.07	71.65	71.36	74.62	74.73	74.67
Fixed	67.24	66.58	66.91	71.35	72.01	71.68

Table 3: Results on hard-RoadBEV under 10 cm standard.

representations on a subset dataset hard-RoadBEV, which comprises only lanes with large scopes and curvatures. We report both the results of parametric curves (without local branch) and final point outputs. The results shown in Tab.3 validate that modeling curves under the adaptive axis helps to fit such scoped curves better than the original ones. Additionally, though traditional representation result is 6.2% lower than our representation, after the local predictions, the gap is narrowed to 4.0%, showing the advantage of local branch in helping with imperfect curve fittings for robust output. Corresponding visualization in Fig.4(a) also shows the validation of our global representation on flexibility.

Contributions of local shape prediction and matching. Tab.2 shows that without local shape prediction, directly outputting global curves will lower F1 by 6.0%. Changing from local shape matching to local point regressing lower F1 by 2.5%. Also, Fig.4 shows that our local shape matching relieves the influence brought by imperfect global prediction, and avoids local errors under high precision requirements.

r	16	24	32	40	48
F1	77.23	77.45	77.57	77.47	77.36

Table 4: Experimental results on RoadBEV under 10cm standard about the size of anchor cells r .

Size of anchor cells. We also experiment on the size of anchor cells r , and the results on RoadBEV under 10 cm standard are in Tab.4. With r increasing, enlarged cells cover more room for local prediction to provide better refinement for potential imperfect global prediction. However, when r is too large, it may cover multiple lines in one cell which sometimes leads to ambiguity, worsening the performance. In general, the shape of cells hardly affects results much so a fixed r is enough for FHLD to get satisfying results.

Conclusion

In this paper, we propose an end-to-end framework FHLD which fully fuses information from global and local visions to output flexible and accurate results. In particular, lane line representations in different levels are proposed, and related hierarchical shape matching strategies are designed accordingly. Experiments on two datasets validate our superiority in improving local error under high accuracy requirements. Additionally, the idea of our work can be easily extended to 2D lane detection tasks and other inputs like multi-modal.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61976170, 91648121, 62088102).

References

- Chen, S.; Zhang, Z.; Zhong, R.; Zhang, L.; Ma, H.; and Liu, L. 2020. A dense feature pyramid network-based deep learning model for road marking instance segmentation using MLS point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1): 784–800.
- Chen, Z.; Liu, Q.; and Lian, C. 2019. Pointlanenet: Efficient end-to-end cnns for accurate real-time lane detection. In *2019 IEEE intelligent vehicles symposium (IV)*, 2563–2568. IEEE.
- Feng, Z.; Guo, S.; Tan, X.; Xu, K.; Wang, M.; and Ma, L. 2022. Rethinking efficient lane detection via curve modeling. In *Computer Vision and Pattern Recognition*.
- He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask R-CNN. *CoRR*, abs/1703.06870.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385.
- Homayounfar, N.; Ma, W.-C.; Lakshmikanth, S. K.; and Urtasun, R. 2018. Hierarchical recurrent attention networks for structured online maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3417–3426.
- Homayounfar, N.; Ma, W.-C.; Liang, J.; Wu, X.; Fan, J.; and Urtasun, R. 2019. Dagmapper: Learning to map by discovering lane topology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2911–2920.
- Hou, Y.; Ma, Z.; Liu, C.; and Loy, C. C. 2019. Learning lightweight lane detection cnns by self attention distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1013–1021.
- Jeong, J.; Cho, Y.; Shin, Y.-S.; Roh, H.; and Kim, A. 2019. Complex Urban Dataset with Multi-level Sensors from Highly Diverse Urban Environments. *International Journal of Robotics Research*, 38(6): 642–657.
- Jung, J.; Che, E.; Olsen, M. J.; and Parrish, C. 2019. Efficient and robust lane marking extraction from mobile lidar point clouds. *ISPRS journal of photogrammetry and remote sensing*, 147: 1–18.
- Ko, Y.; Lee, Y.; Azam, S.; Munir, F.; Jeon, M.; and Pedrycz, W. 2021. Key points estimation and point instance segmentation approach for lane detection. *IEEE Transactions on Intelligent Transportation Systems*.
- Li, Q.; Wang, Y.; Wang, Y.; and Zhao, H. 2021. Hdmapnet: A local semantic map learning and evaluation framework. *arXiv preprint arXiv:2107.06307*.
- 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.; Yuan, Z.; Liu, T.; and Xiong, Z. 2021b. End-to-end Lane Shape Prediction with Transformers. In *WACV*.
- Ma, L.; Li, Y.; Li, J.; Yu, Y.; Junior, J. M.; Gonçalves, W. N.; and Chapman, M. A. 2020. Capsule-based networks for road marking extraction and classification from mobile LiDAR point clouds. *IEEE Transactions on Intelligent Transportation Systems*, 22(4): 1981–1995.
- Pan, Y.; Yang, B.; Li, S.; Yang, H.; Dong, Z.; and Yang, X. 2019. Automatic Road Markings Extraction, Classification and Vectorization from Mobile Laser Scanning Data. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.
- Peng, K.; Fei, J.; Yang, K.; Roitberg, A.; Zhang, J.; Bieder, F.; Heidenreich, P.; Stiller, C.; and Stiefelhagen, R. 2022. MASS: Multi-attentional semantic segmentation of LiDAR data for dense top-view understanding. *IEEE Transactions on Intelligent Transportation Systems*.
- Phillion, J. 2019. Fastdraw: Addressing the long tail of lane detection by adapting a sequential prediction network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11582–11591.
- Qin, Z.; Wang, H.; and Li, X. 2020. Ultra Fast Structure-aware Deep Lane Detection. In *The European Conference on Computer Vision (ECCV)*.
- Su, J.; Chen, C.; Zhang, K.; Luo, J.; Wei, X.; and Wei, X. 2021. Structure guided lane detection. *arXiv preprint arXiv:2105.05403*.
- Sun, T.-Y.; Tsai, S.-J.; and Chan, V. 2006. HSI color model based lane-marking detection. In *2006 IEEE intelligent transportation systems conference*, 1168–1172. IEEE.
- Tabelini, L.; Berriel, R.; Paixao, T. M.; Badue, C.; De Souza, A. F.; and Oliveira-Santos, T. 2021a. Keep your eyes on the lane: Real-time attention-guided lane detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 294–302.
- Tabelini, L.; Berriel, R.; Paixao, T. M.; Badue, C.; De Souza, A. F.; and Oliveira-Santos, T. 2021b. Polylanenet: Lane estimation via deep polynomial regression. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 6150–6156. IEEE.
- Wen, C.; Sun, X.; Li, J.; Wang, C.; Guo, Y.; and Habib, A. 2019. A deep learning framework for road marking extraction, classification and completion from mobile laser scanning point clouds. *ISPRS journal of photogrammetry and remote sensing*, 147: 178–192.
- Yan, L.; Liu, H.; Tan, J.; Li, Z.; Xie, H.; and Chen, C. 2016. Scan line based road marking extraction from mobile LiDAR point clouds. *Sensors*, 16(6): 903.
- Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; and Yan, J. 2021. Learning High-Precision Bounding Box for Rotated Object Detection via Kullback-Leibler Divergence. *CoRR*, abs/2106.01883.
- Zheng, T.; Fang, H.; Zhang, Y.; Tang, W.; Yang, Z.; Liu, H.; and Cai, D. 2021. Resa: Recurrent feature-shift aggregator for lane detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3547–3554.