End-to-end Lane Shape prediction with Transformers

Ruijin Liu   Zejian Yuan   Tie Liu   Zhiliang Xiong

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

College of Information Engineering, Capital Normal University, China

Shenzhen Forward Innovation Digital Technology Co. Ltd, China

WACV 2021

Speaker: Wencong Zhang

Background

- Vision-based lane marking detection is a fundamental module in autonomous driving
- The popular pipeline that solves it in two steps—feature extraction plus post-processing, is too inefficient and flawed in learning the global context and lanes' long and thin structures.

Contributions

- reframe the lane detection output as parameters of a lane shape model

    parameters are derived from a lane shape model which models the road structures and the camera pose. Which have explicit physical meanings

- develop a network built with transformer blocks to reinforce the learning of global context and lanes’ slender structures.

$$X = kZ^3 + mZ^2 + nZ + b, \tag{1}$$

$$u = \frac{k'}{v^2} + \frac{m'}{v} + n' + b' \times v, \tag{2}$$

$$\begin{aligned} u' = &\frac{k' \times \cos^2\phi}{(v' - f\sin\phi)^2} + \frac{m'\cos\phi}{(v' - f\sin\phi)} + n' \\ &+ \frac{b' \times v'}{\cos\phi} - b' \times f\tan\phi, \end{aligned} \tag{3}$$

$$u' = \frac{k''}{(v' - f'')^2} + \frac{m''}{(v' - f'')} + n' + b'' \times v' - b''', \tag{4}$$

$$g_t = (k'', f'', m'', n', b''_t, b'''_t, \alpha_t, \beta_t)$$

Method

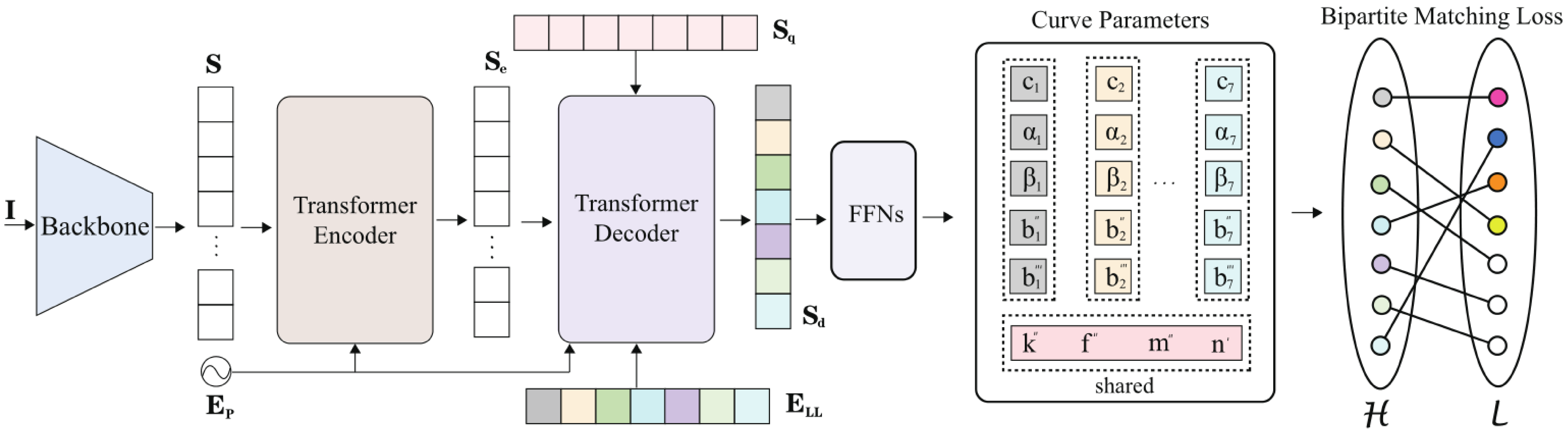


Figure 1. Overall Architecture. The $\mathbf{S}$, $\mathbf{S_e}$ and $\mathbf{E_p}$ indicate flattened feature sequence, encoded sequence and the sinusoidal positional embeddings which are all tensors with shape $HW \times C$. The $\mathbf{S}_q$, $\mathbf{E_{LL}}$ and $\mathbf{S_d}$ represent query sequence, learned lane embedding and the decoded sequence which are all in shape $N \times C$. Different color indicate different output slots. White hollow circles represent ”non-lanes”.

Method

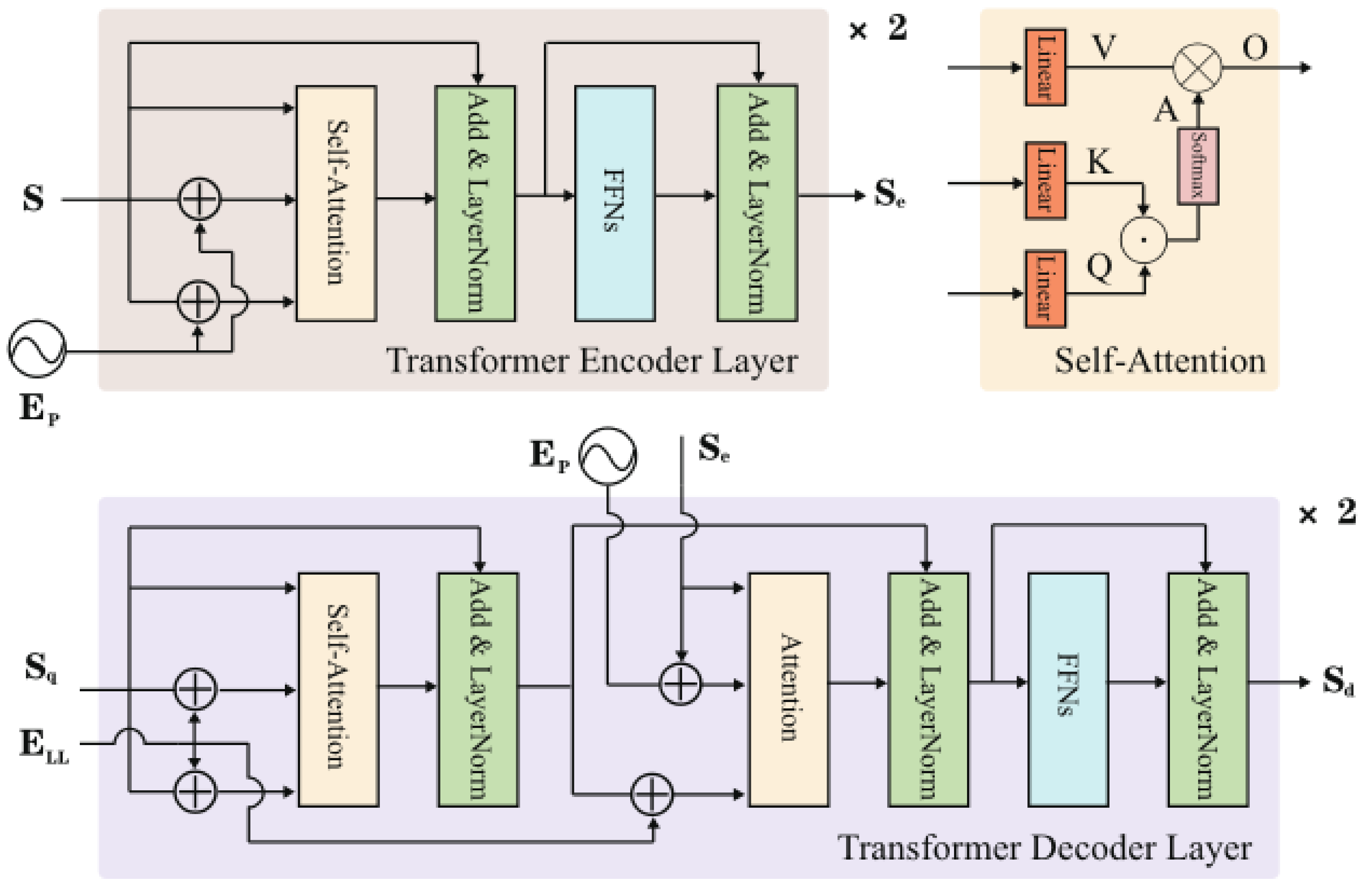


Figure 2. Transformer Encoder and Decoder. The ⊕ and ⊙ represent matrix addition and dot-product operations respectively.

×2

×2

Hungarian Fitting Loss

$$L = \sum_{i=1}^{N} - \omega_1 \log p_{\hat{z}(i)} \left(\hat{c}_i \right) + \mathbb{1} \left(\hat{c}_i = 1 \right) \omega_2 L_1 \left(\hat{\mathbf{s}}_i, \mathbf{s}_{\hat{z}(i)} \right)$$
$$+ \mathbb{1} \left(\hat{c}_i = 1 \right) \omega_3 L_1 \left(\hat{\alpha}_i, \alpha_{\hat{z}(i)}, \hat{\beta}_i, \beta_{\hat{z}(i)} \right), \tag{8}$$

z (i): index of predicted curve

Pz(i)(ci): the probability of class ci (0:non-lane;1:lane)

Sz(i): fitting lane sequece

α, β: vertical starting and ending offset

Ablation studies

Investigation of Shape Model

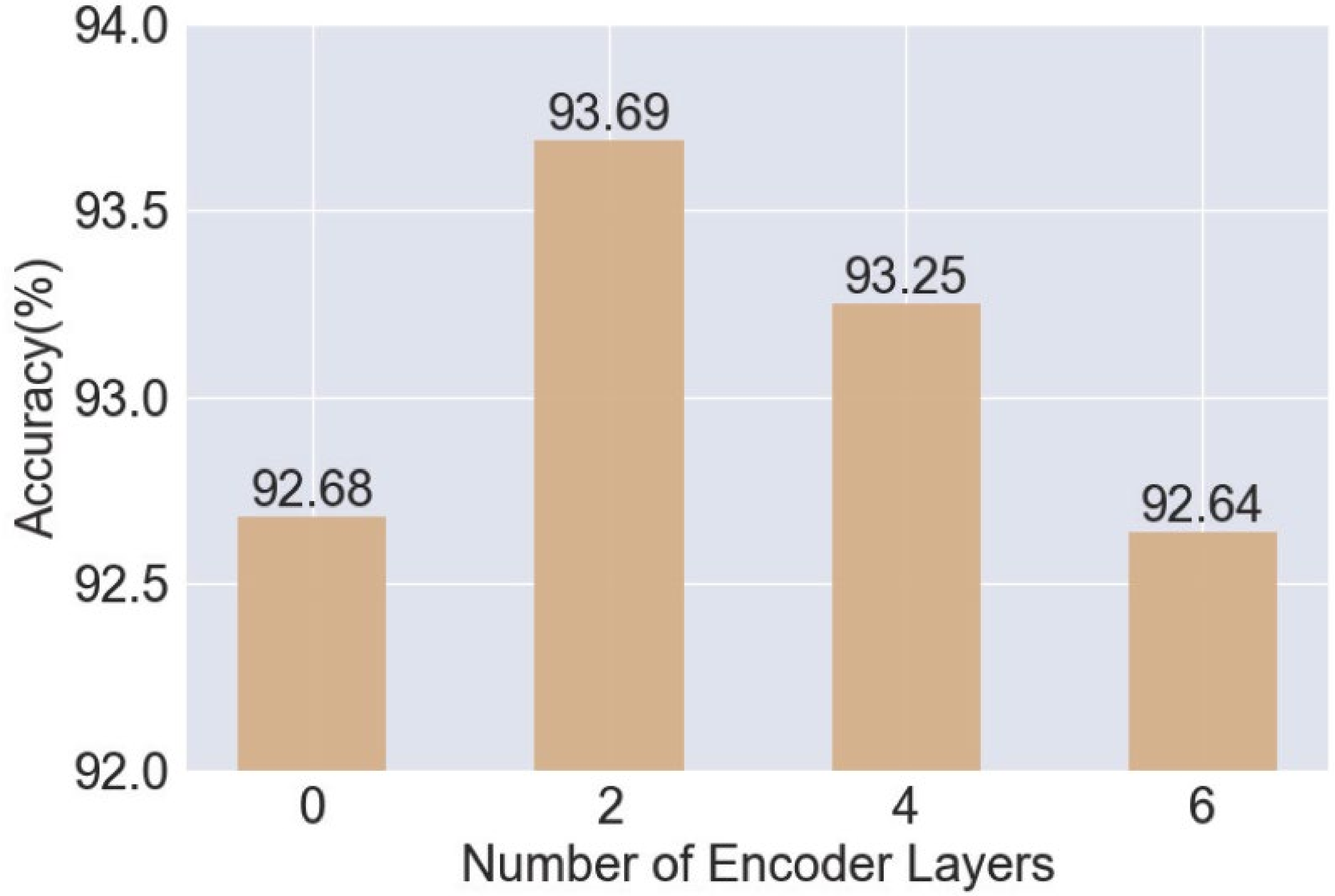
Table 2. Quantitative evaluation of different shape models on TuSimple validation set (%).

Curve Shape	Consistency	Acc	FP	FN
Quadratic	-	91.94	0.1169	0.0975
Quadratic	✓	93.18	0.1046	0.0752
Cubic	-	92.64	0.1068	0.0868
Cubic	✓	**93.69**	**0.0979**	**0.0724**

$$g_t = (k'', f'', m'', n', b_t'', b_t''', \alpha_t, \beta_t)$$

Ablation studies

Number of encoder layers.



(a) Effects of Different Encoder Sizes

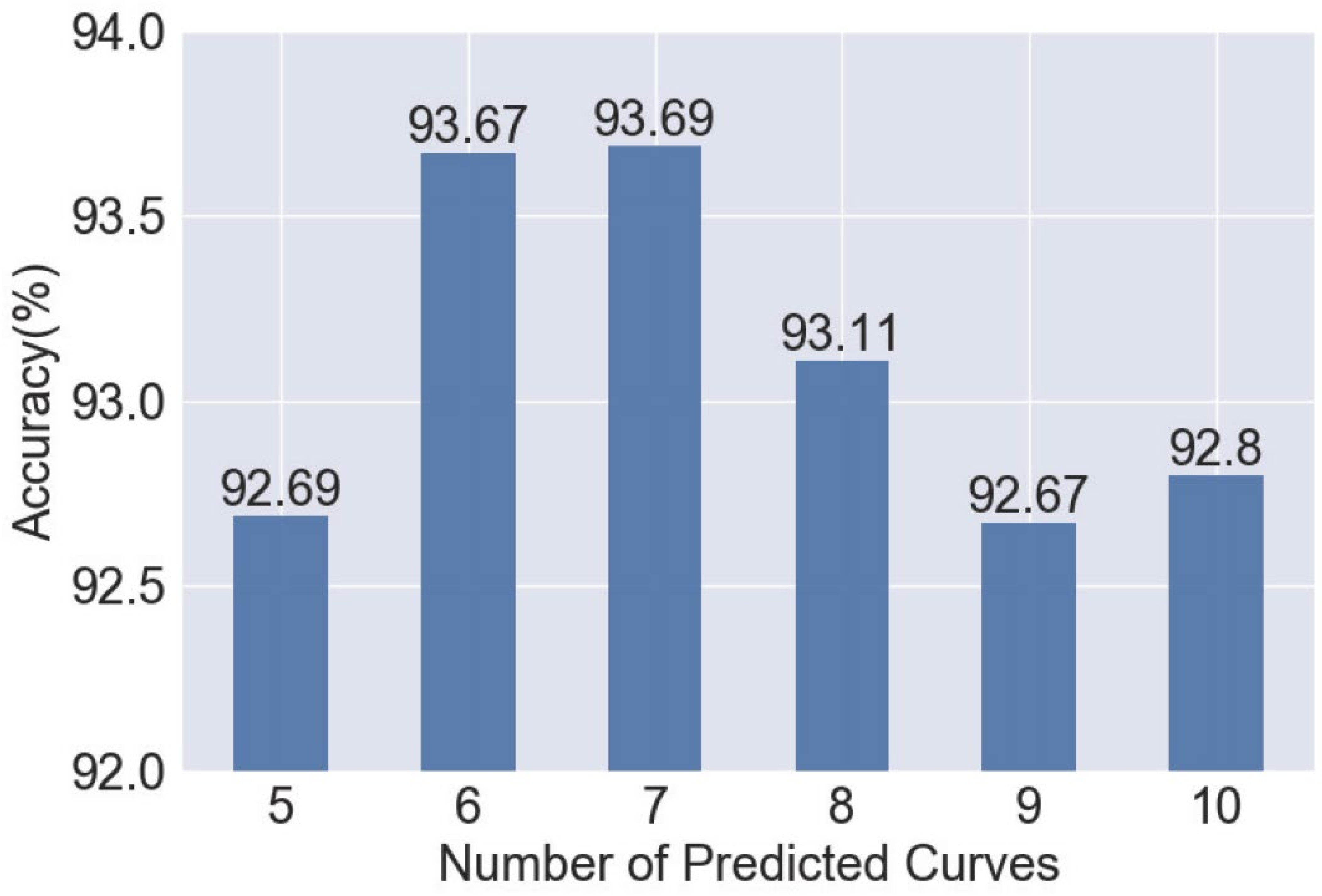
Number of decoder layers.

Table 3. Quantitative evaluation of decoder size and different decoder layer on TuSimple validation set (%). The encoder size is set to be 2.

Size \ Layer	1	2	3	4	5	6
2	93.55	**93.69**	-	-	-	-
4	92.52	93.08	93.15	**93.15**	-	-
6	92.70	93.07	93.05	93.13	93.14	**93.16**

Ablation studies

Number of predicted curves.



(b) Effects of Different Output Quantities

Experiments

Table 1. Comparisons of accuracy (%) on TuSimple testing Set. The number of multiply-accumulate (MAC) operations is given in G. The number of parameters (Para) is given in M (million). The PP means the requirement of post-processing.

Method	FPS	MACs	Para	PP	Acc	FP	FN
FastDraw [15]	90	-	-	✓	95.20	.0760	.0450
SCNN [14]	7	-	20.72	✓	96.53	.0617	**.0180**
ENet-SAD [6]	75	-	0.98	✓	96.64	.0602	.0205
PINet [7]	30	-	4.39	✓	96.70	**.0294**	.0263
Line-CNN [8]	30	-	-	-	**96.87**	.0442	.0197
PolyLaneNet [18]	115	1.784	4.05	-	93.36	.0942	.0933
Ours	**420**	**0.574**	**0.77**	-	**96.18**	**.0291**	**.0338**

Experiments

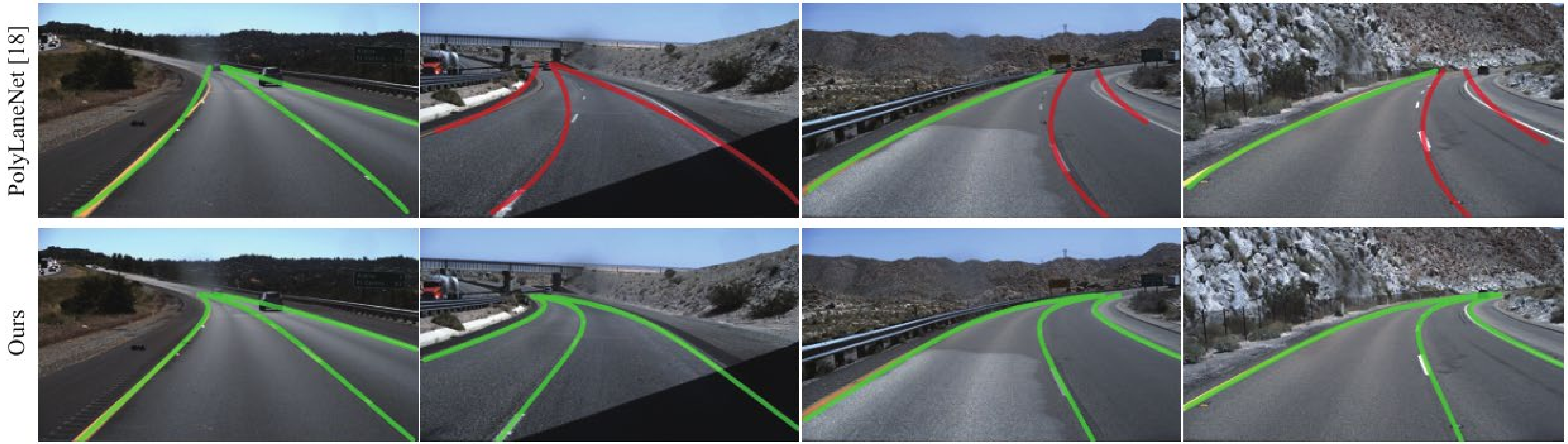


Figure 3. Qualitative comparative results on TuSimple test set. The first row visualizes the predicted curves by the best model of officially public PolyLaneNet resources (red curves means these predictions are mismatched). The second row visualizes our predictions.

Experiments

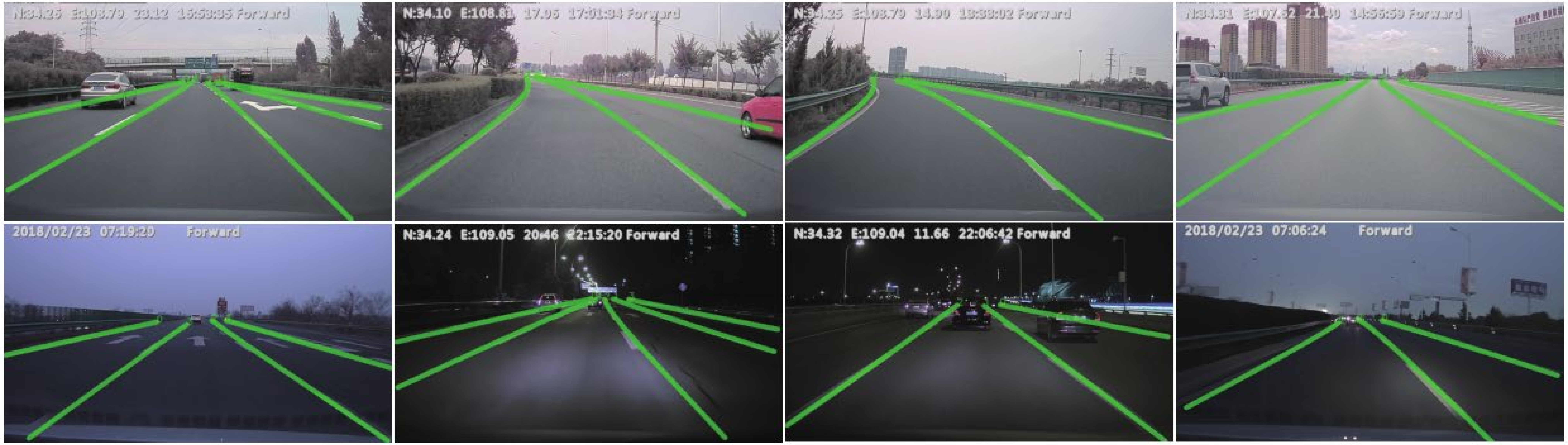


Figure 7. Qualitative transfer results on FVL dataset. Our method even estimates exquisite lane lines without ever seeing the night scene.

Thanks