PriorLane: A Prior Knowledge Enhanced Lane Detection Approach Based on Transformer

Qibo Qiu[1,2], Haiming Gao[1], Wei Hua[1†], Gang Huang[1] and Xiaofei He[2], *Senior Member, IEEE*

Abstract— **Lane detection is one of the fundamental modules in self-driving. In this paper we employ a transformer-only method for lane detection, thus it could benefit from the blooming development of fully vision transformer and achieves the state-of-the-art (SOTA) performance on both CULane and TuSimple benchmarks, by fine-tuning the weight fully pre-trained on large datasets. More importantly, this paper proposes a novel and general framework called PriorLane, which is used to enhance the segmentation performance of the fully vision transformer by introducing the low-cost local prior knowledge. PriorLane utilizes an encoder-only transformer to fuse the feature extracted by a pre-trained segmentation model with prior knowledge embeddings. Note that a Knowledge Embedding Alignment (KEA) module is adapted to enhance the fusion performance by aligning the knowledge embedding. Extensive experiments on our Zjlab dataset show that PriorLane outperforms SOTA lane detection methods by a** 2.82% **mIoU, and the code will be released at:** `https://github.com/vincentqqb/PriorLane`.

Key words–Lane detection, Vision transformer, Prior knowledge, Fusion, Knowledge Alignment

I. INTRODUCTION

Lane detection is one of the most essential tasks of perception for both self-driving and Advanced Driver Assistant System (ADAS), which provides information for localization and planning [1]. Many traditional lane detection methods are designed based on hand-crafted features [2], [3]. With a blooming development of deep learning, lane detection comes towards a new age with the higher robustness and the stronger generalization, because the lane detection task depends on texture features and high-level semantic analysis, and it benefits a lot from the representation ability of deep learning based models.

In order to carry out the efficient lane detection task, the segmentation based method, as one of the most typical methods in the field of lane detection, has attracted much attention [4], [5], while there are several methods designed to take advantage of the strong shape prior of lane markings, e.g., anchor-based methods [6], [7], row-wise detection methods [8], [9], parametric prediction methods [10], [11].

Although those aforementioned methods have made a great progress in the field of autonomous vehicles, the research on robust lane detection is still far from mature.

[1]Zhejiang Lab, Hangzhou, P. R. China, 311121 (Email: qiuqibo.qqb@zhejianglab.com, huawei@zhejianglab.com).

[2]State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou, P. R. China, 310058.

†Corresponding author.

Especially, severe occlusions caused by vehicles and pedestrians are frequently occurring, and lane markings in the distance are always too fuzzy to catch the feature in deep models.

In this paper, a pure vision transformer with a hierarchical encoder is employed to segment lane markings in images, which is called Mixed Transformer (MiT) [12], because the transformer [13] is designed to catch the long-term dependency, and recently it plays a more and more important role in vision tasks. In addition, the MiT blocks could be fully pre-trained on large scale datasets, since the transformer-only architecture used in lane detection and general segmentation are unified in an elegant way.

Furthermore, to enhance the segmentation performance, a general fusion framework called PriorLane is proposed to leverage the prior knowledge to improve the performance of MiT blocks. The local prior knowledge data is represented by a grid map in Bird's Eye View (BEV), then PriorLane cuts the grid map into patches and maps them with a trainable linear projection, which are known as embeddings. And a corresponding Knowledge Embedding Alignment (KEA) module is designed to align the embedding spatially.

Finally, a novel fusion transformer (FT) is used to fuse the image feature extracted via MiT blocks with embeddings of the prior knowledge data. Similar to the segformer [12], an Multi-Layer Perceptron (MLP) block is adapted to merge the fused feature with pure MiT features, and outputs pixel-level results of segmentation prediction.

The contributions of this paper are concluded as follows:

- To our knowledge, the transformer-only architecture is employed to perform lane detection for the first time, and it could benefit from the development of fully vision transformer pre-training, which is supported by our experiments on CULane and TuSimple benchmarks.
- More importantly, a novel and general framework called PriorLane is proposed, which enhances the performance of lane segmentation via fusing the image feature with the low-cost local prior knowledge and obtains the better performance by comparative experiments.
- The KEA block is further employed for the prior knowledge alignment, and experiments on our Zjlab dataset show that our framework achieves SOTA performance, e.g., an 73.78% mIoU (2.82% higher than SOTA).

II. RELATED WORK

A. Segmentation-based Lane Detection

Different from general segmentation tasks, the structure of lane markings is thin and long, thus it is easy to lose

the valid information during feature extracting. To tackle this issue, SCNN [14] proposes a sequential message pass scheme, which is useful for structured objects segmentation. While RESA [15] aggregates global information by shifting the sliced feature map recurrently, and achieves the more efficient performance. Although, segmentation as a kind of pixel-level prediction task suffers from high computation cost and is inflexible for motion planning, it still plays an important role in some applications, e.g., High-Definition map construction and localization [16], [17].

B. Query-based Object Detection

DETR [18] is the first query-based object detection work which makes use of transformer to decode predictions end-to-end, thus there is no post process, e.g., Non-Maximum Suppression (NMS). In addition, LSTR [10] achieves a SOTA performance on TuSimple benchmark by regressing the lane line parameters using a Hungarian loss, which is similar to DETR, however the ResNet backbone and transformer module could not be pre-trained simultaneously.

To mitigate the problem, FP-DETR [19] treats the query as a kind of implicit prompts, thus both the backbone and attention layers could be pre-trained simultaneously for object detection. Different from FP-DETR, our PriorLane framework makes use of prompts to fine-tune the segmentation model, in particularly, the prompt is explicit and no manually annotated labels are used to train it.

C. Multi-View Transformers

BEVFormer [20] use the query initialized in BEV to merge the feature from multi camera views and predict corresponding 3D bounding boxes. [21] also uses the attention mechanism to map the camera view into the BEV representation. Aforementioned approaches depend on intrinsic and extrinsic calibration of cameras to build the relationship between BEV and camera views, moreover the BEV representation is usually initialized randomly and can be learned during training. The query used in the proposed PriorLane framework has a physical meaning, and corresponding intrinsic and extrinsic

parameters are not needed, while high precision positions and orientations are expensive to acquire in real applications.

III. METHOD

A. Architecture Design

Fig. 1 shows the overall architecture of the proposed PriorLane framework, it mainly consists of three successive parts: the MiT block, the KEA module and the FT block.

Mix Transformer. [12] designs a series of MiT blocks, and MiT-B5 is utilized in this paper, which is the largest version of MiT blocks and has the best performance on segmentation. Different from typical ViT [22], the advantage of MiT could generate hierarchical features, due to the design of overlapped patch merging and efficient self-attention.

Knowledge Embedding Alignment. Since the vehicle positions are coarse, it is hard to determine the reference knowledge embedding in BEV consistent with the camera view. To address this issue, a Knowledge Embedding Alignment(KEA) module is needed to align the knowledge embedding with image features spatially.

Fusion Transformer. The knowledge encoder layer and the fusion encoder layer work together as the fusion transformer. Different from FP-DETR, which proposes a novel approach for detection transformer pre-training by taking advantage of implicit vision prompts. In this paper, prompts in PriorLane are generated by explicit prior knowledge, moreover, no annotated labels are used for prompts training.

Architecture on benchmarks. Typical lane segmentation methods on CULane and TuSimple benchmarks leverage the annotation of lane existence, e.g., RESA, SCNN and UFLD [8]. To verify the performance of fully vision transformer on CULane and TuSimple benchmarks, the MiT-Lane is employed, which is an intuitive extension of segformer by adding a bench for existence prediction following the MLP block, as shown in Fig. 2.

B. Local Prior Knowledge Representation

Road network, elevation map, etc., are known as the prior knowledge of traffic scenes. In this paper, the prior knowledge is stored in BEV by a grid map, and Fig. 3 shows
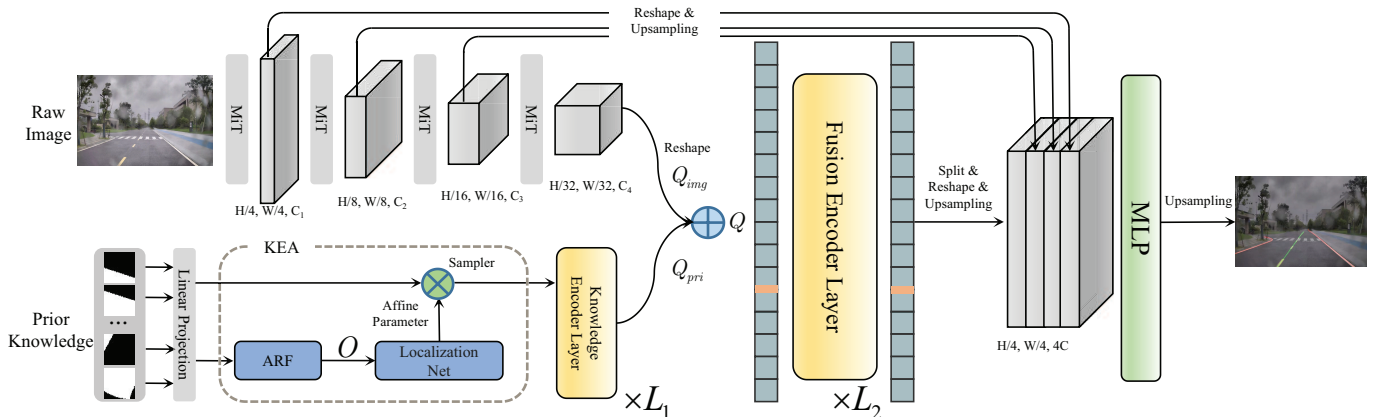

Fig. 1. Overview of the proposed PriorLane framework. L_1 and L_2 indicates the number of knowledge encoder layers and fusion encoder layers, respectively. O is the orientation-invariant feature extracted by ARF.
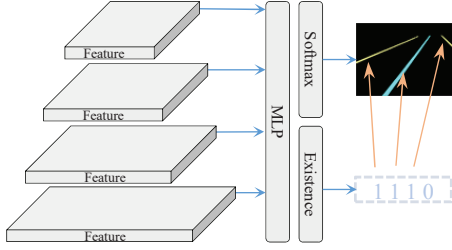
Fig. 2. The segmentation architecture named MiT-Lane on CULane and TuSimple benchmarks.

a grid map with one channel, while 0/1 indicates whether the corresponding grid is in the road freespace or not. More importantly, other kinds of prior knowledge could also be leveraged by adding additional channels. In this way the prior knowledge is rendered into a "big image" and the local prior knowledge could be represented as a smaller one once the perception range is given.

The vehicle positions are collected by a meter-level equipment, and the orientation of the vehicle is randomly rotated to simulate a coarse position, since it is high-cost to obtain the accurate position in an ubran environment. And PriorLane is proposed to enhance the lane segmentation performance without high-definition prior knowledge. Once the resolution, the perception range and the position are given, the sub-image from the "big image" could be obtained, which is named as "prior knowledge" hereinafter.

Fig. 3. The prior knowledge of Zhejiang Lab park represented as a grid map.

Knowledge Embedding. Due to the powerful ability of feature representation by transformer, only linear projection is needed for knowledge embedding. Firstly, we cut the local grid map data $M \in \mathbb{R}^{H \times W \times C}$ into patches and the size of each patch is $P \times P$. Moreover, each patch is converted into a vector of size $1 \times E_p$ by a trainable linear projection, thus the local prior knowledge is represented as the "knowledge embedding", which is denoted by X and the corresponding size is $(H/P) \times (W/P) \times E_p$.

C. Knowledge Embedding Alignment

In this section, Spatial Transformer Network(STN) [23] is adapted to align the embedding in the KEA module. STN could manipulate the prior knowledge embedding according to the affine transform matrix obtained by a localization network, thus the fusion performance of the prior knowledge and the image feature could be enhanced.

Traditional CNN lacks the capability to extract direction information, which is crucial for STN. Active Rotation Filters (ARFs) and ORPooling [24], [25] are adapted to extract orientation information from the prior knowledge embeddings and keep the rotation-invariant information, respectively. An ARF R is made up of N canonical filters, whose size is $N \times K \times K$, and N means rotating the kernel of each filter for $N - 1$ times, the degree is $\frac{2\pi}{N}$ per time. For a given feature map X, the output of an ARF is F denoted as

$$F^{(i)} = \sum_{n=0}^{N-1} R_{\theta_i}^{(n)} \cdot X^{(n)}, \theta_i = i\frac{2\pi}{N}, i = 0, \ldots, N - 1, \quad (1)$$

where $F^{(i)}$ is the i-th orientation of F, and R_{θ_i} means rotating R by a degree of $\theta_i = i\frac{2\pi}{N}$, $R_{\theta_i}^{(n)}$ indicates the n-th channel of R_{θ_i} and $X^{(n)}$ is the n-th orientation channel of the corresponding feature map X. We pool the orientation-sensitive feature F and keep rotation-invariant information O as follows

$$O = max \ F^{(n)}, 0 \le n < N, \quad (2)$$

where $F^{(n)}$ indicates the n-th orientation channel of F. The rotation-invariant feature O is only used for projection parameters estimation, and the output after affine transformation is sampled from the original embedding of the local prior knowledge X, as shown in Fig. 1.

D. Fusion Transformer Layers

As shown in Fig. 1, PriorLane uses knowledge encoder layers with self-attention to refine the embedding aligned by the KEA module. The output of knowledge encoder layers can be treated as a kind of prompts for fusion encoder layers, and inter-prompt relationships can be learned in the fusion encoder layer. As described in [26], the cross-attention usually results in a slow convergence, therefore self-attention is utilized to fuse the prior knowledge embedding with image features in fusion encoder layers, and it is calculated as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (3)$$

where

$$Q = K = V = [Q_{img}; Q_{pri}]. \quad (4)$$

Q is the concatenation of Q_{img} and Q_{pri}, and Q_{pri} is the output of the knowledge encoder layer with the size of $L \times E_p$, while $L = (H/P) \times (W/P)$. Considering that images are collected by different cameras, which have different resolutions, intrinsic parameters and extrinsic parameters, position encoding is not employed in attention calculation.

IV. EXPERIMENTS

A. Datasets

In this paper, we firstly conduct experiments on two famous benchmarks: CULane [14] and TuSimple[1]. Since there is no public dataset contains images, labelled lane markings and prior knowledge, a dataset is collected during

---

[1]https://github.com/TuSimple/tusimple-benchmark/

TABLE I
LANE DETECTION DATASETS DESCRIPTION

Dataset	Train	Test	Visualize	Resolution	Lane	Environment	Labels' Meaning
TuSimple	3268	2782	2782	720×1280	≤ 5	highway	from left to righe
CULane	133235	34680	34680	590×1640	≤ 4	urban\highway	from left to right
Zjlab	695	174	10000	1080×1920	≤ 8	park	yellow\white\stop line

the daily operation of unmanned vehicles in Zhejiang Lab park, which is named Zjlab dataset in Table I.

CULane. For CULane, F1-measure is adapted for evaluation, similar to [14], each lane marking is treated as a line with 30 pixel width, thus intersection-over-union(IoU) could be calculated between the ground truth and the predicted segmentation. Predictions with IoUs larger than 0.5 are treated as true positives(TP) in our experiments. The F1-measure is defined as: $F_1 = \frac{Precision Recall}{Precision + Recall}$, where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$, FP means false positive and FN indicates false negative.

TuSimple. There is an official accuracy evaluation metric for TuSimple, which is defined as: $accuracy = \frac{\sum_{clip} C_{clip}}{\sum_{clip} S_{clip}}$, where C_{clip} is the number of lane key points that predicted correctly and S_{clip} is the number of ground truth lane key points in each clip (only the last frame of each clip is under evaluation). A lane key point is viewed as a true positive only if its distance from the ground truth is within 20 pixels.

Zjlab. To collect Zjlab dataset, we mounted cameras with a resolution of 1920×1080 on 3 vehicles during the daily operation. Finally, 869 typical frames with different vehicle positions and weathers are collected and labelled, e.g., sunny, cloudy and rainy. There are also bends, straights and junctions with lane markings blocked by obstacles. The labelled dataset is shuffled and divided into train and test sets, which have 695 and 174 images respectively. In addition, 10000 images are collected for visualization.

To access the semantic segmentation performance, the mean intersection-over-union (mIoU) is calculated following [27], which averages over IoUs of all classes, and the IoU is defined as: $IoU = \frac{TP}{TP+FP+FN}$, where TP, FP and FN indicate the number of true positive, false positive and false negative pixels. A pixel is treated as a true positive only when the predicted label is equal to the ground truth.

B. Experiment Settings

For Zjlab dataset, three versions of PriorLane are carried out: PriorLane-Imp, PriorLane-KE and PriorLane-KEA. Experiments of PriorLane-KE are conducted to verify the influence of prior knowledge embeddings. Moreover, PriorLane-KEA is carried out by adding the KEA module to PriorLane-KE. In addition, PriorLane-Imp is designed, whose architecture is the same as PriorLane-KEA, while trainable weights with randomly initialization are put into the KEA module instead of the knowledge embedding.

The RESA and SCNN models on Zjlab dataset are implemented base on the pytorch-auto-drive [28] codebase and the detailed configurations stay the same with those on CULane

benchmark published by the codebase. Models on Zjlab dataset are trained on the same server with 8 Tesla V100s and the input size is 540×960, and the main configurations of PriorLane models have no difference between those on CULane and TuSimple benchmarks.

C. Main Results

Results on CULane. To compare the results on CULane dataset with SOTA methods, four popular segmentation based methods are used for comparison, including SCNN, UFLD and RESA. Although UFLD is often classified as a row-wise method, compared with anchor-based methods it is more similar to the segmentation method, and auxiliary segmentation is used during training. Table II shows the F1-measures on the CULane dataset, and the modified method MiT-Lane outperforms other segmentation based methods. In Fig. 4, qualitative results on CULane dataset show that MiT-Lane based on fully vision transformer can predict the lanes occluded severely, because transformer can not only catch the longitudinal dependency, but also the inter-relationship between different lane markings.

Results on TuSimple. For TuSimple benchmark, LSTR is also used for comparison, because it is an early transformer based method in lane detection. Table. III shows the quantitative results on TuSimple benchmark, and MiT-Lane achieves the SOTA performance. Visualization results on TuSimple benchmark are shown in Fig. 4, and the MiT-Lane method can predict a longer lane marking based on the long-term dependency modeling ability of transformer, which is important for the highway scenario. However the longer part negatively affects the quantitative results, because it has no human-annotated labels.

Results on Zjlab. For Zjlab dataset, SCNN, RESA, MiT-Lane are used as baselines. Quantitative results in Table IV show that the proposed framework PriorLane-KEA get the best mIoU of 73.78%, which indicates that prior knowledge can improve the segmentation performance. Visualization results in Fig. 5 show that PriorLane-KEA can predict the blank of dotted line markings and the blocked part more accurately, which benefits from the local prior knowledge. In addition, the model can distinguish stop lines from white lane markings at a intersection.

Comparative experiments show that MiT-Lane outperforms SCNN and RESA on TuSimple, CULane and Zjlab datasets, which verifies that the transformer-only lane segmentation method without domain specific design is a qualified lane detector. Thus lane segmentation could benefit a lot from the fast-growing progress of the fully vision transformer, which is verified to be powerful on large datasets.
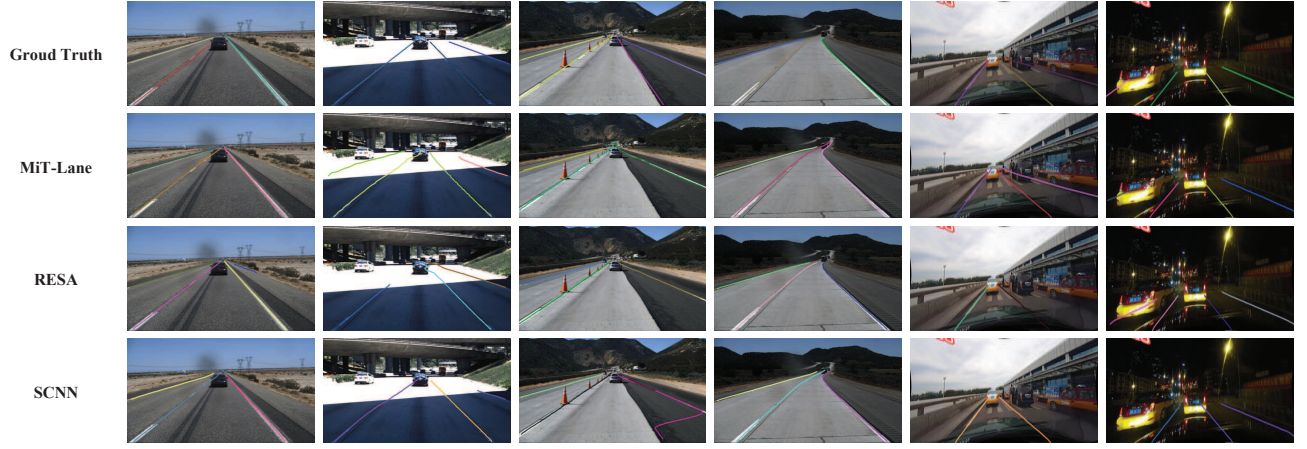
Fig. 4. Visualization results on CULane and TuSimple datasets, the first four columns are results on the TuSimple dataset, while the rest columns are results on the CULane dataset. Different lane marking instances are represented by different colors, note that MiT-Lane can predict a longer lane marking compared with the ground truth, however the longer part negatively affects the quantitative results.
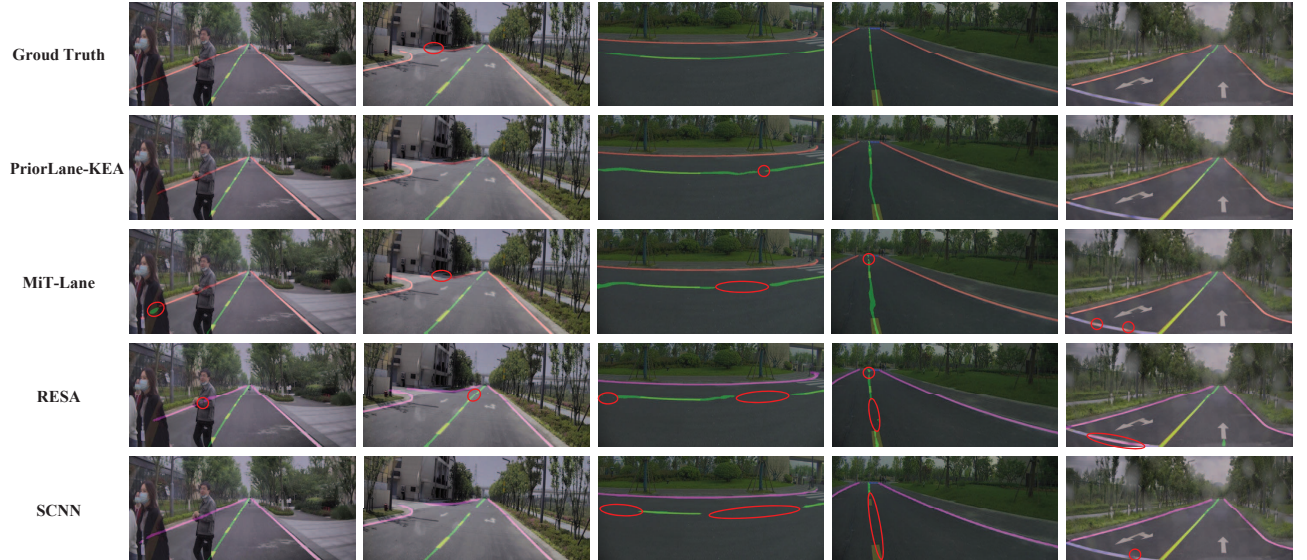

Fig. 5. Visualization results on Zjlab dataset, different types of lane markings are represented by different colors, and red circles indicate failures of ground truths and predictions. Note that PriorLane-KEA can predict the blank of dotted line markings, because of the ability of long-term dependency modeling, and the prior knowledge also contributes to pixel semantic classification.

TABLE II

COMPARISON OF F1-MEASURE WITH SOTA METHODS ON CULANE TEST SET, IOU THRESHOLD IS SETTED 0.5, AND ONLY FP IS SHOWN FOR CROSSROAD. * REPRODUCED BY THE PYTORCH-AUTO-DRIVE FRAMEWORK, WHICH IS THE BEST RESULT OF 3 RANDOM RUNS, ** RESULTS IN OFFICIAL PAPER. THE PERFORMANCE OF MIT-LANE IS THE MEAN VALUE FROM 3 RANDOM RUNS.

Methods	Total	Normal	Crowed	Night	No line	Shadow	Arrow	Dazzle	Curve	Cross
SCNN(ResNet-101)*	73.58	91.10	71.43	68.53	46.39	72.61	86.87	61.95	67.01	1720
UFLD (ResNet-34)**	72.3	90.7	70.2	66.7	44.4	69.3	85.7	59.5	69.5	2037
RESA (ResNet-50)*	74.04	91.45	71.51	69.01	46.54	75.83	87.75	63.90	68.24	**1522**
MiT-Lane	**76.27**	**92.36**	**73.86**	**70.26**	**49.6**	**78.13**	**88.59**	**68.26**	**73.94**	2688.3

D. Ablation Study

Analysis for components. PriorLane-KE surpasses MiT-Lane by a 1.12% mIoU, which indicates the explicit prompts could improve the performance of lane segmentation. Further comparison with PriorLane-KEA shows that the KEA module can improve the result by aligning the local prior knowledge embeddings, supported by a 0.5% mIoU. Note that PriorLane-Imp, which has the same architecture with PriorLane-KEA, outperforms MiT-Lane by a 0.46% mIoU, it is mainly due to the effect of the trainable prompts.

TABLE III

COMPARSION OF ACCURACY, FP, FN ON TUSIMPLE TEST SET, THE MAX
LENGTH OF THE PREDICTION IS SETTED 56. * REPRODUCED BY THE
PYTORCH-AUTO-DRIVE FRAMEWORK, WHICH IS THE BEST RESULT OF 3
RANDOM RUNS, ** RESULTS IN OFFICIAL PAPER. THE PERFORMANCE OF
MIT-LANE IS THE MEAN VALUE FROM 3 RANDOM RUNS.

Method	Accuracy	FP	FN
SCNN(ResNet-101)*	95.69	0.052	0.050
UFLD(ResNet-34)**	96.06	-	-
RESA(ResNet-101)*	95.56	0.058	0.051
LSTR(ResNet-18)*	95.06	0.049	0.042
MiT-Lane	**96.58**	**0.039**	**0.029**

TABLE IV

COMPARISON OF THE IOU ON ZJLAB TEST SET, EACH RESULT IS THE
MEAN VALUE FROM 3 RANDOM RUNS

Method	Mean IoU	Back-ground	White Line	Yellow Line	Stop Line
SCNN	67.89	99.30	66.66	51.39	54.23
RESA	70.96	99.43	71.88	61.12	51.41
MiT-Lane	72.16	99.56	74.01	61.99	53.08
PriorLane-Imp	72.62	99.57	74.95	62.63	53.31
PriorLane-KE	73.28	99.58	**75.25**	62.85	55.42
PriorLane-KEA	**73.78**	**99.59**	75.01	**63.12**	**57.39**

Number of Encoder Layers. There are two kinds of encoder layers: the knowledge encoder layer and the fusion encoder layer, and the number of knowledge encoder layers and fusion encoder layers are indicated by L_1 and L_2, respectively, as shown in Fig. 1.

To investigate the effect of the number of different encoder layers, experiments with different L_1 and L_2 values are carried out. As shown in Fig. 6, mIoUs vary with different L_1 and L_2 values, and there is a peak in the histogram. The knowledge encoder layer can be treated as a feature refiner, and deeper layers produce higher semantic features, which contribute to the segmentation performance. However the larger L_1 degrades the model generalization and leads to a decrease of the mIoU after the peak.

Higher semantic features produced by the knowledge encoder layer need deeper fusion encoder layers to merge the valid information. However the image feature dominate others in deeper fusion encoder layers, thus the performance could be suppressed, as a result, a larger L_2 does not necessarily obtain a higher mIoU under a fixed L_1.

Influence of Perception Range. To investigate the effect of different perception ranges, experiments with different perception ranges are carried out, as shown in Fig. 7. Since the local prior knowledge is represented in a grid map, which is rendered into an "image" with one channel, and the longer perception range results in a larger "image". Accordingly, local prior knowledge with different perception ranges are resized to the same size (200×200). As shown in Fig. 7, with the increase of the perception range, PriorLane-KEA obtains a higher mIoU with the help of the local prior
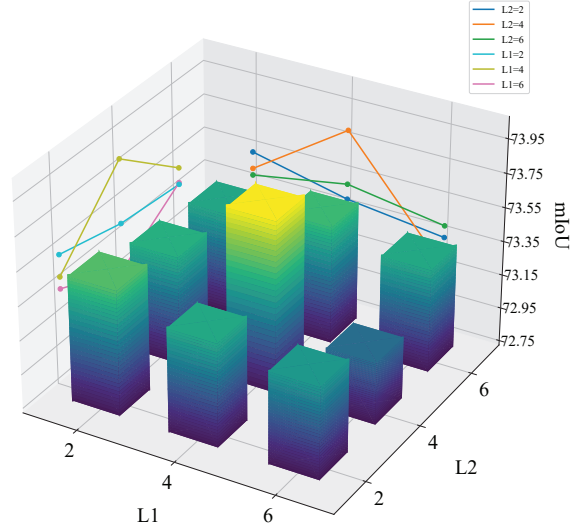

Fig. 6. Quantitative results of different settings of L_1 and L_2, represented by different pillars, and the number of attention heads is fixed to 8. Each result is the mean value from 3 random runs.

knowledge. However the image resolution in the distance is low, and it is difficult for the model to calculate the attention between the low resolution feature and the corresponding prior knowledge, as a result, there is a decrease when the perception range is too large.
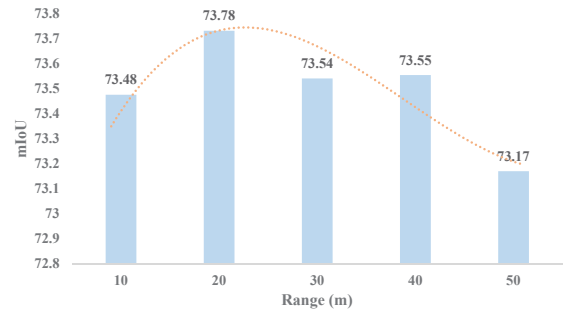

Fig. 7. Quantitative results of different perception ranges (the blue histogram), and the orange trend line indicates a downtrend when the perception range is larger than 20m. The value of L_1 and L_2 are fixed to 4 and 4, respectively, and the number of attention heads is fixed to 8. Each result is the mean value from 3 random runs.

V. CONCLUSION

In this paper, we have verified that the transformer-only architecture on lane detection and general segmentation can be unified in an elegant way. Thus, lane detection task could benefit from the rapid development of vision transformer pre-training. In addition, a novel framework called PriorLane is proposed, which fuses the local prior knowledge of the scene with images. Moreover, a KEA module is adapted to align the data from two different modalities, and experiments show that both the framework and KEA module could enhance the performance of lane segmentation. In the future, a larger dataset with more kinds of low-cost prior knowledge will be collected to enhance the segmentation performance.

REFERENCES

[1] A. Kawasaki and A. Seki, "Multimodal trajectory predictions for urban environments using geometric relationships between a vehicle and lanes," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 9203–9209.

[2] K. Brandes, A. Wang, and R. Shah, "Robust lane detection with binary integer optimization," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 229–235.

[3] M. Beyeler, F. Mirus, and A. Verl, "Vision-based robust road lane detection in urban environments," in *IEEE International Conference on Robotics and Automation)*, 2014, pp. 4920–4925.

[4] W. Jiang, Y. Wu, L. Guan, and J. Zhao, "Dfnet: Semantic segmentation on panoramic images with dynamic loss weights and residual fusion block," in *IEEE International Conference on Robotics and Automation*, 2019, pp. 5887–5892.

[5] H. Abualsaud, S. Liu, D. B. Lu, K. Situ, A. Rangesh, and M. M. Trivedi, "Laneaf: Robust multi-lane detection with affinity fields," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7477–7484, 2021.

[6] X. Li, J. Li, X. Hu, and J. Yang, "Line-cnn: End-to-end traffic line detection with line proposal unit," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 248–258, 2019.

[7] L. Tabelini, R. Berriel, T. M. Paixao, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "Keep your eyes on the lane: Real-time attention-guided lane detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 294–302.

[8] Z. Qin, H. Wang, and X. Li, "Ultra fast structure-aware deep lane detection," in *European Conference on Computer Vision*, Springer, 2020, pp. 276–291.

[9] J. Philion, "Fastdraw: Addressing the long tail of lane detection by adapting a sequential prediction network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11582–11591.

[10] R. Liu, Z. Yuan, T. Liu, and Z. Xiong, "End-to-end lane shape prediction with transformers," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3694–3702.

[11] L. Tabelini, R. Berriel, T. M. Paixao, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "Polylanenet: Lane estimation via deep polynomial regression," in *IEEE International Conference on Pattern Recognition*, 2021, pp. 6150–6156.

[12] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial cnn for traffic scene understanding," in *AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[15] T. Zheng, H. Fang, Y. Zhang, W. Tang, Z. Yang, H. Liu, and D. Cai, "Resa: Recurrent feature-shift aggregator for lane detection," in *AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, pp. 3547–3554, 2021.

[16] R. Asghar, M. Garzón, J. Lussereau, and C. Laugier, "Vehicle localization based on visual lane marking and topological map matching," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 258–264.

[17] K. Petek, K. Sirohi, D. Büscher, and W. Burgard, "Robust monocular localization in sparse hd maps leveraging multi-task uncertainty estimation," in *IEEE International Conference on Robotics and Automation*, 2022, pp. 4163–4169.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, Springer, 2020, pp. 213–229.

[19] W. Wang, Y. Cao, J. Zhang, and D. Tao, "Fp-detr: Detection transformer advanced by fully pre-training," in *International Conference on Learning Representations*, 2021.

[20] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," *arXiv preprint arXiv:2203.17270*, 2022.

[21] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *IEEE International Conference on Robotics and Automation*, 2022, pp. 9200–9206.

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

[23] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[24] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Oriented response networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 519–528.

[25] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[26] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional detr for fast training convergence," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3651–3660.

[27] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[28] Z. Feng, S. Guo, X. Tan, K. Xu, M. Wang, and L. Ma, "Rethinking efficient lane detection via curve modeling," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17062–17070.