Supplementary Material: Gen-LaneNet

Yuliang Guo*, Guang Chen, Peitao Zhao, Weide Zhang,
Jinghao Miao, Jingao Wang, and Tae Eun Choe

Baidu Apollo, Sunnyvale CA 94089, USA

Abstract. The supplementary material is organized as follows:
1. Algebraic Verification of the Geometric Transformation.
2. Features Investigation of Issue Associated 3D-LaneNet.
3. Experiments on Center Lines.
4. Qualitative Comparison.

1 Algebraic Verification of the Geometric Transformation

In this section, we present an algebraic derivation of the geometric transformation between 3D ego-vehicle coordinate frame and virtual top-view coordinate frame. Although a general derivation from geometric perspective has been presented in our main paper, we present this algebraic derivation as a double-verification. The derivation considers a simpler camera setup when only pitch angle is involved in camera orientation.

A 3D point (x, y, z) in ego-vehicle coordinate frame can be projected to a 2D image point (u, v) through a projective transformation. Its corresponding 2D point $(\bar{x}, \bar{y})$ in top-view coordinate frame can be projected to the same 2D image point through a planer homography. Given R as the rotation matrix, T as the translation vector, K indicating camera intrinsic parameters, the described relationship can be written as:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \alpha_1 K \begin{bmatrix} R\ T \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \alpha_2 K \begin{bmatrix} R_{1:2}\ T \end{bmatrix} \begin{bmatrix} \bar{x} \\ \bar{y} \\ 1 \end{bmatrix}, \tag{1}$$

where $R_{1:2}$ indicates the first two columns of R, and α_1, α_2 are two constant coefficients. Given camera height h, pitch angle θ, we can explicitly write R, T as:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\sin\theta & -\cos\theta \\ 0 & \cos\theta & -\sin\theta \end{bmatrix}, T = \begin{bmatrix} 0 \\ \cos\theta \cdot h \\ \sin\theta \cdot h \end{bmatrix}. \tag{2}$$

Given simplified notation of $\sin\theta$ as s, $\cos\theta$ as c, we can rewrite Equation 1 as:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -s & -c & ch \\ 0 & c & -s & sh \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \alpha \begin{bmatrix} 1 & 0 & 0 \\ 0 & -s & ch \\ 0 & c & sh \end{bmatrix} \begin{bmatrix} \bar{x} \\ \bar{y} \\ 1 \end{bmatrix}.$$

This equitation can expended as:

$$x\begin{bmatrix}1\\0\\0\end{bmatrix}+y\begin{bmatrix}0\\-s\\c\end{bmatrix}+z\begin{bmatrix}0\\-c\\-s\end{bmatrix}+\begin{bmatrix}0\\ch\\sh\end{bmatrix}= \alpha\bar{x}\begin{bmatrix}1\\0\\0\end{bmatrix}+\alpha\bar{y}\begin{bmatrix}0\\-s\\c\end{bmatrix}+\alpha\begin{bmatrix}0\\ch\\sh\end{bmatrix},$$

which further leads to three equations in scalars:

$$x = \alpha\bar{x} \tag{3}$$

$$-sy - cz + ch = -\alpha s\bar{y} + \alpha ch \tag{4}$$

$$cy - sz + sh = \alpha c\bar{y} + \alpha sh. \tag{5}$$

Reorganizing Equation 5 in

$$y = \alpha\bar{y} + \frac{s}{c}(\alpha h - h + z), \tag{6}$$

and substituting Equation 6 with y in Equation 4, we derive α step-by-step:

$$\begin{aligned}-\alpha s\bar{y} + \alpha ch &= -s\big[\alpha\bar{y} + \frac{s}{c}(\alpha h - h + z)\big] - cz + ch\\ \Longrightarrow \alpha ch &= -\alpha\frac{s^2h}{c} + \frac{s^2(h-z)}{c} + c(h-z)\\ \Longrightarrow \frac{s^2+c^2}{c}\alpha h &= \frac{s^2+c^2}{c}(h-z)\\ \Longrightarrow \alpha &= \frac{h-z}{h}\end{aligned} \tag{7}$$

Substituting Equation 7 with α in Equation 3 and Equation 6 respectively, we at last derive the equations:

$$x = \bar{x}\cdot\frac{h-z}{h} \tag{8}$$

$$y = \bar{y}\cdot\frac{h-z}{h}, \tag{9}$$

So far we have algebraically derived the geometric transformation between 3D coordinate frame and virtual top-view coordinate frame. The transformation agrees with the geometric proof presented in the main paper.

2 Features Investigation of 3D-LaneNet

As mentioned in the paper, 3D-LaneNet [1] represent anchor points in an inappropriate coordinate frame such that visual features can not be aligned with the

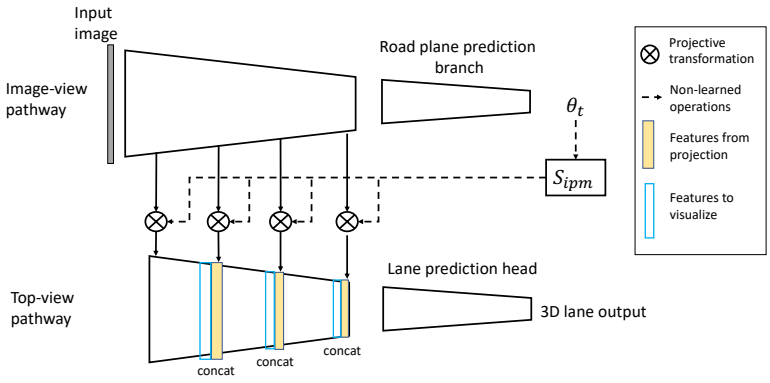


Fig. 1. 3D-LaneNet: An overview pipeline of 3D-LaneNet. The whole network can be decomposed into four sub-networks: image-view pathway, road plane predication branch, top-view pathway and lane prediction head.

prepared ground-truth lane line. We further verify this issue by investigating the key features.

As illustrated in Fig. 1, the network of 3D-LaneNet processes information in two pathways: The *image-view pathway* processes and preserves information from the image while the *top-view pathway* processes features in top-view and uses them to predict the 3D lane output. Information flows from image-view pathway to the top-view pathway through four projective transformation layers.

To confirm the alignment between top-view features and the ground-truth lane, we choose to visualize feature from a few key layers of the top-view pathway, which are marked in blue in Fig. 1. As illustrated in Fig. 2, for a uphill road, image lanes projected to the virtual top-view are expected to appear diverging rather than parallel with each other. When the features from the key layers are visualized, we can observe the same diverging appearance in the feature space. However, the anchor representation from 3D-LaneNet would provide parallel ground-truth lines, which could not align with the diverging features. Although the network learns to focus on lanes, where features are high-lighted, the network can not deform the features to their targeting positions internally. As a result, the misalignment between visual features and ground truth makes the method not generalizable to unobserved scenes.

3 Experiments on Center Lines

Similar to the evaluation of lane line prediction, we conduct evaluation on center line prediction from three perspectives: the effect of new anchor representation, Table 1; the upper bound of two-stage framework, Table 2; and the whole system comparison, Table 3. A candidate method is also evaluated under three different splits of dataset: *Balanced scenes*, *Rarely observed scenes*, and *Scenes with*

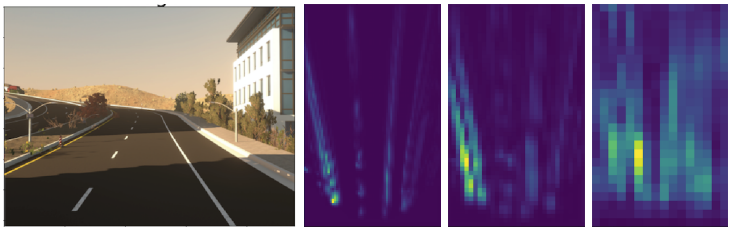


Fig. 2. 3D-LaneNet feature visualization: Given an image captured on a uphill road, imaged lanes are suppose to appear diverging when projected to the virtual top-view. As observed from the visualized features from those blue-marked layers in Figure 1, **top-view features** also form diverging lines. Consequently, diverging visual features will not be in alignment with the parallel lane lines prepared as ground-truth.

visual variations. Observe from the evaluation of center line prediction, similar conclusion can be drawn compared to the evaluation of the lane line prediction.

		balanced scenes			**rarely observed**			**visual variations**		
		w/o	w/	gain	w/o	w/	gain	w/o	w/	gain
3D-LaneNet	**F-score**	89.5	93.3	+3.8	77.0	84.1	+7.1	75.5	86.6	+11.1
	AP	91.4	95.5	+4.1	80.0	85.9	+5.9	77.7	88.7	+11.0
3D-GeoNet	**F-score**	91.2	94.5	+3.3	79.7	85.9	+6.2	87.9	92.3	+4.4
	AP	93.2	96.8	+3.6	83.0	87.7	+4.7	90.6	94.2	+3.6
Gen-LaneNet	**F-score**	88.2	90.8	+2.6	76.1	79.5	+3.4	84.2	88.2	+4.0
	AP	90.8	92.6	+1.8	79.4	80.6	+1.2	87.0	90.0	+3.0

Table 1. (Center line) Comparison of anchor representations. "w/o" represents the integration with anchor design in [1], while "w" represents the integration with our anchor design. For convenience, we also shows the performance gain by integrating our anchor design.

Anchor Effect: As shown in Table 1, the introduction of our new anchor leads to consistent improvement over all candidate methods and over all splits of dataset. Substantial improvement can be observed on 3D-LaneNet on *rarely observed scenes* and *scenes with visual variations* with 7.1% and 11.1% improvements in F-score respectively. This observation verifies the importance of the new anchor and prove that establishing alignment between visual features and lane labels help generalization to unobserved scenes of visual appearance.

The Upper Bound of the Two-Stage Framework: As shown in Table 2, the two-stage framework Gen-LaneNet appears superior to the end-to-end learned method 3D-LaneNet in all three splits of dataset. 3D-GeoNet achieve the highest performance in all cases which shed a light on the upper bound of Gen-LaneNet given perfect image segmentation. Specifically, the margin between 3D-LaneNet and 3D-GeoNet can be significantly large, 16% in both F-score and AP, on scenes with visual variations. Meanwhile, Gen-LaneNet is shown to gain

		balanced scenes	rarely observed	visual variations
3D-LaneNet	**F-score**	89.5	77.0	75.5
	AP	91.4	80.0	77.7
3D-GeoNet	**F-score**	**94.5**	**85.9**	**92.3**
	AP	**96.8**	**87.7**	**94.2**
Gen-LaneNet	**F-score**	**90.8**	**79.5**	**88.2**
	AP	**92.6**	**80.6**	**90.0**

Table 2. (Center line) The upper bound of the two-stage framework. 3D-GeoNet shows potential improvement on Gen-LaneNet when a better image segmentation algorithm is integrated.

significantly when provided with more available 2D labels and better segmentation network.

Whole System Comparison: As observed from Table 3, Gen-LaneNet surpasses 3D-LaneNet over all splits of dataset. The most significant improvement appears under scenes with visual variations (13% in both AP and F-score), where the 3D labels have not included certain illumination but 2D labels have. Besides F-score and AP, x, z errors from close range (0-40 m) and far range (40 - 80 m) are also reported. Although Gen-LaneNet compute these errors over more matched pairs of predicted lanes and ground-truth lanes, the localization errors of its result are maintained lower or on par with 3D-LaneNet.

Dataset Splits	Method	F-Score	AP	x error near (m)	x error far (m)	z error near (m)	z error far (m)
balanced scenes	3D-LaneNet	89.5	91.4	0.066	0.456	0.015	0.179
	Gen-LaneNet	90.8	92.6	0.055	0.457	0.011	0.176
rarely observed	3D-LaneNet	77.0	80.0	0.162	0.883	0.040	0.557
	Gen-LaneNet	79.5	80.6	0.121	0.885	0.026	0.547
visual variations	3D-LaneNet	75.5	77.7	0.120	0.636	0.030	0.227
	Gen-LaneNet	88.2	90.0	0.072	0.438	0.015	0.187

Table 3. (Center line) Whole system comparison between 3D-LaneNet [1] and Gen-LaneNet.

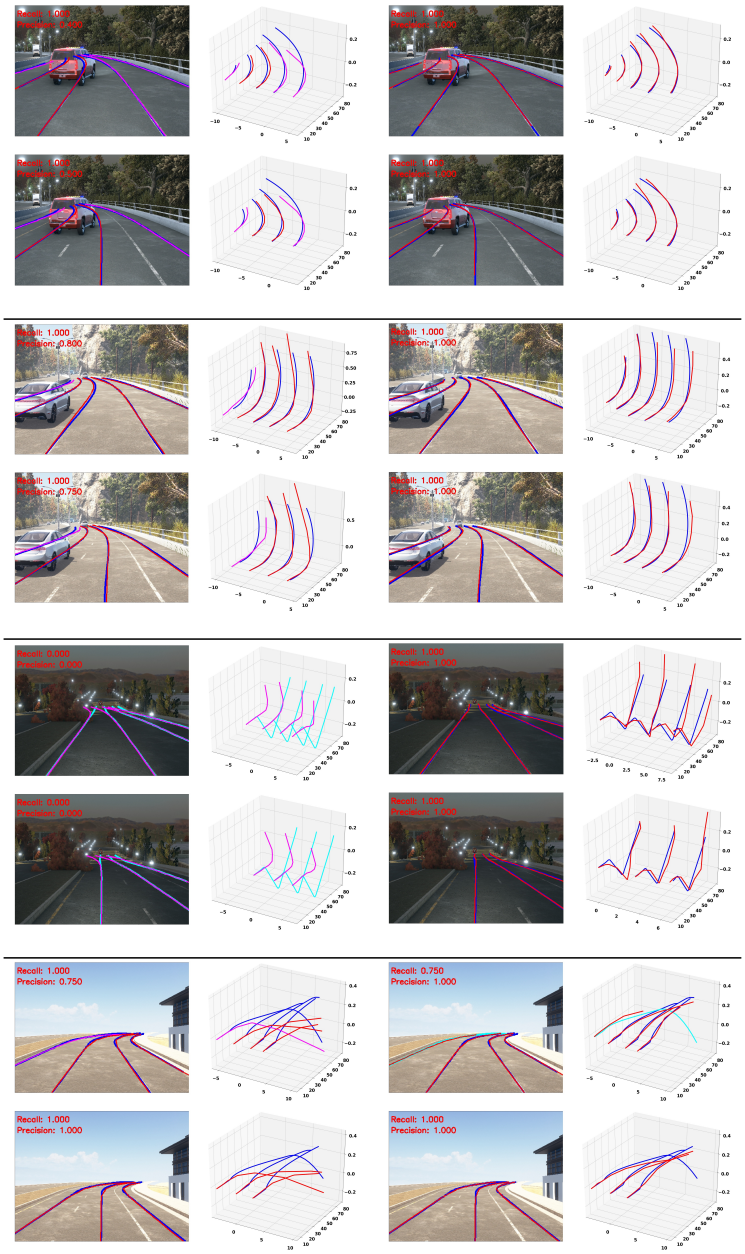
4 Qualitative Comparison

We provide qualitative comparison of both lane lines and center lines. For each example, lane line results are shown in the top row, and center line results are shown in the bottom row. The matching result between the detection and the ground-truth is color-coded: the recovered ground truth in blue; the correct detection in red; the missed ground-truth in purple; and the false-alarm detection in cyan. The visual comparisons are conduct in two sets. First, we compare the original 3D-LaneNet and its improved version adopting our new anchor. This

set of comparison is meant to emphasize the effect of our new anchor. The visualized examples are selected from the test set of the standard five-fold split of dataset. Observed from Figure 3, the new anchor leads to consistency improvement over hilly and sharp-turning roads. Second, we present visual comparison of 3D-LaneNet and Gen-LaneNet as whole systems. The examples are chosen from the split of dataset considering seances with visual variation. As observed from Figure 4, 3D-LaneNet can be very unstable encountering unobserved illumination, however Gen-LaneNet is rather robust.

References

1. Garnett, N., Cohen, R., Pe'er, T., Lahav, R., Levi, D.: 3d-lanenet: end-to-end 3d multiple lane detection. In: IEEE International Conference on Computer Vision, ICCV (2019)



3D-LaneNet 3D-LaneNet (new anchor)

Fig. 3. Effect of the new anchor. Predicted lanes from 3D-LaneNet and from the extended version with our new anchor are visually compared. Examples are chosen from the test set given the standard five-fold split of the whole dataset. Observe that adopting our new anchor consistently improves the localization of laneline and in turn leads to better prediction.

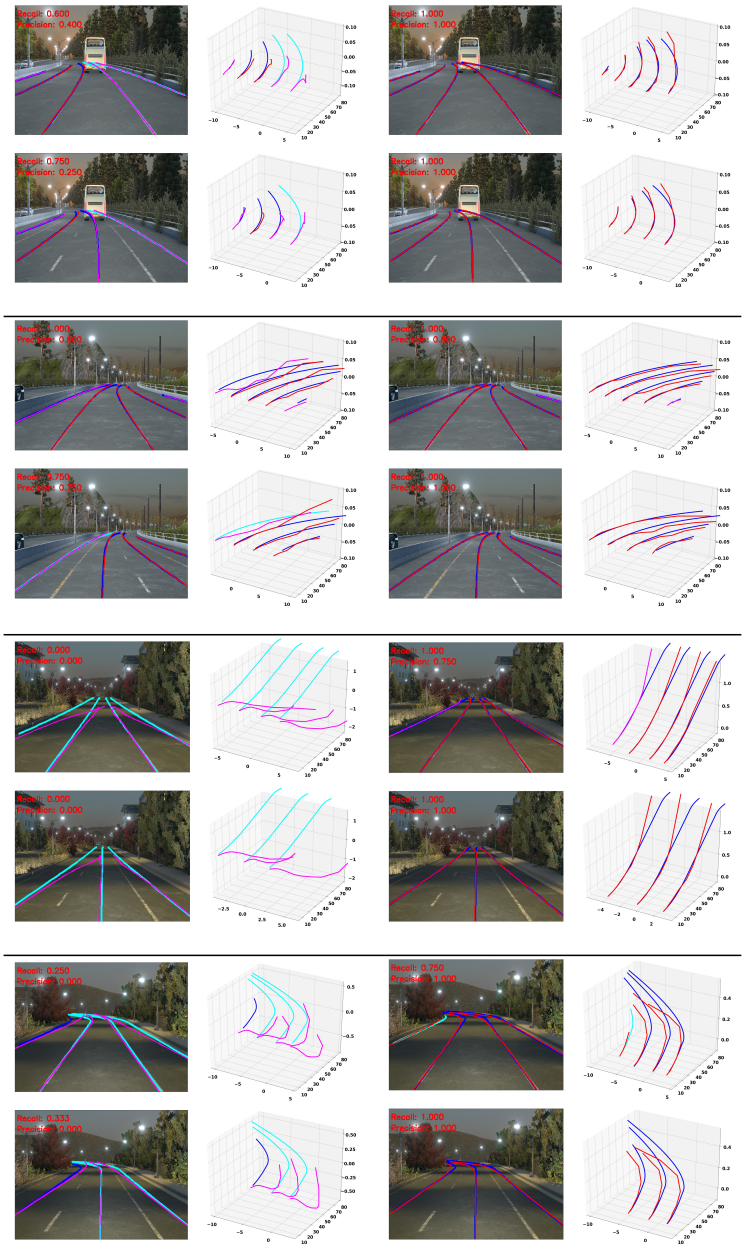


Fig. 4. Visual comparison between **3D-LaneNet and Gen-LaneNet** are show on four examples. Examples are chosen from the data split evaluating an algorithm's robustness to illumination change. Observe that 3D-LaneNet is rather sensitive to illumination change while Gen-LaneNet is not.