Multi-task Attention Network for Lane Detection and Fitting

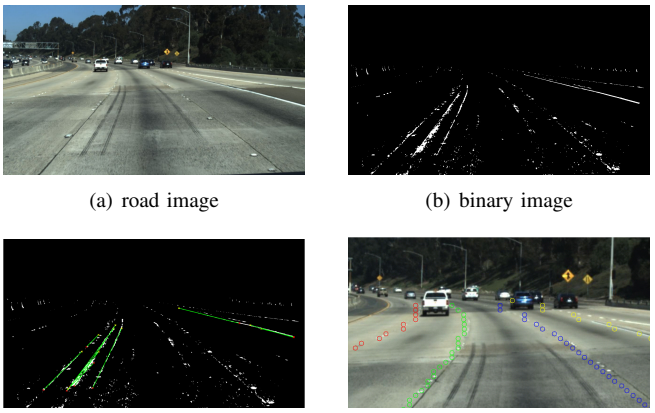
Qi Wang, *Senior Member, IEEE*, Tao Han, Zequn Qin, Junyu Gao, *Student Member, IEEE*, Xuelong Li, *Fellow, IEEE*

***Abstract*—Many CNN-based segmentation methods have been applied in lane marking detection recently and gain excellent success for a strong ability in modeling semantic information. Although the accuracy of lane line prediction is getting better and better, lane markings' localization ability is relatively weak, especially when the lane marking point is remote. Traditional lane detection methods usually utilize highly specialized hand-crafted features and carefully designed post-processing to detect the lanes. However, these methods are based on strong assumptions and thus are prone to scalability. In this work, we propose a novel multi-task method, which 1) integrates the ability to model semantic information of CNN and the strong localization ability provided by handcrafted features and 2) predicts the position of vanishing line. A novel lane fitting method based on vanishing line prediction is also proposed for sharp curves and non-flat road in this paper. By integrating segmentation, specialized handcrafted features and fitting, the accuracy of location and the convergence speed of networks are improved. Extensive experimental results on four lane marking detection datasets show that our method achieves state-of-the-art performance.**

I. Introduction

In 1998, the GOLD [1] system is proposed to detect obstacles and lanes in a structured environment, which is the first well-known method in lane detection to the best of our knowledge. However, lane marking detection in the open world and unstructured road (without distinct lane markings) has been a long-standing problem in the last few decades. In general, the primary difficulties of lane detection can be summarized into two parts, which are feature extraction and lane modeling. The feature extraction part can be used for locating the lane markings. Then the lane modeling procedure summarizes the detection results in the mathematical form.

Mainstream lane marking detection methods usually utilize hand-crafted low-level image processing and carefully designed post-processing to tackle the difficulties of feature extraction and lane modeling. This kind of method is straightforward and integrates much prior knowledge based on certain assumptions. In a restricted environment, these methods could achieve good performance. However, due to the unsatisfactory generalization ability of the hand-crafted features and assumptions, the scale of these methods is always not large enough for practical use. Many assumptions like parallel lane, flat ground, limited directions, fixed ROI, and lane width might fail.



(a) road image (b) binary image

(c) conventional baseline hough transform (d) deep method results at the far end of road

Fig. 1. A demonstration of some conventional results from low-level image processing method and deep method. Figure (a) shows the original road image, and Figure (b) is the binary image of dark-light-dark (DLD) features, which can be regarded as a pre-processing process to preliminarily extract lane line by detecting the change of light and shade of the pixel. Figure (c) shows the top 10 results of the Hough transform, which has been widely used in the low-level image processing method. The deep method results at the far end of the road are shown in Figure (d).

Moreover, the inherent pipeline structure of lane marking detection could spread and enlarge the error caused by inappropriate assumptions. Another drawback of these methods is that the detection range is usually small. Thus, the results could not extend to the end of the road. It is crucial to detect as far as possible because further detection results can bring more reaction time in autonomous driving. As shown in Fig. 1(c), although the hand-crafted low-level image processing method gives good feature extraction results of lane markings, the analysis on these intermediate results is tough. Even with the top 10 results, the Hough transform, a widely used method in lane marking detection, still could not find the correct lane markings. Although many methods utilize post-processing like inverse perspective mapping (IPM) [2], [3], [4], a transformation to eliminate the curvature of parallel objects in the image caused by perspective effect, and temporal information to achieve better performance, the difficulties of analyzing low-level image processing still remain.

Another kind of method proposed recently utilizes deep segmentation networks to estimate the location of lane mark-

Manuscript received December 03, 2019; revised August 02, 2020; accepted November 13, 2020. This work was supported by the National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, and 61871470.

The authors are with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@gmail.com; hantao10200@mail.nwpu.edu.cn; zequnqin@gmail.com; gjy3035@gmail.com; xuelong_li@nwpu.edu.cn). Xuelong Li is the corresponding author.

ing directly. The major advantage of this method is the strong semantic ability. In the high semantic level, the deep segmentation method achieves remarkable performance when estimates the general lane marking directions. However, there is a potential problem that the output stride of modern segmentation architecture is large. The output stride usually varies from 8 to 32 [5]. As a result, the detail of segmentation cannot be well captured. Although some methods [6], [7] utilize encoder-decoder structure to address the problem of output stride, the above problem still exists. When it comes to the lane marking detection, the estimated locations of the far end lane markings might be poor. As shown in Fig. 1(d), when we zoom in the result of the deep segmentation method at the far end of the road, the localization ability becomes terrible.

In this paper, we integrate the localization ability of hand-crafted low-level image processing method and the semantic analysis ability of the deep segmentation method. A multi-task lane segmentation network and a novel vanishing line fitting method are proposed to tackle the above problems. To combine the advantages of deep segmentation method and traditional method, we propose an adaptive dark-light-dark (ADLD) method for low-level feature extraction. Then the proposed ADLD features are used as spatial attention information in the multi-task network. The added ADLD spatial attention could not only gain higher performance but also speed up the training process. To further refine the results, we propose a single parameter inverse perspective mapping method, which uses the position of vanishing line position as the only parameter. By using the proposed method, the number of IPM parameters drops from 6 [8] to one, which can be obtained from the previous multi-task network. Moreover, the proposed method could run at 20-50 fps depending on different backbone networks.

The main contributions of our work are summarized as follows.

- Propose a novel multi-task framework for lane detection. It consists of lane marking segmentation and vanishing line prediction tasks. The vanishing point prediction task makes the inverse perspective mapping can be implemented without the camera parameters, which bring a refined segmentation result.
- Utilize the ADLD features to guide the training process, directly providing pixel-level spatial attention for the front end and DULR module. The ADLD features not only enlighten the network to focus on the lane line but also make the model works faster than other attention modules.
- Conduct extensive ablation studies to verify that all the proposed components contribute to the detection results, and experiments on four datasets (Caltech lanes [11], KITTI [26], Tusimple Benchmark [38] and CULanes [30].) show that the proposed framework achieves comparable performance with a faster speed.

The remainder of the paper is structured as follows. In Sec. II, a brief review of many traditional and deep methods is listed. The proposed method is described in Sec. III. Extensive experiments and ablation studies are shown in Sec. IV. At the end of this paper, Sec. V summarizes this work.

II. Related Work

In the past many years, many methods are proposed for lane marking detection in many aspects. In general, these methods can be divided into two categories, which are deep segmentation method and traditional method using hand-crafted low-level image processing and carefully designed post-processing.

A. Traditional Method

The GOLD system [1] utilizes stereo IPM to detect lanes and obstacles. Due to the limited computing resources, this system is just a customized prototype that runs at 10fps. In [9], a stochastic road shape estimation system is proposed. A two-stage lane marking extracting algorithm is applied in this work. Meanwhile, an estimation using particle filtering is also proposed. Unlike the cubic spline used in [9], [10] proposed a lane detection and tracking method using the B-Snake model.

Then Hough transform and random sample consensus, a combination of IPM, is proposed in [11]. In this work, the feature extraction method is a specially designed Gaussian filter, which has a similar shape with the ridge. However, the direction of the Gaussian filter is fixed. In [12], gabor filters are introduced to describe multiple orientations. The post-processing method of [12] is still a Hough transform. Then a Gaussian sum particle filter is proposed in [13], which is based on an assumption of the vanishing point. In [14], a hierarchical approach is proposed, which utilizes low-level classifiers and spatial layout features.

Unlike previous works [11], [12], [15] focusing on improving feature extraction, [16] proposes a novel method that eliminates the shadow on the road to improve performance and robustness. [17] proposed an invariant illumination lane marking detection method. This paper utilizes YCbCr color space to achieve illumination invariant processing of road images. Some works try to adopt more prior knowledge. In [18], an optimization technique with a conditional random field (CRF) is proposed in this paper. The original lane marking detection problem is transformed into an optimization problem. In [19], a lane marking ground truth generation method is proposed based on the Time-Sliced image, which is stacked from multiple frames. Later on, [20] introduces a similar technique called spatiotemporal image to predict lane markings. From the above, we can see that most traditional methods focus on improving feature extraction and utilizing prior knowledge, but scalability problems always exist.

B. Deep Method

With the rising of deep segmentation methods [21], many deep methods are proposed for lane marking detection. [22] gave a primacy attempt that combines deep networks and random sample consensus. The present deep networks in [22] is auxiliary. Following the idea of [22], [23] introduces a region-based lane marking classification method, which also uses the networks as auxiliary means. In [24], a multi-task network aims to evaluate the effectiveness of deep networks

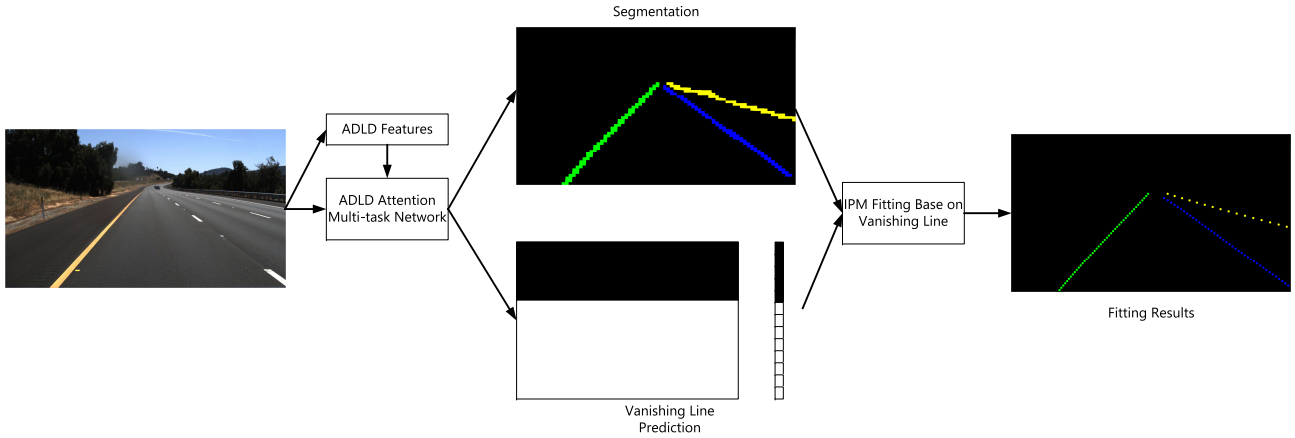


Fig. 2. Overall structure of the proposed method. The proposed ADLD attention multi-task network first segments the lane markings and predicts the position of vanishing line. Then a IPM method is proposed to utilize the position of vanishing line. By integrating these results, the final fitting results are obtained.

on highway driving is proposed. In [25], a network architecture that aims to find the best trade-off between segmentation quality and runtime and achieves top accuracies on the KITTI dataset [26]. A Bayesian network is proposed in [27], which can detect both road and road boundaries in a single process.

After proving the effectiveness of multi-task learning, [28] proposes a lane and road marking detection network, which utilizes the multi-task information derived from the vanishing point. Besides convolutional neural networks, [29] introduces a long short-term memory network that regards lane marking detection as a sequential problem. Another method that captures the sequential information of lane is the SCNN [30]. The SCNN method regards the spatial information as the deep information and changes the direction of convolution. Such spatial convolution benefits the information flow along the lane and thus gains better performance. When the segmentation accuracy of lane markings becomes high, more methods start to extract lane markings from segmentation. [8] proposed an instance segmentation network which also outputs the estimation of camera parameters. These parameters are used for lane fitting. A potential problem of this method is that the training of the network is based on an ill-conditioned problem. The calculation of the loss function contains an inverse matrix when conduct mapping and inverse mapping utilizing these camera parameters. In this way, for a tiny difference in the input, the inverse operation could enlarge this difference. As a result, the training of this network might be tough. Different from [8] adopts clustering, [31] introduces an instance segmentation method that directly solves the relationship problem defined by instances and uses graph coloring to assign instance ID. The above works are in visual image-based lane detection. It is still challenging to accurately identify road areas in visual images, such as illumination changes and blurred images. Because LIDAR data is less likely to be susceptible to visual persuasion, some works [32], [33] are integrating LIDAR data to improve the visual image-based road detection. Although these methods achieve great performance, the mentioned location problem still exists.

III. Methodology

In this section, we demonstrate the proposed ADLD attention multi-task network and the lane fitting method based on the vanishing line. The overall structure is shown in Fig. 2.

A. Adaptive dark-light-dark features

The main idea of adaptive DLD features is that the lane markings are bright and the surrounding are dark. Denote $D^+(x,y)$ as the right light-dark part image and $D^-(x,y)$ as the left dark-light part image. We have:

$$D^+(x,y) = I(x,y) - I(x+m,y), \tag{1}$$

$$D^-(x,y) = I(x,y) - I(x-m,y), \tag{2}$$

where $I(x,y)$ is the value of a gray image, and m is the neighbor distance parameter. Generally, the m is set as the width of the lane. Inspired by canny [34] edge detection, we adopt an adaptive two-stage threshold method to obtain binary masks from the dark-light-dark features. The high threshold is defined as the top $q\%$ quantile value from corresponding features, and the low threshold is defined as the top $2q\%$ quantile value. In this work, q is set to 10. Thus we have:

$$D_{high} = \begin{cases} 1 & D^+(x,y) > ht^+ \wedge D^-(x,y) > ht^-, \\ 0 & else \end{cases} \tag{3}$$

$$D_{low} = \begin{cases} 1 & D^+(x,y) > lt^+ \wedge D^-(x,y) > lt^-, \\ 0 & else \end{cases} \tag{4}$$

where the ht and lt are high and low thresholds obtained from top $q\%$ and $2q\%$ quantile values, respectively. Then, we use connected component in D_{high} to select the weak features in D_{low}. The final binary image is generated in this way.

B. ADLD Attention

Because ADLD features provide pixel-wise labeling, its localization ability is good enough to be a heuristic for spatial attention. We use such heuristic information as spatial attention [35], [36], which could guide the training process. Inspired by SCNN [30], we also utilize the DULR module, an information

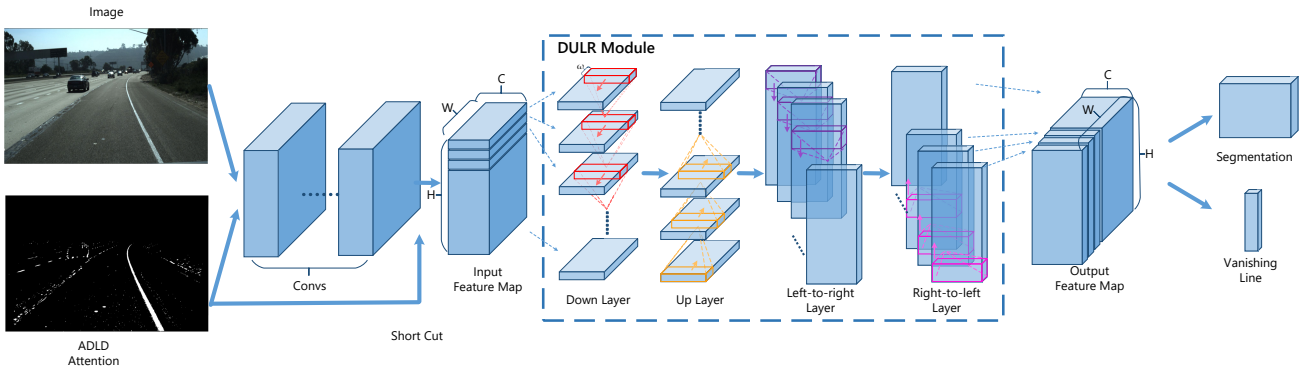


Fig. 3. The structure of the multi-task attention network. The ADLD attention is both added to the input layer and the DULR module. The DULR module is used for better performance as [30]. D, U, R, and L refer to the four feature passing units, which are similar in structure. The spatial information is conveyed from up to down, down to up, left to right, right to left. The output of the network is composed of segmentation and vanishing line prediction. As shown in Fig. 4, the vanishing line prediction task is accomplished with indicator vector.

pass module from row and column directions with convolution operation in slice features, which could capture the spatial information to optimize the segmentation of lane markings. Unlike common spatial and channel-wise attention obtained from networks by itself, the ADLD attention can be seen as a preprocessing of original data. There are two parts that utilize ADLD attention in the proposed segmentation network.

As the long and thin lane markings are hard to learn for a network, the first parts where we add the attention information is the input layer. In this work, the ADLD attention is concatenated with the original inputs. In this way, the added information could speed up the training process. As we all know, the shallow layers of neural networks represents low-level edge and texture information. Because the added ADLD features are already composed of low-level information, the learning of shallow layers could benefit from the ADLD features.

The second part where we add the attention information is the DULR module. The ADLD attention is scaled and concatenated with the features generated from previous convolutional layers. Because the DULR module encodes the spatial information, it is convenient to introduce the ADLD attention in the DULR module. With the ADLD features added, the salient part in the attention, which is the lane marking part in most cases, could contribute to the message passing within the DULR module. As a result, the DULR module encodes more information guided by the ADLD attention, and the localization ability of the network is improved. More discussions about the impact of ADLD attention can be seen in Sec. IV.

C. Vanishing Line Estimation Task

In this paper, the vanishing line is defined as the horizontal line crossing vanishing point. Due to the vanishing point's inherent geometry information, there are many methods [28], [10] utilize it as auxiliary knowledge to improve performance. For the same reason, the vanishing line could help other tasks like lane marking segmentation and fitting in this work. The common ways of using deep neural networks to predict vanishing points are segmentation-based methods. As shown in Fig. 4(b), the idea of segmenting vanishing point image

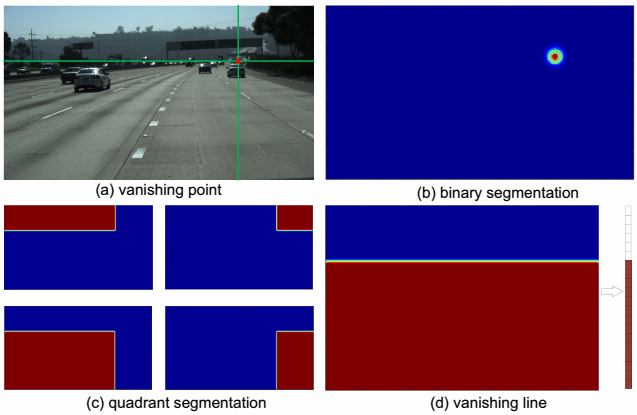


Fig. 4. Comparison of vanishing point segmentation and the proposed vanishing line prediction task. Figure (a) shows the location of the vanishing point in a road image. In figure (b) and (c), two kinds of segmentation targets are shown. The ideal binary segmentation in figure (b) only treats the vanishing point as the foreground. Figure (c) shows a variant that divides the image into four parts according to the vanishing point location. The proposed vanishing line prediction task is shown in figure (d). The prediction of the vanishing line can be simplified by using an indicator vector.

is straightforward, but this kind of task might fail due to the heavily imbalanced data [28]. The area of the vanishing point in this task is too small to dominate the training process, and the output of the network would converge to the class of background. An improved variant in Fig. 4(c) is to segment four parts divided by vanishing point. However, the determination of the vanishing point's location is relatively tough. The proposed vanishing line estimation task only predicts the location of the horizontal line. In this way, the original task can be simplified with an indicator vector along the vertical direction shown in Fig. 4(d). Such simplification could improve the performance of vanishing line prediction and eliminate the difficulties in determining the location from a segmentation mask. Furthermore, the computational cost is lower than the quadrant segmentation.

The overall structure of the proposed multi-task attention network is shown in Fig. 3. Our network adopts raw image and ADLD features as inputs and outputs the segmentation of lane markings and the prediction of vanishing line position.

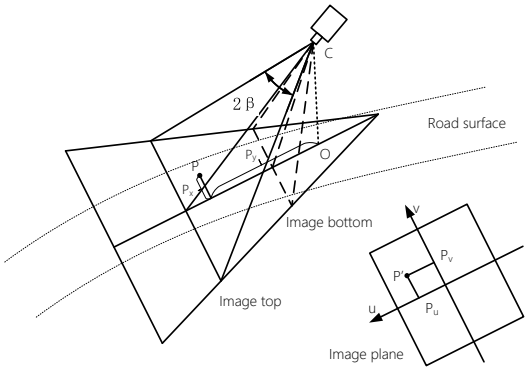


Fig. 5. Geometric model of IPM. β is the horizontal angular aperture of the camera. The point C is the optic center, and the point O is the projection of C on the ground.

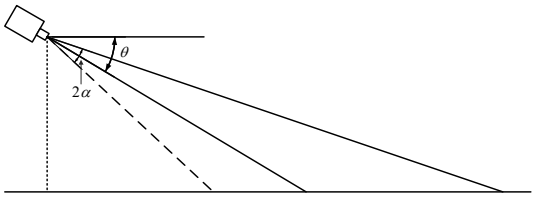


Fig. 6. Side view of the IPM Geometric model. θ is the pitch angle, and α is the vertical angular aperture of the camera.

D. Lane Fitting based on Vanishing Line

Due to the perspective effect, the curve lane trends have bigger curvature at the far end of the road while the lane in the lower part of the image is fairly straight. An excellent lane marking fitting method should generally fit a straight line in the majority part of the lane. Meanwhile, if the lane is not straight and has large curvature caused by perspective effect, the method should fit as bent as possible at the far end part of the road image. This phenomenon causes the main difficulty in lane marking fitting.

One way to eliminate such difficult is to conduct inverse perspective mapping and get rid of the effect produced by perspective. However, the IPM process is parametric and requires 6 intrinsic or extrinsic camera parameters. It should be noted that these parameters are not easy to obtain, and the extrinsic camera parameters vary if the position of the camera changes. Besides, many lane marking detection datasets don't provide these parameters. To tackle the above problems, we propose a single parameter IPM method that the required parameter can be easily obtained from the previous vanishing line prediction network.

Based on the principle of pinhole imaging, the geometric model of IPM is shown in Fig. 5. Denote α as the vertical angular aperture of the camera and β as the horizontal angular aperture of the camera. θ means the pitch angle of the camera, and h is the height of the camera. The image size is denoted as $m \times n$. The relations between a pixel (u, v) and the world coordinate (x, y) can be formulated as:

$$y = hcot\left\{\theta - atan[tan\alpha(1 - \frac{2u}{m-1})]\right\} \tag{5}$$

and

$$x = \sqrt{h^2 + y^2}\frac{tan\beta(\frac{2v}{n-1} - 1)}{\sqrt{1 + [tan\alpha(1 - \frac{2u}{m-1})]^2}}. \tag{6}$$

Then the corresponding inverse transformation is:

$$u = \frac{m-1}{2}[1 - \frac{tan\theta - atan(\frac{y}{h})}{tan\alpha}] \tag{7}$$

and

$$v = \frac{n-1}{2}\left\{1 + \frac{x}{tan\beta\sqrt{h^2 + y^2}}\sqrt{1 + tan^2[\theta - acot(\frac{y}{h})]}\right\}. \tag{8}$$

From Eq. 5 and 6, we can see that for different parameter of h, the ratio of mapped x and y are constant. Meanwhile, the effects of β, m, and n are linear according to the above equations. By this way, after the rasterization and scaling of the inverse perspective image, these parameters would no longer affect the generated image. As a result, there are only two parameters pitch angle θ and vertical angular aperture of the camera α that could affect the IPM. Fortunately, the pitch angle can be obtained from the previously introduced vanishing line prediction network. In [37], the pitch angle can be estimated from the position of vanishing point:

$$\theta = atan[tan\alpha(1 - \frac{2\bar{y}}{n})], \tag{9}$$

where $\bar{y}$ is the vertical position of the vanishing point and also the position of the vanishing line. It can be verified that the vertical angular aperture of the camera α would no longer affect the IPM transformation by using the estimated pitch angle. As described above, the parameters of m, n, and h would not affect the results of IPM. In this way, as long as we set meaningful but not have to be exact values to these parameters, the proposed method could generate a correct IPM image.

We can now use the proposed IPM method based on vanishing line prediction to refine the results of lane marking segmentation. Because the perspective effect can be eliminated using IPM, the lane marking points obtained from the segmentation mask are transformed with the proposed mapping method. In the transformed space, the lane marking points is fitted with cubic polynomials. Then the fitted points are transformed into the original image space. The overall method is summarized in Algorithm 1.

IV. Experiment

In this section, we first show the datasets and the corresponding experimental setup we used. Then, the effectiveness of the proposed method is examined using ablation studies, and the experimental results on the four lane marking detection datasets are shown.

A. Data preparation and Setup

Dataset. There are four datasets which are Caltech lanes dataset [11], KITTI dataset [26], Tusimple Benchmark dataset [38] and CULanes dataset [30] used in this work. The Caltech lanes dataset contains 1,224 labeled frames and 4,172 marked

Algorithm 1 Lane fitting based on vanishing line

Input:
The image size m and n;
Meaningful arbitrary parameters of height h, vertical angular aperture of the camera α and horizontal angular aperture of the camera β;
Lane marking points P obtained from segmentation mask;
Estimated vanishing line position $\bar{y}$;
Output:
Fitted Lane marking points P_{fit};
1: Calculate the pitch angle using Eq. 9;
2: Transform lane marking points P to P' in the IPM space using Eq. 5 and 6;
3: Fit lane marking point P'_{fit} using P' with cubic polynomials;
4: Transform fitted point P'_{fit} to P_{fit} in the original image space using Eq. 7 and 8;
5: **return** P_{fit};

lanes. The two urban scenarios in Cordova and Washington are included in the Caltech lanes dataset. The KITTI lane detection dataset is part of the KITTI road benchmark, with 95 images for training and 96 images for testing. The Tusimple Benchmark dataset is released for the CVPR 2017 workshop on the autonomous driving challenge, which is composed of about 10,000 one-second-long video clips. Among them, 6,408 images are annotated with lane markings. Each frame contains 2 to 5 lanes. There are 3,626 images used for training and 2,782 images used for testing. The CULanes dataset is collected in Beijing with 88,880 frames for the training set, 9,675 frames for the validation set, and 34,680 frames for the testing set. The dataset is divided into nine scenarios: normal, crowded, night, shadow, and curve, etc. The information of the four datasets is listed in Table I.

TABLE I
BRIEF COMPARISON OF THE FOUR DATASETS.

Dataset	Training	Testing	Size	Scenarios
Caltech	N/A	1224	640x480	urban streets
KITTI	95	96	1242x375	urban streets
Tusimple	3626	2782	1280x720	high way
CULanes	88880	34680	1640x590	urban streets

Vanish point annotation. In our multi-task framework, for lane marking segmentation task, supervised labels are already provided. However, for the vanishing line prediction task, we have no relevant supervised information to use. Therefore, we manufacture vanishing point labels for the above four datasets. First of all, we use the ground truth of lane lines for automatic labeling. By detecting where the lane lines intersect, we annotate the crossover position as the vanishing point. For some images with large curvature, we manually label them. Since each image only needs to be marked with one point, it does not consume much labor. In total, the entire annotation process costs 5 human hours.

Training details. For the Caltech, Tusimple Benchmark and CULanes dataset, Adam [39] optimizer is utilized with a base learning rate of 5e-4 and a weight decay of 1e-4. The learning rate of KITTI dataset is set to 2e-4. The learning rate policy is the poly with a power of 0.9. For the Tusimple Benchmark dataset and CULanes dataset, the image is resized and cropped to 800×288. Random flip and rotation augmentation are also utilized.

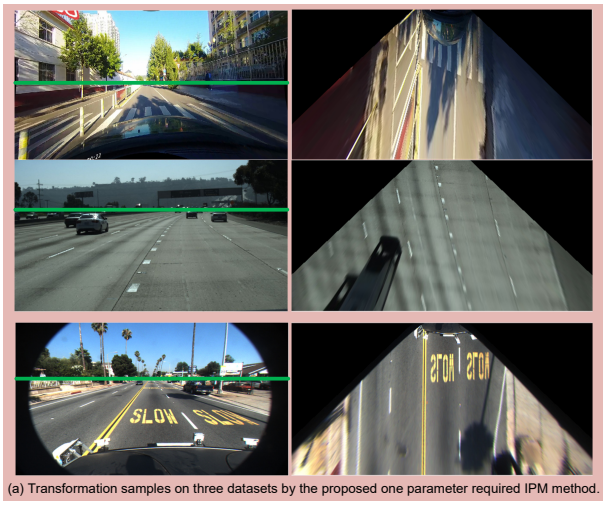


Fig. 7. (a) The proposed IPM method is applied on the three different datasets, where the experiment configurations are the same. (b) The original IPM is conducted on the Caltech Lane dataset with its camera parameters.

B. Evaluation

In this work, we follow the evaluation metrics defined by the corresponding datasets. For the Caltech lanes dataset, the criterion of the correct lane is defined as: $min(\hat{d}_1, \hat{d}_2) \leq t_1$ and $min(\bar{d}_1, \bar{d}_2) \leq t_2$, in which $\hat{d}$ is the median value of the nearest distance, and $\bar{d}$ represents the mean distance. The subscript indicates the lane number used for calculation. Following the implementation of Caltech lanes dataset, t_1 and t_2 are set to 20 and 15 respectively. The evaluation metric is the rate of correctly detected lanes.

For the KITTI dataset, we use the metrics provided by the evaluation benchmark, which are maximum F-measure (MaxF), average precision (AP), precision, recall, false positive rate (FPR) and false negative rate (FNR). Suppose τ is the threshold of classification. The primary metric MaxF is defined as:

$$MaxF = \max_{\tau} F - measure. \tag{10}$$

F-measure is defined as:

$$F - measure = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}, \tag{11}$$

in which P is the precision, R is the recall and β is set to 1 as the F1-measure.

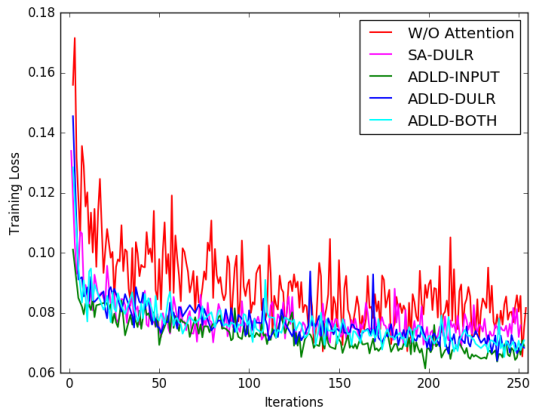


Fig. 8. Training losses of segmentation task with different settings of attention. W/O Attention is the baseline method without attention. SA-DULR means the self-attention module is added to the DULR module. ADLD-INPUT and ADLD-DULR are methods that ADLD attention is added to the input layer and the DULR module respectively. ADLD-BOTH means that ADLD attention is both added to input layer and DULR module.

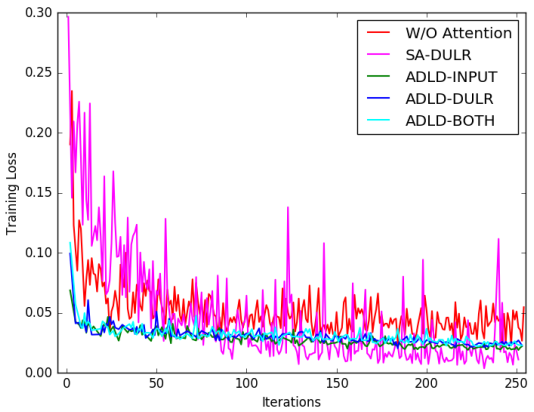


Fig. 9. Training losses of vanishing line prediction task with different settings of attention.

For the Tusimple Benchmark dataset, the evaluation metric is defined according to the correctly detected lane marking points. The formula is:

$$acc = \frac{\sum_{clip} C_{clip}}{\sum_{clip} S_{clip}}. \tag{12}$$

where C_{clip} is the number of correct points in the last frame of the clip, and S_{clip} is the requested points in the last frame of the clip. If the distance between the sampled detection lane point and the ground truth is below 20 pixels, it is considered as a correct match. If the number of detected lanes beyond $N + 2$ or the detection time is larger than 200ms, all of the points in this clip are considered as false detection. N is the number of lanes in the ground truth.

For the CULanes dataset, the evaluation metric is based on intersection-over-union (IoU). The ground truth is composed of 30 pixels wide lane segmentation mask. If the IoU of detection is larger than a certain threshold, it is regarded as a true positive. There are two thresholds, which are 0.3 and 0.5, corresponding to loose and strict evaluation. Then the final evaluation metric is F1-measure.

C. Ablation Study

1) Effectiveness of IPM based on vanishing line: In Sec. III-D, we propose the fitting method based on vanishing line. One key step in this method is IPM based on vanishing line. As described in Sec. III-D, compared with other IPM methods, the proposed IPM method only requires the vanishing line's position, and other parameters are arbitrary. To prove the effectiveness of this method, 1) we conduct the proposed IPM on three entirely different cameras under the same experiment configuration. As shown in Fig. 7(a), the proposed method on three datasets captured by different camera parameters achieves consistent and quite good results. 2) Most Lane line datasets do not provide the camera parameters. Among the dataset used in the paper, only the Caltech Lane dataset contains the camera parameters. To prove that the IPM method based on the vanishing line and camera parameters have the same performance, we conduct the comparison experiments on the Caltech Lane dataset. The third row of Fig. 7(a) shows that the proposed IPM achieves the approximate results comparing with the original IPM method shown in Fig. 7(b).

2) Effectiveness of ADLD Attention: In our framework, two parts are combined with the proposed ADLD attention. We evaluate the effectiveness of the proposed ADLD attention by comparing the losses of different tasks. In this experiment, segmentation and vanishing line losses are shown in Fig. 8 and 9. Each task contains five experimental settings. The first one is the original network without ADLD attention. The second and the third settings are the network added with only one ADLD attention to the input layer or the DULR module. The fourth one is the setting, in which ADLD attention is added to the input layer and the DULR module. In order to compare the differences of ADLD attention and other self-attention methods, another method Point-wise Spatial Attention Network (PSANet) [40] is also included. Due to the complexity of PSANet, the self-attention module is only added to the DULR module to replace the original ADLD attention, which has a smaller size. All of the experiments are conducted on Tusimple Benchmark Dataset with the same parameters.

From Fig. 8 and 9, we can see that all of the networks combined with the ADLD attention or self-attention achieve better performance in both lane marking segmentation and vanishing line prediction tasks. However, the training losses of the second task of the self-attention method are less stable. ADLD attention is slightly better than self-attention in the segmentation task, while self-attention has better results in vanishing line prediction task. The reason is that ADLD attention is designed for lane markings instead of objects like the sky. So it has less effect on the vanishing line prediction task than self-attention. These results also confirm that multi-task training could achieve better performance.

Another comparison of these different attention settings is shown in the first row of Table II, which is the lane marking point accuracy on the Tusimple Benchmark testing set. We can

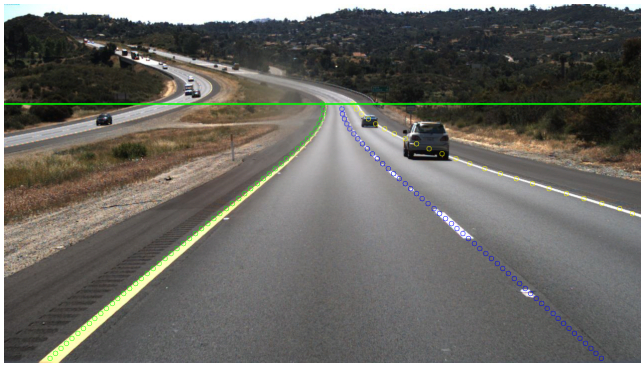


Fig. 10. Fitting results on non-flatten ground.

see that all of the attention methods achieve higher accuracy than the baseline method without attention. Although the self-attention method could also utilize spatial attention, the ADLD attention is specialized for the segmentation task so it could gain higher performance for this task. Besides, the computational complexity of ADLD features is linear, and the speed is much faster than convolutional layers. In this way, the proposed attention method is much simple and efficient for lane marking segmentation.

TABLE II
LANE MARKING POINT ACCURACY ON TUSIMPLE BENCHMARK TESTING SET. PLAIN NETWORK MEANS THE SETTING WITHOUT DULR MODULE AND ATTENTION. FIRST ROW IS THE LANE MARKING POINT ACCURACY OF THE ORIGINAL NETWORK OUTPUTS. SECOND ROW SHOWS THE RESULTS OF LEAST SQUARES FITTING. THE LAST ROW CONTAINS THE RESULTS OF THE PROPOSED FITTING METHOD.

Method	Plain Network	W/O Attention	SA-DULR	ADLD-INPUT	ADLD-DULR	ADLD-BOTH
W/O Fitting	95.49	95.49	95.52	95.53	95.55	95.73
W/ LS Fitting	95.53	95.89	95.84	95.93	95.97	96.00
W/ Propose Fitting	96.06	96.19	96.29	96.27	96.33	96.37

3) Effectiveness of Lane Marking Fitting: In this subsection, we demonstrate the effectiveness of the proposed fitting method. The proposed lane marking fitting method uses the prediction of vanishing line as a parameter. Generally, the higher accuracy of vanishing line prediction results in fewer errors of the proposed fitting method. However, because the proposed transformation method of IPM is mathematically invertible, the transformed lane marking points with the inaccurate parameter of vanishing line prediction could still be inversely transformed. In this way, the demand for accuracy of vanishing line prediction is not very strict. The average mean intersection over union (mIoU) of vanishing line prediction is 98.35% when training on the Tusimple Benchmark dataset. It is good enough for subsequent tasks.

Because only the Tusimple Benchmark dataset uses the accuracy of lane marking points as the evaluation metric directly, the experiments of lane fitting are carried out on the Tusimple Benchmark dataset. The experiment with a plain network that excludes the DULR module and attention is also added. The rest of the settings are the same as Sec. IV-C2. The results of lane marking fitting are shown in Table II.

The fitted results for all of the settings are better than the raw outputs and the results of the least-squares. Moreover, the estimated vanishing line could alleviate the perspective effect from the non-flat ground, as shown in Fig. 10.

4) Computational Cost of Different Steps: The proposed method is composed of several steps. It is necessary to show the detailed computational cost of every step. Because different settings of attention at different places have little effects on computational time, we only use the network that ADLD attention is both added to the input layer and DULR module. We have reduced the font of the formula in the revised version. We use VGG16 as our backbone network with an input size of 800×288, and this experiment is carried out with a single GTX 1080Ti. Therefore, the runtime is computed on the GPU. The results are shown in Table III. We can observe that the generation of ADLD features and the fitting process cost little time compared with the network itself. The proposed ADLD features and fitting process are efficient.

TABLE III
RUNTIME COMPARISON BETWEEN DIFFERENT STEPS. THE RUNTIME OF IPM TRANSFORMATION AND FITTING STEP VARIES WITH DIFFERENT NUMBER OF DETECTED LANE MARKING POINTS.

Step	ADLD Features	Network	IPM & Fitting
Runtime(ms)	0.5	78.4	0.0-2.2

D. Evaluation Results

In this section, we demonstrate the results of the proposed method. For the Caltech lanes dataset, [11] is used for comparison. For the KITTI dataset, RBNet [27], Up-Conv-Poly [25] and SPRAY [14] are used for comparison. For the Tusimple Benchmark dataset [30], [8], [31] and some not published methods are used for comparison. For the CULanes dataset, VGG [41], ResNet [42], Renet [43], DenseCRF [44] and SCNN [30] are used for comparison. For VGG and ResNet, we use them as the backbone network of Deeplab [45]. All of the settings of our method can be seen in Sec. IV-A.

TABLE IV
EXPERIMENTAL RESULTS ON CALTECH LANES DATASET.

Method	Clips	Correct Rate	False Positive
Caltech	Cordova1	97.21	3.00
	Cordova2	96.16	38.38
	Washington1	96.70	4.72
	Washington2	95.13	2.21
Proposed	Cordova1	**98.71**	**0.64**
	Cordova2	**97.25**	**0.21**
	Washington1	**98.75**	**0.78**
	Washington2	**98.89**	**0.00**

1) Caltech Lanes dataset: Because the size of the Caltech Lanes dataset is relatively small, we utilize a pre-trained model on the CULanes dataset to finetune the results on the Caltech Lanes dataset. In this experiment, 80% of the original testing set are randomly selected as the training set, and the remains are used for testing. The results are shown in Table IV. Furthermore, some visual results on the validation test are shown in Fig. 11. The third column is our prediction results,

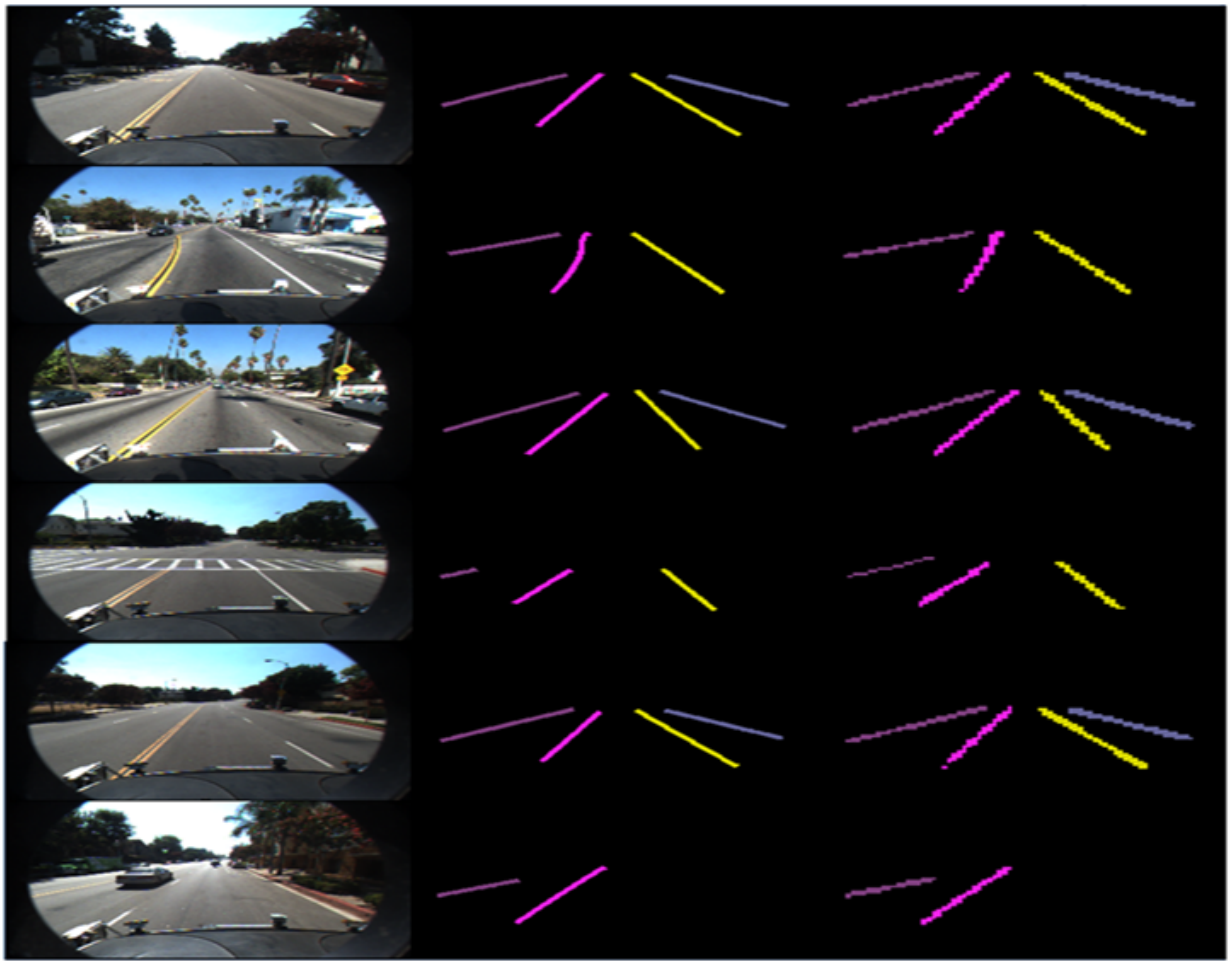


Fig. 11. Visual results on the Caltech Lanes dataset. For each example, from left to right are: input image, ground truth, proposed method. The four colors from left to right are left lane, current lanes, and right lane.

which are jagged due to down-sampling in the network. From the overall fitting effect, it is very close to the ground truth in the second column. Moreover, it can be seen that the prediction even fits the hidden lane in some cases.

From Table IV, we can observe that our method achieves satisfactory results compared with the Caltech method. Meanwhile, the false-positive rates of the proposed method are very low. In fact, there are some false positives in the outer lane's results. In this case, the ability of our method is quite good.

2) KITTI dataset: The results of the KITTI dataset are shown in Table V. Because the ground truth format of the KITTI dataset is the segmentation of road instead of lane markings, which is not compatible with our method, the fitting process and DULR module are removed. The output of the network is modified from the lane markings prediction to the road area segmentation. The proposed method achieves similar results with the first and second methods. Besides, it has a very fast speed.

3) Tusimple Benchmark dataset: The results of the Tusimple Benchmark dataset are shown in Table VI. Because the first method is not published, we do not know whether the authors utilize extra data. Without considering the unknown method, the proposed method achieves promising performance on the Tusimple Benchmark dataset with no extra data. It is also very close to SCNN, which utilizes extra data.

The visualization results can be seen in Fig. 12. From it, we can see that the proposed method shows good performance

TABLE V
LANE ESTIMATION EVALUATION ON KITTI DATASET (%).

Method	**MaxF**	**AP**	**PRE**	**REC**	**FPR**	**FNR**	**Runtime**
NVLaneNet	91.86	91.42	90.89	92.85	1.64	7.15	0.08 s
ILN	91.62	91.17	91.98	91.26	1.40	8.74	0.24 s
Proposed	91.37	91.40	93.08	89.71	1.17	10.29	0.05 s
RBNet	90.54	82.03	94.92	86.56	0.82	13.44	0.18 s
Up-Conv-Poly	89.88	87.52	92.01	87.84	1.34	12.16	0.08 s
SPRAY	83.42	86.84	84.76	82.12	2.60	17.88	0.05s

TABLE VI
TUSIMPLE BENCHMARK LEADERBOARD.

Rank	Method	Published	Extra Data	Accuracy	FP	FN
1	leonardoli	No	N/A	96.87	0.0442	0.0197
2	XingangPan	Yes	Yes	96.53	0.0617	0.0180
3	Proposed	N/A	No	96.51	0.2393	0.0316
4	aslarry	Yes	No	96.50	0.0851	0.0269
6	DavyNeven	Yes	No	96.38	0.0780	0.0244

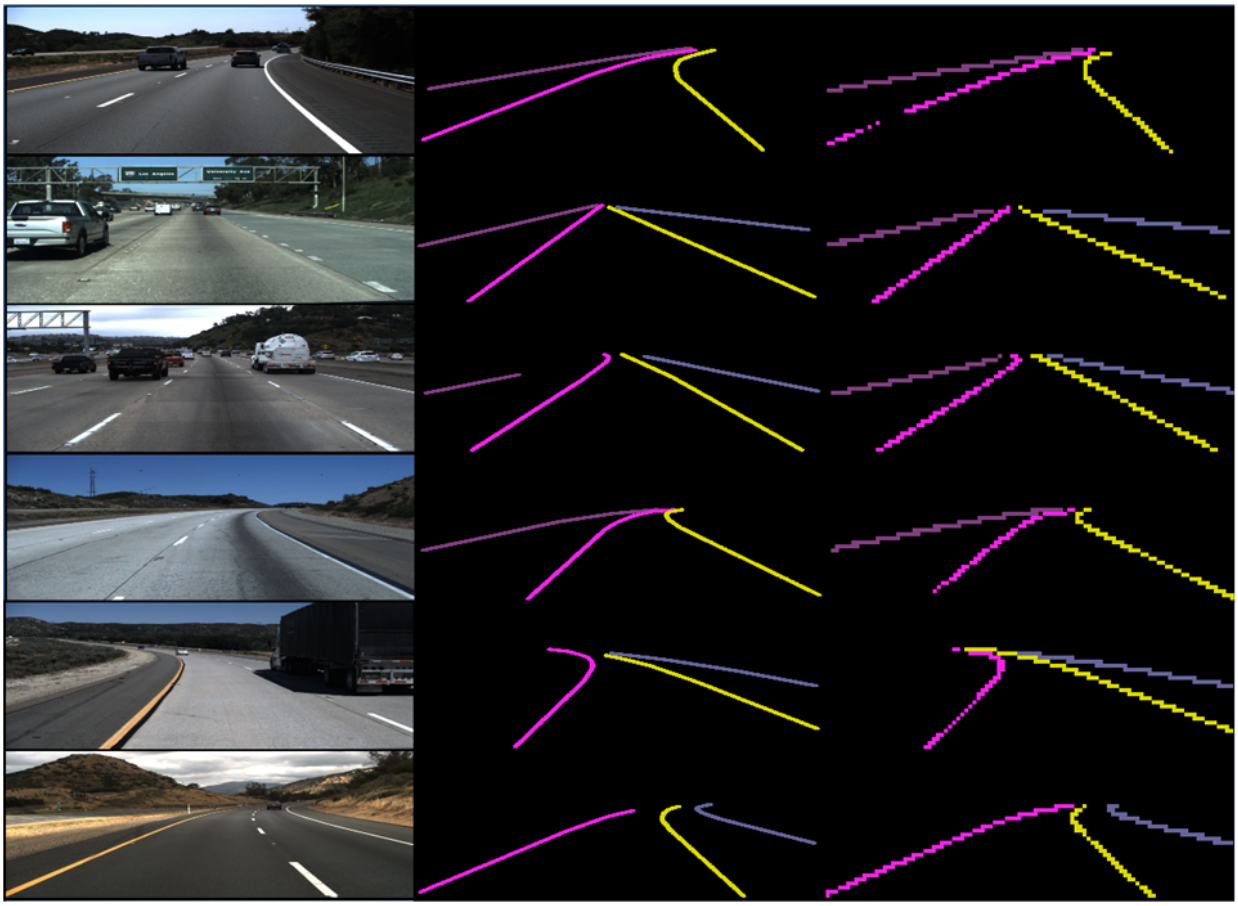


Fig. 12. Visual results on the Tusimple validation set. For each example, from left to right are: input image, ground truth, proposed method. The four colors from left to right are left lane, current lanes, and right lane.

when dealing with sharp curves at the far end part of the road. Meanwhile, the localization ability is better due to the effect of ADLD attention.

4) CULanes dataset: The results of the CULanes dataset are shown in Table VII. The whole test set is divided into 9 categories. As we can see, in most cases, our method achieves the best results. For some tough scenarios like curve and shadow, our method outperforms other methods because the proposed ADLD attention makes the network have better localization ability. For the crossroad, the ADLD attention could filter its features and has fewer false positives. The visualization results are shown in Fig. 13. From the figure, the fitting performance in the CULanse dataset is still excellent. Notice that in the first row, there is no prediction result. The reason is that in this dataset, no prediction is made whenever a zebra crossing is encountered. In this image, there are many zebra crossings, so there are no predictions here.

By testing on four datasets, the advantages of the proposed method are summarized as follows. 1) Reduce the false-positive rates, which is contributed by the ADLD attention module. 2) Because ADLD does not require self-learning, so it makes the detection framework has a faster speed. 3) From the visualization results, the proposed method outperforms others when dealing with sharp curves at the far end of the road. In conclusion, our method takes into account both speed and accuracy. It contributes to the practical application of lane detection.

TABLE VII
COMPARISON ON CULANES DATASET. FOR CROSSROAD, THE NUMBER OF FALSE POSITIVES IS SHOWN.

Category	VGG16	ResNet50	ResNet101	ReNet	DenseCRF	SCNN	Ours
Normal	83.1	87.4	90.2	83.3	81.3	**90.6**	90.2
Crowded	61.0	64.1	68.2	60.5	58.8	**69.7**	**69.7**
Night	56.9	60.6	65.9	56.3	54.2	66.1	**67.3**
No line	34.0	38.1	41.7	34.5	31.9	43.4	**44.7**
Shadow	54.7	60.7	64.6	55.0	56.3	66.9	**68.5**
Arrow	74.0	79.0	84.0	74.1	71.2	84.1	**84.8**
DazzleLight	49.9	54.1	**59.8**	48.2	46.2	58.5	59.7
Curve	61.0	59.8	65.5	59.9	57.8	64.4	**69.6**
Crossroad	2060	2505	2183	2296	2253	1990	**1933**

V. CONCLUSION

In this paper, we propose a novel multi-task attention method for lane marking detection, which combines deep and traditional methods. Specifically, the proposed method improves the localization ability by ADLD attention that could also benefit the performance of segmentation and vanishing line prediction. Moreover, a lane fitting method based on vanishing line prediction is introduced, which benefits the proposed network. For the proposed components such as

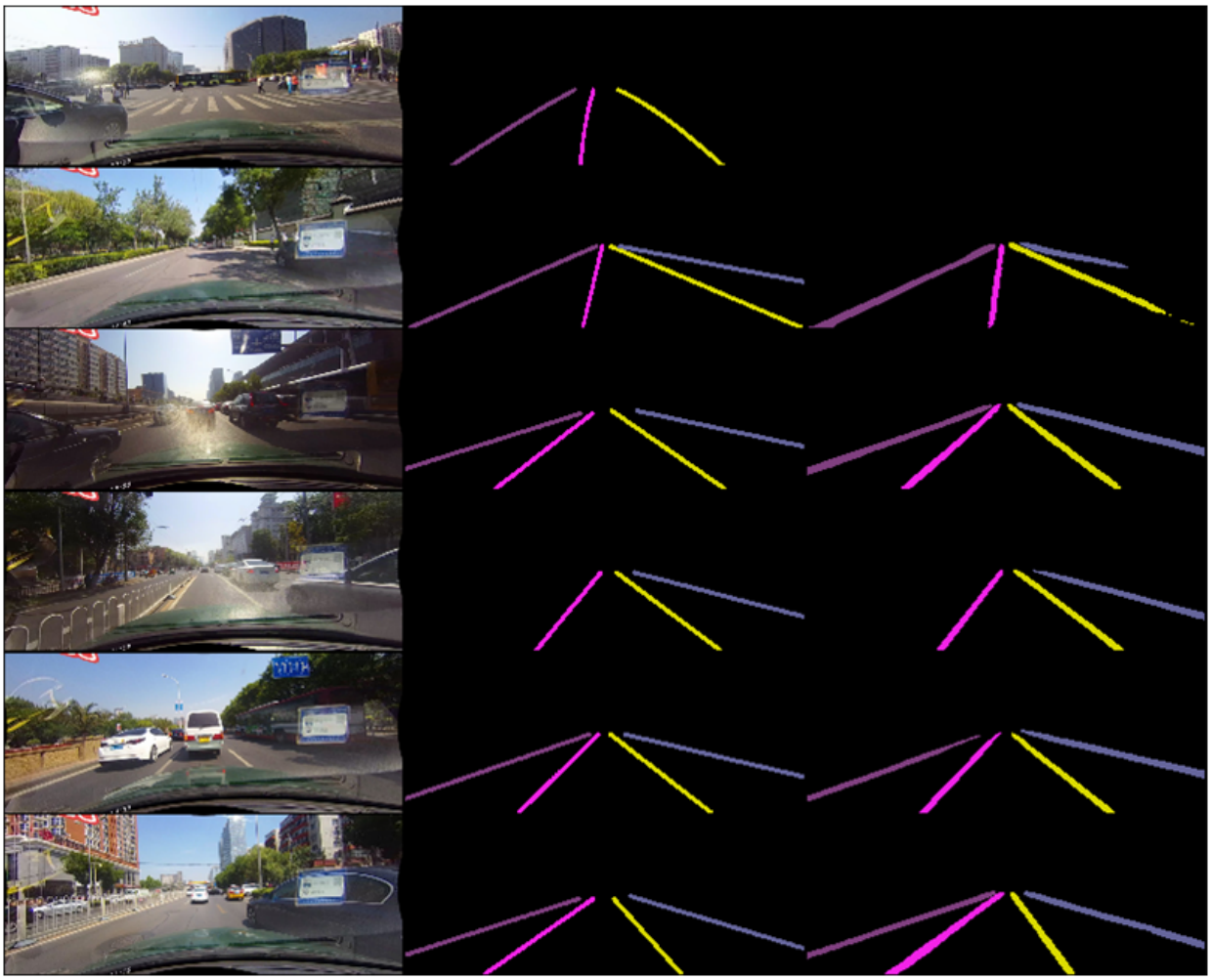


Fig. 13. Visual results on the CULanse dataset. For each example, from left to right are: input image, ground truth, proposed method. The four colors from left to right are left lane, current lanes, and right lane.

ADLD attention, IPM based on vanishing line and lane fitting, comprehensive ablation studies are carried out to verify their effectiveness. These experimental results show the effectiveness of the prosed method. Meanwhile, extensive experiments on four datasets which are Caltech Lanes dataset, KITTI dataset, Tusimple Benchmark dataset and CULanes dataset also demonstrate the effectiveness of our method.

REFERENCES

[1] M. Bertozzi, A. Broggi, and zhitong xiong, "Gold: A parallel real-time stereo vision system for generic obstacle and lane detection," *IEEE Trans. on image processing*, vol. 7, no. 1, pp. 62–81, 1998.

[2] M. Bertozz, A. Broggi, and A. Fascioli, "Stereo inverse perspective mapping: theory and applications," *Image and vision computing*, vol. 16, no. 8, pp. 585–590, 1998.

[3] D. Zhang, B. Fang, W. Yang, X. Luo, and Y. Tang, "Robust inverse perspective mapping based on vanishing point," in *Proc. IEEE International Conference on Security, Pattern Analysis, and Cybernetics*. IEEE, 2014, pp. 458–463.

[4] J. Jeong and A. Kim, "Adaptive inverse perspective mapping for lane map generation with slam," in *Proc. International Conference on Ubiquitous Robots and Ambient Intelligence*. IEEE, 2016, pp. 38–41.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, no. 12, pp. 2481–2495, 2017.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

[8] D. Neven, B. De Brabandere, S. Georgoulis, M. Proesmans, and L. Van Gool, "Towards end-to-end lane detection: an instance segmentation approach," *arXiv preprint arXiv:1802.05591*, 2018.

[9] B. Southall and C. J. Taylor, "Stochastic road shape estimation," in *Proc. IEEE International Conference on Computer Vision*, vol. 1. IEEE, 2001, pp. 205–212.

[10] Y. Wang, E. K. Teoh, and D. Shen, "Lane detection and tracking using b-snake," *Image and Vision computing*, vol. 22, no. 4, pp. 269–280, 2004.

[11] M. Aly, "Real time detection of lane markers in urban streets," in *Intelligent Vehicles Symposium*. IEEE, 2008, pp. 7–12.

[12] S. Zhou, Y. Jiang, J. Xi, J. Gong, G. Xiong, and H. Chen, "A novel lane detection based on geometrical model and gabor filter," in *Intelligent vehicles symposium*. IEEE, 2010, pp. 59–64.

[13] Y. Wang, N. Dahnoun, and A. Achim, "A novel system for robust lane detection and tracking," *Signal Processing*, vol. 92, no. 2, pp. 319–334, 2012.

[14] T. Kühnl, F. Kummert, and J. Fritsch, "Spatial ray features for real-time ego-lane extraction," in *Proc. IEEE Conference on Intelligent Transportation Systems*. IEEE, 2012, pp. 288–293.

[15] Z. Wang, F. Nie, L. Tian, R. Wang, and X. Li, "Discriminative feature

selection via a structured sparse subspace learning module," in *Proc. International Joint Conference on Artificial Intelligence*, 2020, pp. 3009–3015.
[16] J. M. Á. Alvarez and A. M. Lopez, "Road detection based on illuminant invariance," *IEEE Trans. on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 184–193, 2011.
[17] J. Son, H. Yoo, S. Kim, and K. Sohn, "Real-time illumination invariant lane detection for lane departure warning system," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1816–1824, 2015.
[18] J. Hur, S.-N. Kang, and S.-W. Seo, "Multi-lane detection in urban driving environments using conditional random fields," in *Intelligent Vehicles Symposium*. IEEE, 2013, pp. 1297–1302.
[19] A. Borkar, M. Hayes, and M. T. Smith, "A novel lane detection system with efficient ground truth generation," *IEEE Trans. on Intelligent Transportation Systems*, vol. 13, no. 1, pp. 365–374, 2012.
[20] S. Jung, J. Youn, and S. Sull, "Efficient lane detection based on spatiotemporal images," *IEEE Trans. on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 289–295, 2016.
[21] Q. Wang, J. Gao, and Y. Yuan, "A joint convolutional neural networks and context transfer for street scenes labeling," *IEEE Trans. on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1457–1470, 2017.
[22] J. Kim and M. Lee, "Robust lane detection based on convolutional neural network and random sample consensus," in *Proc. International Conference on Neural Information Processing*. Springer, 2014, pp. 454–461.
[23] B. He, R. Ai, Y. Yan, and X. Lang, "Accurate and robust lane detection based on dual-view convolutional neutral network," in *Intelligent Vehicles Symposium*. IEEE, 2016, pp. 1041–1046.
[24] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue *et al.*, "An empirical evaluation of deep learning on highway driving," *arXiv preprint arXiv:1504.01716*, 2015.
[25] G. L. Oliveira, W. Burgard, and T. Brox, "Efficient deep models for monocular road segmentation," in *Proc. IEEE International Conference on Intelligent Robots and Systems*. IEEE, 2016, pp. 4885–4891.
[26] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *Proc. International Conference on Intelligent Transportation Systems*, 2013.
[27] Z. Chen and Z. Chen, "Rbnet: A deep neural network for unified road and road boundary detection," in *Proc. International Conference on Neural Information Processing*. Springer, 2017, pp. 677–687.
[28] S. Lee, I. S. Kweon, J. Kim, J. S. Yoon, S. Shin, O. Bailo, N. Kim, T.-H. Lee, H. S. Hong, and S.-H. Han, "Vpgnet: Vanishing point guided network for lane and road marking detection and recognition," in *Proc. IEEE International Conference on Computer Vision*. IEEE, 2017, pp. 1965–1973.
[29] J. Li, X. Mei, D. Prokhorov, and D. Tao, "Deep neural network for structural prediction and lane detection in traffic scene," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 690–703, 2017.
[30] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial cnn for traffic scene understanding," *arXiv preprint arXiv:1712.06080*, 2017.
[31] Y.-C. Hsu, Z. Xu, Z. Kira, and J. Huang, "Learning to cluster for proposal-free instance segmentation," *arXiv preprint arXiv:1803.06459*, 2018.
[32] Z. Chen, J. Zhang, and D. Tao, "Progressive lidar adaptation for road detection," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, pp. 693–702, 2019.
[33] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "Lidar–camera fusion for road detection using fully convolutional neural networks," *Robotics and Autonomous Systems*, vol. 111, pp. 125–131, 2019.
[34] J. Canny, "A computational approach to edge detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, no. 6, pp. 679–698, 1986.
[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. International Conference on Machine Learning*, 2015, pp. 2048–2057.
[36] Q. Wang, J. Gao, and Y. Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," *IEEE Trans. on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 230–241, 2018.
[37] M. Nieto, L. Salgado, F. Jaureguizar, and J. Cabrera, "Stabilization of inverse perspective mapping images based on robust vanishing point estimation," in *Intelligent Vehicles Symposium*. IEEE, 2007, pp. 315–320.
[38] TuSimple, "Tusimple benchmark," http://benchmark.tusimple.ai/, 2017, accessed June 15, 2018.
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[40] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proc. European Conference on Computer Vision*. Springer, 2018, pp. 270–286.
[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[43] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio, "Renet: A recurrent neural network based alternative to convolutional networks," *arXiv preprint arXiv:1505.00393*, 2015.
[44] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Proc. Advances in Neural Information Processing Systems*, 2011, pp. 109–117.
[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Tao Han received the B.E. degree in transportation equipment and control engineering from the Northwestern Polytechnical University, Xi'an, China, in 2019. he is currently working toward the M.S. degree in computer science and technology in the Center for OPTical IMagery Analysis and Learning, School of Computer Science, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Zequn Qin received the B.E. degree in computer science technology from Northwestern Polytechnical University, Xi'an, China, in 2016. He is currently pursuing the M.E. degree with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His current research intersts include computer vision and machine learning.

Junyu Gao received the B.E. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China, in 2015. He is currently pursuing the Ph.D. degree from Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Xuelong Li (M'02-SM'07-F'12) is a full professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.

APPENDIX A
DERIVATION OF THE IPM BASED ON VANISHING LINE

In this section, we provide a complete derivation of Eq. 5 and 6.

The detailed illustrations of geometry model are shown in Fig. 14 and 15. In order to establish the relations between point $P(P_x, P_y)$ and $P'(P_u, P_v)$, we have:

$$P_y = \overline{OC} tan(\angle P_j CO). \tag{13}$$

The $\angle P_j CO$ can be represented as:

$$\angle P_j CO = \angle P_j CM + \angle MCO, \angle MCO = \frac{\pi}{2} - \theta. \tag{14}$$

From another view, we have:

$$\angle P_j CM = \angle P_u CM'. \tag{15}$$

Denote d_u and d_v are the pixel size, then

$$d_u = \frac{2ftan\alpha}{m-1}, tan(\angle P_u CM') = \frac{(m-1-2P_u)d_u}{2f}. \tag{16}$$

The final representation of P_y can be:

$$P_y = hcot\left\{\theta - atan[tan\alpha(1 - \frac{2P_u}{m-1})]\right\} \tag{17}$$

which is the derivation of Eq. 5. Similarly, we have:

$$P_x = \overline{CP_j} tan(\angle P_j CP) \tag{18}$$

The $\angle P_j CP$ is equal to:

$$\angle P_j CP = \angle P' CP_u. \tag{19}$$

Similar to Eq. 16, we can get:

$$d_v = \frac{2ftan\beta}{n-1}, tan(\angle P_v CM') = \frac{(m-1-2P_v)d_v}{2f}, \tag{20}$$

and

$$tan(\angle P' CP_u) = tan(\angle P_v CM')cos(\angle P_u CM') \tag{21}$$

Substituting Eq. 19, 20 and 21 to Eq. 18, we have:

$$P_x = \sqrt{h^2 + y^2} \frac{tan\beta(\frac{2v}{n-1} - 1)}{\sqrt{1 + [tan\alpha(1 - \frac{2u}{m-1})]^2}}. \tag{22}$$

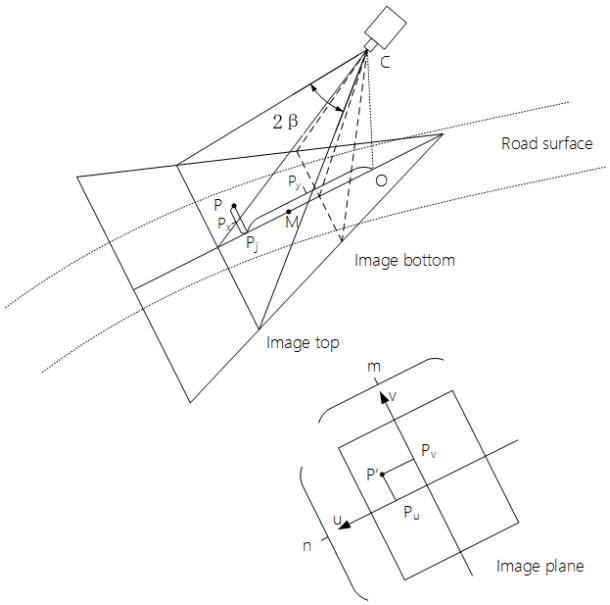


Fig. 14. The detailed geometric model of IPM. β is the horizontal angular aperture of the camera. The point C is the optic center and the point O is the projection of C on the ground. M is the crossover point of optic axis and the ground. P_j is the projection of point P along the optic axis and the ground. m and n are the size of image.

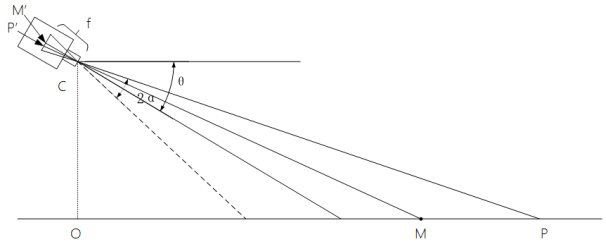


Fig. 15. The detailed side view of the IPM Geometric model. θ is the pitch angle and α is the vertical angular aperture of the camera. The definitions of point O, M and P are the same as Fig. 14. f is the focal distance. M' and P' are the image point of M and P.