A Hybrid Global-Local Perception Network for Lane Detection

Qing Chang, Yifei Tong*

School of Mechanical Engineering, Nanjing University of Science and Technology, China
qingchang@njust.edu.cn, tyf51129@aliyun.com

Abstract

Lane detection is a critical task in autonomous driving, which requires accurately predicting the complex topology of lanes in various scenarios. While previous methods of lane detection have shown success, challenges still exist, especially in scenarios where lane markings are absent. In this paper, we analyze the role of global and local features in accurately detecting lanes and propose a Hybrid Global-Local Perception Network (HGLNet) to leverage them. Global and local features play distinct roles in lane detection by respectively aiding in the detection of lane instances and the localization of corresponding lanes. HGLNet extracts global semantic context by utilizing a global extraction head that aggregates information about adaptive sampling points around lanes, achieving an optimal trade-off between performance and efficiency. Moreover, we introduce a Multi-hierarchy feature aggregator (MFA) to capture feature hierarchies in both regional and local ranges, elevating the representation of local features. The proposed Hybrid architecture can simultaneously focus on global and local features at different depth levels and efficiently integrate them to sense the global presence of lanes and accurately regress their locations. Experimental results demonstrate that our proposed method improves detection accuracy in various challenging scenarios, outperforming the state-of-the-art lane detection methods.

Introduction

Lane detection is a fundamental component of autonomous driving systems. It requires the detector to accurately determine the shape of each lane line from a front-view image obtained by a vehicle camera. The detection of lanes assists the vehicle in accurately determining its present location and the subsequent decision-making process.

Recently, deep-learning-based methods for lane detection have shown promising results (Pan et al. 2018; Zheng et al. 2021, 2022). However, there are still some challenges in detecting accurate lanes, such as faded lane lines, adverse illumination or weather situations, occlusion of lane markings by other objects like cars, or the absence of lane markings.

To address these challenges, we perform an extensive analysis of contemporary approaches to lane detection and observe three significant characteristics of lane representation: globality, locality, and consistency. When the majority of the lanes are invisible, utilizing the globality of lane representation (cyan rectangles in Fig. 1) becomes necessary for predicting the existence of the lanes by using all available visible portions. The locality refers to the inherent various features hierarchies of lanes, *e.g.*, the straight edges of lane markings, the same color, and regular shapes within regions(orange squares in Fig. 1). It contributes to correcting the position of lanes by enhancing and associating different hierarchies of features. The consistency means that many types of lanes are a smooth thin line, which is an a priori shape that is crucial for lane prediction.

Fig. 1(a) illustrates that the majority of lane instances are obscured by vehicles, but some are still discernible, thereby necessitating the detector to incorporate global context to

Figure 1: Illustrations of hard cases for lane detection. (a) The case that lane is almost occupied by the cars. The global features (the cyan rectangles) assist in determining the presence of lane lines at occluded areas. (b) Roads without lane marking. (c) The lane line is blurred under night conditions, subsequently requiring various hierarchical features, *e.g.*, the local features (edges, colors), and an extended range of region features (the orange squares), to precisely locate the lane line. (d) The case that the extreme lighting blurs lane.

*Yifei Tong is the corresponding author.

determine lanes. Detecting the absence of lanes, as shown in Fig. 1(b), poses an even more arduous task since no lane information is present in the image. In such situations, the network must furnish an ample supply of global semantic context to recognize the existence of lanes. Nonetheless, there is still a challenge in proficiently extracting global features for lanes. Prior research either constructs a message-passing mechanism to gather global context only from convolution channels (Pan et al. 2018) or designs an anchor-based attention mechanism to compute global information from local information (Tabelini et al. 2021a). These approaches have demonstrated the significance of global features for lane detection, but their global context information for the complete feature map extraction is still insufficient, resulting in poor performance in challenging scenarios.

Another common problem in lane detection is lane line blurring, *i.e.*, missing feature representation in part of the region. As shown in Fig. 1(c), the lane is blurred at night with adverse illumination, while in Fig. 1(d), the lane is hard to recognize due to the extreme lighting condition. FOLOLine (Qu et al. 2021) has demonstrated that accurate lane localization can be achieved through the regression of keypoints to create local features, highlighting the importance of local modeling. However, it did not comprehensively investigate the multi-hierarchical characteristics of local features, leaving ample room for improving the accuracy of lane detection in challenging scenarios.

Besides, the robustness of the model to adverse scenarios is facilitated by the effective integration of both local and global features. CLRNet (Zheng et al. 2022) introduces the ROIGather model that gathers both global and local features simultaneously at the same level during the parameter refinement process. Nevertheless, global features are commonly associated with high-level semantics while local features are generally represented in low-level features. Consequently, the method may fail to capture either global or local features entirely at the same level.

In this paper, we propose a novel Hybrid Global-Local Perception Network (HGLNet), which focuses on global and local features in parallel and fully integrates them. Specifically, our proposed Hybrid architecture extracts global context and models local features concurrently via two heads. To obtain the higher-level global semantic context for lanes while maintaining efficiency, we designed the global extraction head. The global extraction head selectively attends to a small set of key sampling points around a lane and feeds the sampling sequence into self-attention layers to efficiently obtain the global context for lane detection. Additionally, to address the issue of accurately detecting the position of occluded and blurred lane lines, we introduce a Multi-hierarchy feature aggregator (MFA), which explicitly and efficiently models feature hierarchies in the regional and local ranges. Our proposed method achieves state-of-the-art results on three lane detection benchmarks, *i.e.*, CULane (Pan et al. 2018), Tusimple (TuSimple 2020), and LLAMAS (Behrendt and Soussan 2019). In summary, our main contributions are:

- We demonstrate how global and local features play a significant role in lane detection, and we propose a novel approach called Hybrid Global-Local Perception Network (HGLNet) that effectively extracts the global semantic context and models the local features while fully utilizing them for lane detection.
- Our proposed global extraction head combines adaptive sampling points with the capacity to focus critical features of the attention mechanism to model global relationships, achieving a superior trade-off between performance and efficiency.
- Our proposal involves the explicit modeling of object hierarchies at both regional and local ranges using MFA. This model facilitates accurate regression of lane locations and can be integrated into other networks.

Related Work

Lane Detection Methods

Deep-learning-based lane detection methods have shown promising prospects and can be divided into five categories: segmentation-based, keypoint-based, curve-based, row-anchor-based, and line-anchor-based methods.

Segmentation-based methods. These methods treat lane detection as a per-pixel classification problem, with each pixel classified as either lane area or background (Pan et al. 2018; Zheng et al. 2021; Lu et al. 2021; Ding et al. 2020). To distinguish lanes instance, SCNN (Pan et al. 2018) obtains global features from a message-passing mechanism that helps each pixel get long-range information to complete the invisible parts of lanes. Nevertheless, the method is slow for real-time applications. LaneNet (Neven et al. 2018) adopts a different way of lane representation by casting lane detection as an instance segmentation problem. RESA (Zheng et al. 2021) optimizes the efficiency of the message-passing mechanism, and the accuracy is improved. However, these methods are ineffective and time-consuming since they perform pixel-wise prediction on the whole image and ignore specific local features.

Keypoint-based methods. Inspired by human pose estimation, some works consider lane detection as a keypoint estimation and regression problem. FOLOLane (Qu et al. 2021) proposes a bottom-up lane detection method, which estimates the presence and offset of local lane keypoints through the network to generate the final lane lines. GANet (Wang et al. 2022) proposes a feature enhancement module for lane line perception to enhance the local keypoints association of lane lines and improve the local continuity of lane lines. Despite their ability to associate local features effectively, keypoint-based methods overlook the global semantic context of feature maps, causing difficulty in detecting the existence of lanes in challenging scenarios.

Curve-based methods. Different from keypoints regression, Curve-based methods model the lane lines with curved polynomial parameters and regress these parameters to detect lanes. PolyLaneNet (Tabelini et al. 2021b) uses polynomial curves as lane representation which results in high efficiency. LSTR (Liu et al. 2021b) studies the shape characteristics of lane lines and employs the geometry constraints of lanes to enhance detection performance. Although curve-based methods regress fewer parameters

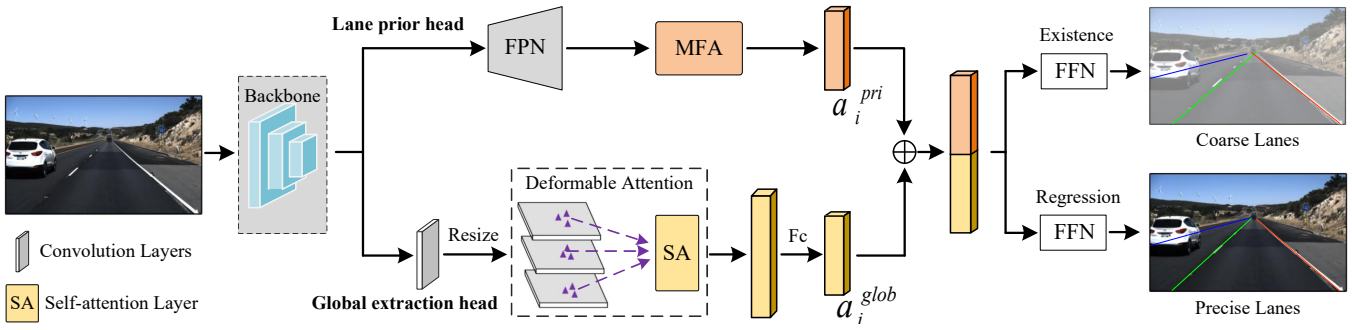

Figure 2: The overall architecture of HGLNet. An image is fed into the backbone to get the feature maps. The lane prior head generates the lane prior from multi-scale visual features generated by FPN (Lin et al. 2017), and MFA explicitly models local features. The global extraction head feeds sparse space samples across multiple feature layers into self-attention to get the global semantic features. Finally, both features are concatenated and then passed through FFN layers to predict the presence of lanes and accurately regress their positions.

and make use of the lane line continuity prior property which makes them fast, they lack the ability to detect lane lines with complicated topology.

Row-anchor-based methods. Row-anchor-based methods are defined as the identification and classification of certain on-row positions in the image. The first row-anchor-based method was proposed by UFLD (Qin, Wang, and Li 2020), which acquires global context features through a fully connected network. CondLaneNet (Liu et al. 2021a) presents a lane detection approach that integrates conditional convolution and row anchor-based formulation and addresses complex topology lane concerns through RIM. Despite the simplicity and speed of these methods, combining global and local features proves to be a difficult challenge, which results in poor overall performance.

Line-anchor-based methods. Building upon region anchors in Faster R-CNN (Ren et al. 2015), Line-CNN (Li et al. 2019) introduces a novel line anchor representation optimized for lane detection. To further improve detection accuracy, LaneATT (Tabelini et al. 2021a) combines local features with global features generated by a simplified attention module. CLRNet (Zheng et al. 2022) proposes a cross-layer refinement mechanism to utilize low-level and high-level features. Although these methods achieve high accuracy in predicting lanes by estimating offsets to line anchors, their performance heavily relies on the position of line anchors and lacks global information. To address this limitation, we propose HGLNet, which employs line anchor priors to adapt lane continuity property and introduces a global extraction head to extract global semantic context.

Deformable Attention Module

Compared with CNN, the attention mechanism is global instead of a two-dimensional locality structure and has much less image-specific inductive bias (Wen, Wang, and Hu 2023). In recent studies, researchers have proposed various attention mechanisms to broaden the boundaries of applications in computer vision (Dosovitskiy et al. 2020; Liu et al. 2021c). As a pioneer in applying Transformer to object detection, DETR (Carion et al. 2020) presents a new method that views object detection as a direct set prediction problem. However, slow convergence and high computational complexities arise because the multi-head attention module exhibits a quadratic complexity growth in relation to the feature map size. To overcome this limitation, Deformable DETR (Zhu et al. 2020) proposes the deformable attention module, which only attends to a small set of key sampling points around a reference, merging the benefits of sparse spatial sampling of deformable convolution with the relation modeling capabilities of transformers. The deformable attention module can be extended naturally for multi-scale feature maps. Inspired by PersFormer (Chen et al. 2022), we constructed a global extraction head that acquires global features by employing adaptive sampling from the resized multi-scale feature maps and the attention mechanism. This adaptive sampling strategy effectively restricts feature aggregation to the region near the slender lane structure, which significantly minimizes computational costs.

Method

The Lane Representation

Unlike conventional object detection methods that employ a rectangular bounding box to capture an object, the line anchor (Li et al. 2019) is appropriate for the representation of the lane line. Lane lines are characterized by slender shapes with well-defined shape priors, thus a predefined lane prior can help the detector better localize lanes. We denote the line anchor with prior information as a **Lane Prior**. Following (Tabelini et al. 2021a) and (Zheng et al. 2022), the lane is represented by a sequence of 2D points. Specifically, the y-coordinate of the points making up the lane boundary is uniformly sampled along the vertical image directly using the following formula: $y_i = i * \frac{H_I}{N-1}$, where H_I denotes image height. Given that the y-coordinate is fixed, a lane prior is uniquely defined by its x-coordinate, which corresponds to the associated y_i. We determine the final predicted lane's positional coordinates by adding the predicted horizontal offset to the lane prior's x coordinate. Fig. 3 illustrates the process of obtaining the final predicted

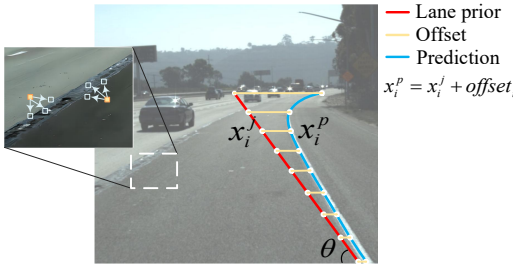

Figure 3: Illustration of the lane representation and global adaptive sampling. The predicted lane is obtained by adding prior lane and offsets, and global features are extracted from adaptive sampling points around the lane.

lane, which involves predicting the lane prior and its corresponding N offsets. The final predicted lane is represented as a sequence of points, $P = \{(x_1^p, y_1^p), \cdots, (x_N^p, y_N^p)\}$. Each lane prior consists of three parts: front and rear background probability, length, and starting point coordinate with the angle between the x-axis of the lane prior.

Global and Local Hybrid Architecture

Motivation. Object detection typically involves both the classification and localization of objects within an image or video sequence. Localization expects more local features to accurately regress the bounding box (Wang et al. 2022), while a broader global semantic context is preferred for object classification (Zhuang et al. 2023). Previous studies have highlighted the significance of integrating both global and local features in lane detection. LaneATT (Tabelini et al. 2021a) calculates the attention between individual feature vectors and combines them to gather both global and local information. However, this method fails to effectively extract the crucial global semantic context from the feature map. In our work, we employ a multi-head mechanism that focuses on global and local features individually.

Lane Prior Head Structure. Inspired by CLRNet (Zheng et al. 2022), the lane prior head utilizes a feature pyramidal hierarchy (FPN) for cross-layer refinement of the lane prior. Specifically, the features extracted by the backbone are transmitted to the FPN neck to derive three levels of feature maps. FPN output layers are set to 3 based on considerations of inference cost in multiple experiments. Then, we employ MFA to enhance each layer of features, resulting in $\{\mathbf{F_0}, \mathbf{F_1}, \mathbf{F_2}\}$, all of them with C_p channels. MFA is utilized to further reinforce the association between features within a local range at each level of the feature map. The detector then learns the lane prior in a top-down manner, starting from the highest layer $\mathbf{F_2}$, and acquires the parameters of the final lane prior (including the start point coordinates x and y, as well as the angle θ) at the lowest layer $\mathbf{F_0}$. Each final lane prior i will possess a corresponding feature vector $\boldsymbol{a}_i^{pri} \in \mathbb{R}^{C_p}$, carrying local feature information.

Global Extraction Head Structure. Both depth-based (Yu et al. 2018) and transformer-based methods (Vaswani et al. 2017) contribute to exploring global semantics. However, the ability of depth-based networks to extract adaptable global features from slender lanes is limited due to the fixed geometric structure and sample points in the convolution kernel. Additionally, the substantial computational burden imposed by transformer-based methods is unsuitable for deployment on autonomous driving chips. Inspired by Deformable DETR (Zhu et al. 2020), we designed a global extraction head that combines adaptive sampling points with the capacity to focus on the essential features of the attention to enable adaptive global feature extraction from slender objects, *i.e.*, lanes.

Firstly, a 1×1 convolution is applied to active feature map levels generated by the backbone, and the feature maps are resized to the same size 30×30 to save memory. Let $\{\boldsymbol{x}^l\}_{l=1}^L$ be the input feature map levels, where $\boldsymbol{x}^l \in \mathbb{R}^{C_g \times H_g \times W_g}$. To each layer, we assign K reference points and define $\boldsymbol{p}_q \in [0,1]^2$ as the normalized coordinates of the reference point for each query element q. The sampling offset of each reference point at each layer can be computed as follows:

$$d = \left(R_l\left(\boldsymbol{p}_q\right) + \triangle \boldsymbol{p}_{mlqk}\right) \tag{1}$$

where $\triangle \boldsymbol{p}_{mlqk}$ indexes the sampling offset of the k^{th} sampling point in the l^{th} feature level and the m^{th} attention head. The function $R_l\left(\boldsymbol{p}_q\right)$ re-scales the normalized coordinates $\boldsymbol{p}_q$ to the input feature map of the l^{th} level. Subsequently, we compute the global features as follows:

$$\boldsymbol{A}^{glob} = \varphi\left(\sum_{m=1}^{M} \boldsymbol{W}_m \left[\sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlqk} \boldsymbol{W}'_m \phi\left(x^l; d\right)\right]\right) \tag{2}$$

where $\boldsymbol{W}'_m \in \mathbb{R}^{C_v \times C_g}$ and $\boldsymbol{W}_m \in \mathbb{R}^{C_g \times C_v}$ are of learnable weights ($C_v = C_g/M$). A_{mlqk} denotes the attention weight of each reference point. The function $\phi(\cdot;\cdot)$ denotes a bilinear interpolation to calculate the interpolated feature for each reference point with sampling offset d. φ denotes a fully connected layer, which is used to activate the information obtained by sampling. The matrix $\boldsymbol{A}^{glob} = \left[\boldsymbol{a}_0^{glob}, ..., \boldsymbol{a}_{N_p-1}^{glob}\right]^T$ contains the global feature vectors, and N_p denotes the number of lane priors. Since lanes have a slender structure and occupy only a small image area, the global extraction head only focuses on a limited number of reference points around a lane line, as shown in Fig. 3.

Multi-Hierarchy Feature Aggregator

Motivation. Lanes typically have multiple hierarchies of characteristics. Recent studies (Wang et al. 2018; Li et al. 2023) have shown that the local hierarchy range exhibits characteristic features, such as edges and colors, that span several pixels. In contrast, the regional range is identified using a tens-of-pixel window size. Lane markings may become partially obscured or blurred. To achieve precise position detection, the lane detector should increase the detection window size to include adjacent pixels in the case of partial occlusion (Chen et al. 2021). Conversely, for blurred lane markings, local features (*e.g.*, lane edges) should be emphasized. As a result, we propose a Multi-hierarchy feature

aggregator (MFA) module to explicitly and efficiently model feature hierarchies in the regional and local ranges.

MFA structure. The illustration of the MFA structure is shown in Fig. 4(a). Dilated Convolution (Yu and Koltun 2015) efficiently expands the receptive field while minimizing the number of parameters. A greater receptive field can capture more context information and enhance the hierarchy of regional range. Let F_n be the input feature map and the operation is $F'_n = \mathrm{SiLU}\left[D\left(F_n\right)\right]$, where D is the dilated convolution operation and SiLU denotes SiLU activation function. In order to perceive vital information in the context, we then feed F_n and F'_n into the Context-aware Attention, shown in Fig. 4(b), which integrates two different receptive fields obtained using the dilated convolution and typical convolution, respectively. This integration is achieved by employing the long-range dependencies method introduced in the non-local (Wang et al. 2018) operation to effectively exploit both local and regional information. The formula for Context-aware Attention is as follows:

$$F''_n = \mathrm{ConAttn}\left(\theta\left(F_n\left(x_i\right)\right), \phi\left(F'_n\left(x_j\right)\right), g\left(F'_n\left(x_j\right)\right)\right) \tag{3}$$

where $\mathrm{ConAttn}$ indexes that the Context-aware Attention and the resulting F''_n is obtained from the weighted sum of F_n. $F_n\left(x_i\right)$ indexes the value at position i of feature map F_n, and $F'_n\left(x_j\right)$ indexes the value at position j of F'_n. θ, ϕ, and g denote 1×1 convolution operations, which are the weight matrices to be learned.

Then, Channel Attention (CA) is used to transmit the spatial attention feature map generated by Context-aware Attention. Inspired by (Hu, Shen, and Sun 2018), we squeeze the feature map size by max-pooling and average-pooling, then obtain the weights on each channel of the feature map by a fully connected layer. The formula is defined as follows:

$$\boldsymbol{W}_c = \sigma\left(\varphi_1\left(\mathrm{Avg}\, P_{ch}\left(F''_n\right)\right) + \varphi_2\left(\mathrm{Max}\, P_{ch}\left(F''_n\right)\right)\right) \tag{4}$$

where $\mathrm{Avg}\, P_{ch}$ and $\mathrm{Max}\, P_{ch}$ denote max-pooling and average-pooling, respectively. φ_1 and φ_2 denote fully connection layers and σ indicates the sigmoid function. The $\boldsymbol{W}_c$ is the channel-wise weights.

Then the channel-wise weights are applied to the feature map F''_n via element-wise multiplication, helping to highlight the features that contribute most to the representation of slender objects (Wen, Wang, and Hu 2023), *i.e.*, the lane lines. Finally, the new features are concatenated with the original features on the channel axis, and then the dimension is reduced to obtain F'''_n. The final aggregated feature is computed as:

$$F'''_n = \Phi\left(F_n \oplus \left(\boldsymbol{W}_c \otimes F''_n\right)\right) \tag{5}$$

where $\oplus$ indexes the concatenate operation and the symbol $\otimes$ indexes the element-wise multiplication on each channel. Φ indexes 1×1 convolution to reduce the dimensionality. The size of F'''_n is consistent with F_n, and it reflects the bonus of 3-D weights to F_n and strengthens the connection between different receptive fields.

The total loss consists of classification loss and regression loss:

$$\mathcal{L}_{total} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{aux}\mathcal{L}_{aux} \tag{6}$$

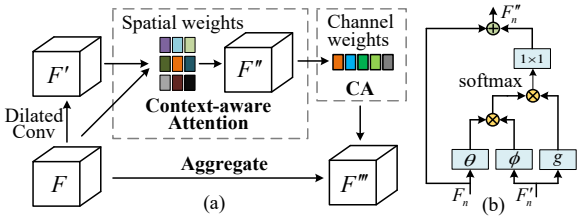

Figure 4: Illustration of MFA module. (a) MFA has a straightforward structure and learns the weights of the spatial and channel directions of the feature map to form 3D weights. (b) Context-aware Attention. The input is two different receptive field feature layers.

where $\mathcal{L}_{cls}$ is the focal loss between predictions and ground truth. $\mathcal{L}_{reg}$ is the smooth-l_1 loss for the start point coordinate, angle, and lane length regression. $\mathcal{L}_{aux}$ is the auxiliary segmentation loss following (Qin, Wang, and Li 2020), which is only used in the training phase and has no cost in inference. The weights λ_{cls}, λ_{reg}, λ_{aux} are determined by experiments respectively.

Experimental

Datasets and Evaluation Metrics

We conduct experiments on three widely used benchmarks: CULane (Pan et al. 2018), LLAMAS (Behrendt and Soussan 2019), and TuSimple (TuSimple 2020).

CULane: CULane is a large-scale lane detection dataset with 88k training images and 34k testing images. The test images are classified into nine scenarios: crowded, night, no line, etc. These scenario tests can reflect the robustness of the model. All the images have 1640×590 pixels.

LLAMAS: LLAMAS is a large-scale highway lane detection dataset with over 100k images. Its test set's label is not public, so the testing result will be given after uploading results to their website. For CULane and LLAMAS, We adopt the F1-measure as an evaluation metric, which is based on Intersection-over-union (IoU). The F_1 is defined as:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

where $Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}$. A predicted lane whose IoU is greater than 0.5 is judged as true positive (TP), otherwise false positive (FP), or false negative (FN). To better compare the positioning performance of the algorithms (Zheng et al. 2022), we use the mF1 metric following COCO (Lin et al. 2014). The metric mF1 reports the average F1 score across multiple IoU thresholds, *i.e.*, IoU threshold $= 0.5, 0.55, \cdots, 0.95$. The mF1 is defined as:

$$\mathrm{mF1} = (\mathrm{F1@50} + \mathrm{F1@55} + \cdots + \mathrm{F1@95})/10 \tag{8}$$

TuSimple: TuSimple is a highway dataset comprising $3,626$ training images and $2,782$ testing images. Its image resolution is 720×1280. For the Tusimple dataset, the evaluation formula is accuracy, which is formulated as follows:

$$Accuracy = \frac{\sum_{clip} C_{clip}}{\sum_{clip} S_{clip}} \tag{9}$$

Method	mF1	F1	Normal	Crowd	Dazzle	Shadow	No line	Arrow	Curve	Cross	Night	FPS
SCNN(VGG16)	38.84	71.60	90.60	69.70	58.50	66.90	43.40	84.10	64.40	1990	66.10	7.5
RESA(ResNet-34)	-	74.50	91.90	72.40	66.50	72.00	46.30	88.10	68.60	1896	69.80	45.5
RESA(ResNet-50)	47.86	75.30	92.10	73.10	69.20	72.80	47.70	88.30	70.30	1503	69.90	35.7
UFLD(ResNet-18)	38.94	68.40	87.70	66.00	58.40	62.80	40.20	81.00	57.90	1743	62.10	**282**
UFLD(ResNet-34)	-	72.30	90.70	70.20	59.50	69.30	44.40	85.70	69.50	2037	66.70	170
LaneATT(ResNet-122)	51.48	77.02	91.74	76.16	69.47	76.31	50.46	86.29	64.05	1264	70.81	20
LaneAF(ERFNet)	48.60	75.63	91.10	73.32	69.71	75.81	50.62	86.86	65.02	1844	70.90	24
LaneAF(DLA-34)	50.42	77.41	91.80	75.61	71.78	79.12	51.38	86.88	72.70	1360	73.03	20
CLRNet(ResNet-18)	55.23	79.58	93.30	78.33	73.71	79.66	53.14	90.25	71.56	1321	75.11	119
CLRNet(DLA-34)	55.64	80.47	93.73	79.59	75.30	82.51	54.58	**90.62**	74.13	1155	75.37	94
HGLNet(ResNet-18)	55.83	80.65	93.48	78.31	75.13	81.75	53.74	89.98	73.27	**959**	75.06	116
HGLNet(ResNet-34)	56.07	81.23	93.76	78.89	75.29	82.21	54.95	90.43	74.95	1023	75.47	133
HGLNet(ResNet-101)	56.24	81.40	93.74	**79.91**	75.81	**83.34**	55.61	90.78	75.65	1240	76.01	58
HGLNet(DLA-34)	**56.63**	**81.83**	**93.96**	79.78	**76.20**	83.27	**55.89**	**90.83**	**75.77**	1208	**76.44**	104

Table 1: Comparison with popular methods on the CULane test set. The evaluation metric for all scenarios is F1 score with IoU threshold=0.5. For the Cross scenario, only false positives are shown. FPS is measured based on the Pytorch framework.

Method	Valid		Test		
	mF1	F1	F1	Pre	Rec
PolyLaneNet(Eff-B0)	48.82	90.20	88.40	88.87	87.93
LaneATT(ResNet-18)	69.22	94.64	93.46	96.92	90.24
LaneATT(ResNet-34)	69.63	94.96	93.74	96.79	90.88
LaneATT(ResNet-122)	70.80	95.17	93.54	96.82	90.47
HGLNet(ResNet-18)	71.46	96.74	95.99	96.72	95.27
HGLNet(DLA-34)	**71.66**	**97.98**	**96.20**	**97.01**	**95.41**

Table 2: Comparison with popular methods on LLAMAS.

where C_{clip} is the number of correct points and S_{clip} is the number of ground truth points of an image. A predicted point is considered correct only if it is within 20 pixels of the ground truth point. We also calculate the F1 score for Tusimple, and the predicted lane with accuracy greater than 85% is considered a true positive.

Implementation Details

We adopt the ResNet (He et al. 2016) and DLA (Yu et al. 2018) as our pre-trained backbones. All input images are resized to 320×800 during the training and testing phases. Similar to (Qu et al. 2021; Zheng et al. 2022), we perform data augmentation using methods including random affine transformation (translation, rotation, and scaling) and random horizontal flips. In the training process, we use AdamW (Loshchilov and Hutter 2017) optimizer with an initial learning rate of 1e-3 and cosine decay learning rate strategy (Loshchilov and Hutter 2016) with power set to 0.9. For the lane prior head, we set the number of lane prior proposals $N_p = 192$, proposals' feature dimension $C_p = 64$, and the number of points of each lane prior $N = 72$. For the global extraction head, the resized H_g, W_g are $10, 25$, respectively, the channel $C_g = 192$ and reference points $K = 4$. The dilation rate in MHA is set as $d = 2$. The training numbers of epochs for CULane, Tusimple, and LLAMAS are 15, 80, and 20. Training and testing are both performed on Pytorch with one Tesla-V100 GPU.

Method	F1	Acc	FP	FN
SCNN(VGG16)	95.97	96.53	6.17	1.80
RESA(ResNet-34)	96.93	96.82	3.63	2.48
UFLD(ResNet-34)	88.02	95.86	18.91	3.75
PolyLaneNet(Eff-B0)	90.62	93.36	9.42	9.33
LaneATT(ResNet-122)	96.06	96.10	5.64	2.17
CondLaneNet(ResNet-101)	97.24	96.54	2.01	3.50
HGLNet(ResNet-18)	97.72	**96.89**	3.38	**1.55**
HGLNet(ResNet-34)	97.71	96.68	1.93	2.65
HGLNet(ResNet-101)	**97.82**	96.74	**1.81**	2.57

Table 3: Comparison with popular methods on TuSimple.

Quantiative Results

Performance on CULane. As illustrated in Table 1, we show the results of our method on the CULane test set and compare them with other popular lane detection methods. Our proposed method achieves state-of-the-art results on CULane with an 81.83% F1 measure and 104 FPS. Our HGLNet delivers superior performance and a higher efficiency trade-off than the DLA34 version of CLRNet, achieving better performance and operating at 104 FPS as opposed to 94 FPS. Notably, HGLNet has improved detection accuracy for several challenging scenarios, *e.g.*, Shadow, No line, and Night. The ResNet18 version of our method achieves 75.13% F1 on the Dazzle scenario, which is even higher than getting 2.64 points higher than CLRNet (ResNet101) while getting 4.41 points higher than CondLaneNet (ResNet18). In particular, when using the same backbone, our method improves over CLRNet in mF1 metrics. This indicates our method regresses lanes with high localization accuracy.

We show the qualitative results for the four scenarios on the CULane dataset in Fig. 5. In difficult scenes, the segmentation-based methods, such as UFLD and RESA, predict the lanes for each pixel, causing the predicted lane to lose its own characteristics. By effectively extracting global features, our method predicts smoother curved lines and does not miss some lane instances. In low-light and crowded

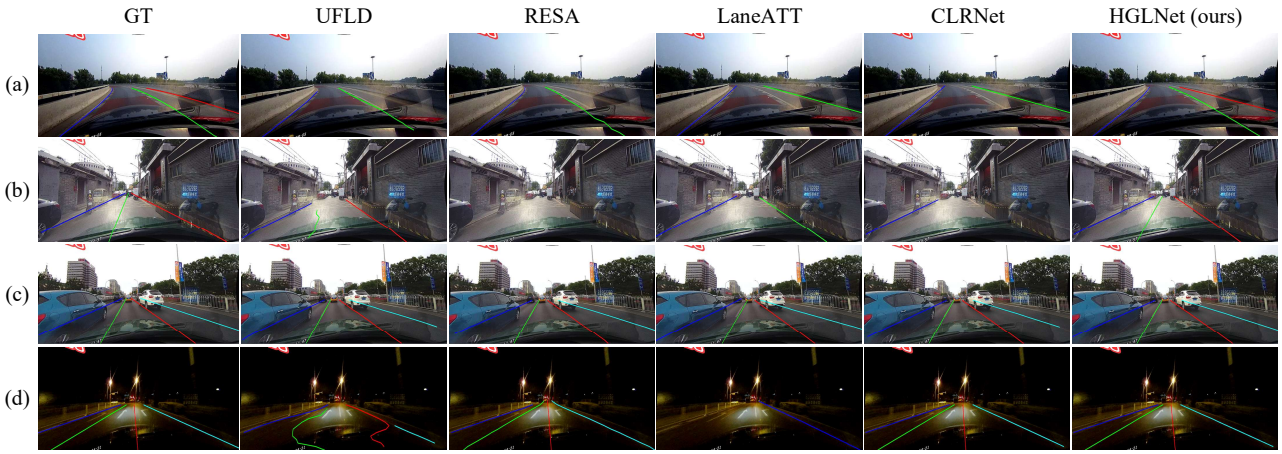

Figure 5: Visualization of different scenario results of UFLD, RESA, LaneATT, CLRNet, and our method on the CULane test set: (a) Curve, (b) No line, (c) Crowded, (d) Night.

scenarios, HGLNet predicts the location of the lanes more accurately. This result shows that our method can enhance the local features of lanes by linking the feature hierarchy.

Performance on LLAMAS. The result on the LLAMAS is shown in Table 2. Since the dataset test labels are confidential, we show both valid and test results to better demonstrate the performance of the different methods. Our method achieves a new state-of-the-art on LLAMAS test set with a 96.20% F1 measure and outperforms PolyLaneNet and LaneATT by 7.8% F1 and 2.66% F1 respectively, which is a significant improvement. Meanwhile, our method has the highest Precision and Recall score and outperforms LaneATT by a 4.94% Recall score, which indicates that our model effectively extracts global features to discriminate lane information. Moreover, our method achieves a 71.66% mF1 measure in the valid set, which further demonstrates that our method can better locate the lane position precisely.

Performance on Tusimple. As shown in Table 3, the difference in performance between the different methods is tiny, which indicates that the accuracy of the dataset seems to be already saturated. Despite this, our method achieves a new start-of-the-art with a 96.89% Accuracy score and surpasses the previous SOTA with a 0.25% FN score. In the meantime, different versions of our method all have lower FP and FN, which firmly demonstrates that HGLNet can validly predict the presence of lane lines even in complex scenarios.

Ablation Study

To verify the effectiveness of the proposed components, we report the overall ablation studies in Table 4. All experiments of the ablation study are based on the ResNet-18 version of HGLNet. We gradually add MFA and Hybrid architecture on the ResNet18 baseline and demonstrate the F1 measure for some of the challenge scenarios.

Ablation study on MFA. To further demonstrate the effect of MHA on lane detection, we performed ablation experiments on MHA. MHA improves the mF1 measure from 52.84% to 54.58% and greatly improves performance in blurred lane lines such as Shadow and Dazzle scenarios. The results reinforce that expanding the local and regional hierarchies of feature maps enables to overcome of lane ambiguity and improves localization accuracy.

Baseline	MFA	Hybrid	mF1	F1	No line	Shadow
✓			52.84	77.54	50.16	79.30
✓	✓		54.58	78.25	52.90	80.64
✓	✓	✓	**55.83**	**80.65**	**53.74**	**81.75**

Table 4: Effects of each component in our method. Results are reported on CULane.

Ablation study on Hybrid architecture. Ablation studies of Hybrid architecture are shown in Table 4. The proposed hybrid architecture is straightforward and can be easily transferred to different networks. With the introduction of the hybrid architecture, the network has been greatly improved, improving the performance in difficult scenarios. The results demonstrate the advantage of our Hybrid architecture to sense the global presence of lane lines and accurately regress their locations. Notably, the proposed Hybrid architecture can (i) efficiently focus global features and local features in parallel, (ii) fully integrate the two features to judge the existence of lane lines and correct their positions.

Conclusion

In this paper, we propose a Hybrid Global-Local Perception Network (HGLNet) that performs efficient parallel extraction of global context and local feature modeling for lane detection. Our analysis focuses on the roles played by globality, locality, and consistency in lane representation. To address the absence of visual evidence for lane presence, we propose using a global extraction head to extract global context from multi-layer feature maps. To more accurately regress the position of the lane, we propose MFA to enhance the hierarchy of local feature ranges. Our proposed method has been demonstrated to improve lane detection accuracy and outperforms current state-of-the-art methods in multiple challenging scenarios.

Acknowledgments

This work was financially supported by the Key R&D Program of Jiangsu Province (BE2023352). The supports are gratefully acknowledged.

References

Behrendt, K.; and Soussan, R. 2019. Unsupervised labeled lane markers using maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

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 *European Conference on Computer Vision*, 550–567. Springer.

Chen, Z.; Yang, C.; Li, Q.; Zhao, F.; Zha, Z.-J.; and Wu, F. 2021. Disentangle your dense object detector. In *Proceedings of the 29th ACM international conference on multimedia*, 4939–4948.

Ding, L.; Zhang, H.; Xiao, J.; Shu, C.; and Lu, S. 2020. A lane detection method based on semantic segmentation. *Computer Modeling in Engineering & Sciences*, 122(3): 1039–1053.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Li, X.; Li, J.; Hu, X.; and Yang, J. 2019. Line-cnn: End-to-end traffic line detection with line proposal unit. *IEEE Transactions on Intelligent Transportation Systems*, 21(1): 248–258.

Li, Y.; Fan, Y.; Xiang, X.; Demandolx, D.; Ranjan, R.; Timofte, R.; and Van Gool, L. 2023. Efficient and Explicit Modelling of Image Hierarchies for Image Restoration. *arXiv preprint arXiv:2303.00748*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

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 *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3694–3702.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021c. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, S.; Luo, Z.; Gao, F.; Liu, M.; Chang, K.; and Piao, C. 2021. A fast and robust lane detection method based on semantic segmentation and optical flow estimation. *Sensors*, 21(2): 400.

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 Artificial Intelligence*, volume 32.

Qin, Z.; Wang, H.; and Li, X. 2020. Ultra fast structure-aware deep lane detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, 276–291. Springer.

Qu, Z.; Jin, H.; Zhou, Y.; Yang, Z.; and Zhang, W. 2021. Focus on local: Detecting lane marker from bottom up via key point. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14122–14130.

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.

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.

TuSimple. 2020. Tusimple benchmark. https://github.com/TuSimple/tusimple-benchmark/.

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, J.; Ma, Y.; Huang, S.; Hui, T.; Wang, F.; Qian, C.; and Zhang, T. 2022. A keypoint-based global association network for lane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1392–1401.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wen, F.; Wang, M.; and Hu, X. 2023. DFAM-DETR: Deformable feature based attention mechanism DETR on slender object detection. *IEICE TRANSACTIONS on Information and Systems*, 106(3): 401–409.

Yu, F.; and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Yu, F.; Wang, D.; Shelhamer, E.; and Darrell, T. 2018. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2403–2412.

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.

Zheng, T.; Huang, Y.; Liu, Y.; Tang, W.; Yang, Z.; Cai, D.; and He, X. 2022. Clrnet: Cross layer refinement network for lane detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 898–907.

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

Zhuang, J.; Qin, Z.; Yu, H.; and Chen, X. 2023. Task-Specific Context Decoupling for Object Detection. *arXiv preprint arXiv:2303.01047*.