BMVC 2022

Spatio-Temporal Fusion-based Monocular 3D Lane Detection

Yin Wang[1], Qiuyi Guo[2], Peiwen Lin[2], Guangliang Cheng[2], Jian Wu[1]

[1] Jilin University
Changchun, China
[2] SenseTime Research

Background

Lane detection is a classic yet challenging computer vision task and an intuitive and effective way for autonomous vehicles to perceive their surroundings. Traditional 2D lane detection methods focus on obtaining the exact lane position on the image. However, due to the lack of image depth information, the 2D results are challenging to apply to downstream tasks directly. Monocular 3D lane detection methods are proposed to solve these issues, which can obtain the lane position from the 3D world space in an end-to-end manner without road planar assumption.

Motivation & Introduction

We aim to provide a **simple** and **spatio-temporal** fusion-based 3D lane detection method abbreviated as **STLane3D**.

- We propose a **multi-frame fusion** mechanism by making use of the strong spatio-temporal continuity in consecutive frames instead of focusing on single frame.
- We utilize a **pre-alignment operation** to align the space information on the different frames. For example, some parts of lanes are invisible in frame *t* - 1 or frame *t* due to the occlusion of surrounding vehicles in Fig .1(a). By fusing multiple frames, the position information can be recognized in Fig .1(b), marked by the green circles.
- We perform 3D lane detection under the **camera coordinate system**.

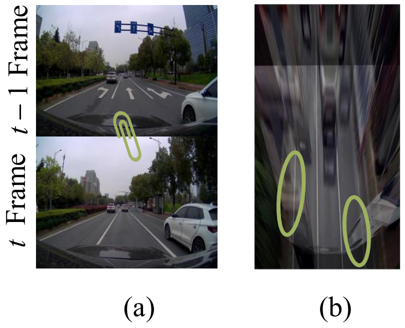

Fig 1 (a) the original picture of frame *t* - 1 and frame *t* in same sequence, (b) the two frames after alignment under BEV.

Overview

The overall paradigm of our method is shown in Fig. 2.

- **Encoder layer** We adopt ResNext-50 combined with RESA block as backbone. The BEV converter projects the extracted features to the BEV plane through bilinear interpolation. The encoder is shared at different timestamps to reduce inference time by saving the feature map of historical frames.
- **Fusion layer** We use channel-wise convolution to fuse the information of historical and current frames as key-value pairs . We use the ego-motion from RTK to pre-align the features of the historical frame before BEV pre-alignment.
- **Decoder layer** We decode the feature by the 3D lane head through a series of convolutions with no padding in the *y* dimension. For each preset anchor, we predict, (1) lateral position offset dx relative to the anchor, (2) height *z* relative to the virtual plane, (3) visibility *vis* of the sampled point, (4) category *type* of the lane.

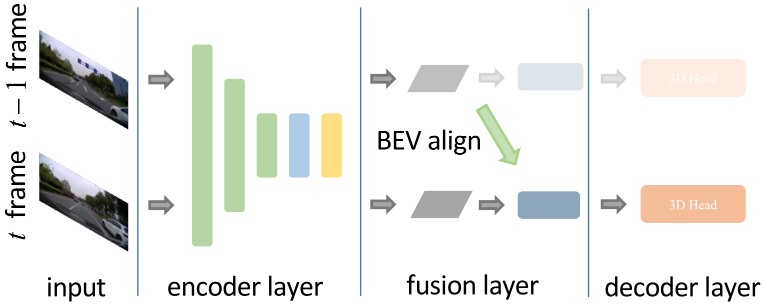

Fig 2 The overall paradigm of STLane3D.

Results

As shown in Tab. 1, we compare the proposed STLane3D with 3DLaneNet, GenLaneNet, SALAD and PersFormer. and achieved the new state-of-art F1-score of 77.53%.

Tab 1 Comparison with the state-of-the-art methods.

Methods	F1-score(%)↑	Precision(%)↑	Recall(%)↑	CD-error(m)↓
3DLaneNet [4]	44.73	61.46	35.16	0.127
GenLaneNet [5]	45.59	63.95	35.42	0.121
SALAD [10]	64.07	75.90	55.42	0.098
PersFormer [3]	74.33	80.30	69.18	0.074
STLane3D-single(ours)	74.05	76.63	71.64	0.085
STLane3D-multi(ours)	**77.53**	**81.54**	**73.90**	**0.066**

Fig. 3 shows a scene where the lane line is obscured by a vehicle on the left. Part of the lane line is not visible in current frame due to the occlusion of the left car, while this lane is observable from history frames.

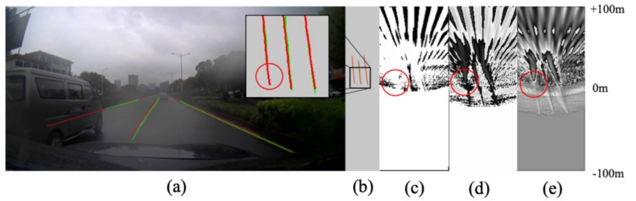

Fig 3 Illustration of an obscured scene.