Reconstruct from Top View: A 3D Lane Detection Approach based on Geometry Structure Prior

Chenguang Li[*1], Jia Shi[*†1,2], Ya Wang[†1,3], Guangliang Cheng[‡1,4]
[1]SenseTime Research [2]Robotics Institute, Carnegie Mellon University
[3]University of Tuebingen [4]Shanghai AI Laboratory
{lichenguang, wangya}@senseauto.com jiashi@andrew.cmu.edu guangliangcheng2014@gmail.com

Abstract

In this paper, we propose an advanced approach in targeting the problem of monocular 3D lane detection by leveraging geometry structure underneath the process of 2D to 3D lane reconstruction. Inspired by previous methods, we first analyze the geometry heuristic between the 3D lane and its 2D representation on the ground and propose to impose explicit supervision based on the structure prior, which makes it achievable to build inter-lane and intra-lane relationships to facilitate the reconstruction of 3D lanes from local to global. Second, to reduce the structure loss in 2D lane representation, we directly extract top view lane information from front view images, which tremendously eases the confusion of distant lane features in previous methods. Furthermore, we propose a novel task-specific data augmentation method by synthesizing new training data for both segmentation and reconstruction tasks in our pipeline, to counter the imbalanced data distribution of camera pose and ground slope to improve generalization on unseen data. Our work marks the first attempt to employ the geometry prior information into DNN-based 3D lane detection and makes it achievable for detecting lanes in an extra-long distance, doubling the original detection range. The proposed method can be smoothly adopted by other frameworks without extra costs. Experimental results show that our work outperforms state-of-the-art approaches by 3.8% F-Score on Apollo 3D synthetic dataset at real-time speed of 82 FPS without introducing extra parameters.

1. Introduction

Lane detection is a fundamental and challenging task for autonomous vehicles. Recently, numerous researches [4, 5, 15, 26, 33] are conducted in this field for the extensive landing of high-level driving automation. Accurate lane detection will provide essential signals for the localization & mapping, decision making and path planning in autonomous driving systems (ADS), while most of the existing methods are built on the assumption of a flat ground plane [3, 17], which is not necessarily satisfied in practice. For instance, to restore the lane information, inverse perspective transformation (IPM) is often applied for projecting lanes from front view to top view [1, 19, 22, 27]. However, the homography matrix for IPM is usually calibrated by point pairs on the flat ground [24]. When encountering roads with non-zero slope like uphill or downhill situations, lane detection on 2D planar geometry would not accurately reflect the actual structure in real world, thus the ADS may malfunction due to the unexpected coordinate mapping. To solve this, it is critical to restore the height information by detecting lanes in 3D space. With accurate 3D lane estima-

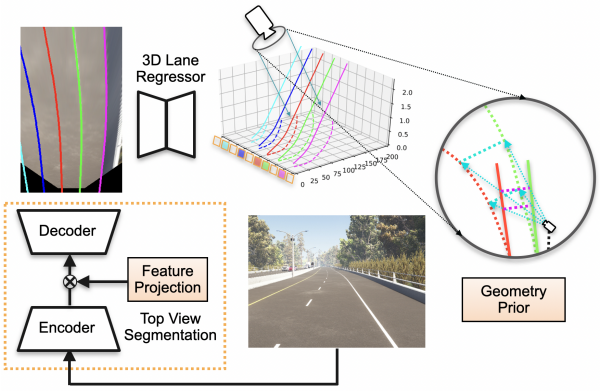

Figure 1. **Framework for 3D lane detection.** The framework is composed of a top view segmentation network and an anchor-based 3D lane regressor. The features extracted from a front view image are projected onto top view to generate the segmentation mask directly under top view supervision. The anchor representation of 3D lanes is then estimated from the top view lane mask. Moreover, we explicitly utilize the geometry prior from 3D to 2D projection for 3D lane reconstruction.

*Both authors contributed equally.
†This work was done during internship at SenseTime Research.
‡Guangliang Cheng is the corresponding author.

tion, most applications like lane keeping assistant (LKA) and lane departure warning (LDW) would be more reliable and robust, especially in non-flat ground scenarios.

In order to bridge the discrepancy between 2D and 3D lane detection, methods like [2, 6–8, 23, 32] were proposed. With the rapid development of the Convolutional Neural Network, deep learning based approaches [10, 12, 13, 18] have made tremendous progress in resolving this problem. Although the existing DNN-based methods have achieved promising results, they attempt to approach 3D lane detection as a point-wise regression task while the underlying geometry structure of 3D lanes seems to be ignored, which may result in unstable prediction with inaccurate structure.

In this paper, we consider 3D lane detection as a reconstruction problem from the 2D image to the 3D space. We propose that the geometry prior of 3D lanes should be explicitly imposed during the training process for fully utilizing the structural constraints of the inter-lane and intra-lane relationship, and the height information of 3D lanes can be extracted from the 2D lane representation. We first analyze the geometry relationship between 3D lane and its 2D representation, and propose an auxiliary loss function based on the geometry structure prior. We also demonstrate that the explicit geometry supervision would boost noise elimination, outlier rejection, and structure preservation for 3D lanes.

Second, in order to reduce the structural information loss in 2D plane, we redefine the pipeline by conducting lane segmentation with top view supervision instead of front view supervision, which addresses the issue of feature confusion due to the perspective distortion [25] on the far side. Lastly, we propose a novel task-specific data augmentation method on 3D lanes, which synthesizes new data by applying pitch, roll and yaw rotation on the original data. This augmentation could generate new data with various 3D ground plane steepness and road structure patterns, which eases the imbalanced distribution of ground plane slope and camera pose. Figure 1 shows the proposed framework, and a more detailed pipeline is shown in Figure 2.

In general, our contributions can be summarized as follows: 1) A novel loss function that enables explicit supervision based on the geometry structure prior of lanes in 3D space for stable reconstruction from local to global. 2) A 2D lane feature extraction module with direct supervision from the top view for maximum retention of lane structural information especially in the far end. 3) A task-specific data augmentation method for 3D lane detection aiming to counter the imbalanced data distribution of ground slope and camera pose to improve generalization on rare cases.

2. Related Work

There are several conventional methods [2, 6–8, 23, 32] that trickle the problem of lane detection in the 3D space. Dickmanns *et al.* [8] utilize a multi-sensor approach with temporal filtering to estimate the vertical curvature of 3D lane. Xiong *et al.* [32] adopt a B-spline curve to model lanes in 3D space, and resolve the depth ambiguity with known constant lane width prior. Coulombeau *et al.* [6] model 3D road with camera pose and road curvature to estimate 3D lane shape from road points with optimization algorithm. Nedevschi *et al.* [23] apply stereo-vision for estimating 3D road surface iteratively with tracking process. Bai *et al.* [2] utilize multi-view sensor approach by combining LiDAR and camera for detecting lanes in real world. Dementhon *et al.* [7] model roads as a Brooks ribbon in space with constant length assumption and iteratively construct the lane plane with horizontal lane width generator. However, most of these geometry based methods build up parametric models for lanes which are not robust enough for complex roads and would fall short on detection range.

Recently, DNN-based methods [10, 12, 13, 18] have shown impressive results on monocular 3D lane detection. Garnett *et al.* propose 3D-LaneNet [12], an end to end dual-pathway [14] network bridging IPM features and 3D lane regression. The image-view pathway extracts and preserves dense features from images to estimate the camera pose, and the top-view pathway then predicts the 3D lanes from the top view features projected from the image-view pathway. Furthermore, as an analogy to single-shot detection [21, 28], a column-based anchor representation on the ground plane is proposed to encode each 3D lane instance.

On the top of 3D-LaneNet [12], following recent anchor-free methods [9, 31], Efrat *et al.* [10] propose a more generalized semi-local tiled anchor representation to replace the previous column-based anchors, which divides continuous lane markings into short segments and encapsulates each of them into a non-overlapping cell. This novel anchor representation enables the generalization of complex topology like split, merge and horizontal lanes.

Gen-LaneNet [13] decouples 3D lane extraction from the image features by proposing a two-stage pipeline, which allows 3D lane estimation to be only dependent on 2D lane masks rather than the original images. This pipeline makes it possible to reduce the demand for expansive 3D lane labeling and utilize large 2D lane detection datasets to train a more robust lane feature extractor. Gen-LaneNet [13] also set up the anchor under virtual top view projection to refine the feature misalignment in 3D-LaneNet [12].

Besides, Jin *et al.* [18] utilize a dual attention module [11] with the framework of Gen-LaneNet [13] to capture lane-to-lane and pixel-wise attention. Also, a linear interpolation loss is proposed to sample lane points within each anchor more densely. This method improves the accuracy yet introduces extra parameters and computational cost.

The up-to-date DNN-based methods are able to show promising results in monocular 3D lane detection. How-

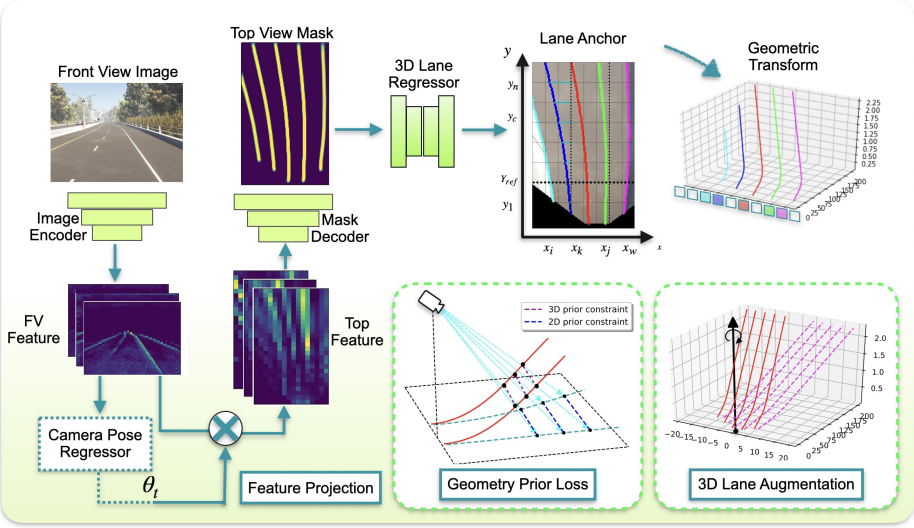

Figure 2. **Overall pipeline of the proposed method.**

ever, most of the existing methods seem to treat this problem only as a local lane regression task with 2D anchors, lacking explicit guidance on reconstructing the height of lanes from the 2D representation. 3D-LaneNet [12] tries to bridge image feature encoding directly with lane decoding with a heavy VGG [30] encoder, without deeper analysis of how the 3D lanes are detected from the image encoding. Gen-LaneNet [13] points out that the estimation of 3D lane height is equivalent to estimating the vector field moving nonparallel lane points on the virtual top view to corresponding 3D positions. This heuristic releases the demand of utilizing a heavy network to encode image features in previous work. However, no further quantitative analysis or explicit supervision is applied from this perspective. Similarly, 3D-LaneNet+ [10] proposes a novel semi-local cell-based anchor, but this only enriches the structure representation on 2D space, rather than targeting the recovery of 3D lane information directly.

Monocular 3D lane detection relies on the accurate reconstruction of the 3D lane structure, in which the height information can be decoded from the 2D features. Previous DNN-based methods [10, 12, 13, 18] choose to apply supervision on 2D anchors and a height dimension independently, however the underlying geometry structure information of lanes in the 3D space seems not to be fully utilized, which makes it challenging to accurately recover the 3D lanes. We propose that the structure of 3D lanes and the 2D-3D relationship following geometry prior should be jointly optimized in the whole process of 3D lane detection.

3. Method

The proposed method takes a single RGB image from the front-view camera, and outputs a group of lane instances in the 3D world space. Following the basic assumption in the previous literature [12, 13] and existing dataset [13], we assume that the camera is installed with zero roll or yaw respect to the world coordinate, and only has pitch variance due to vehicle fluctuation. We establish the world coordinate as the ego-vehicle coordinate with starting points as the perpendicular projection of camera center on the road. Figure 4 shows the world coordinate center at point O, and camera center at point C, with camera pitch θ.

3.1. Geometry in 3D Lane Detection

In most cases, lanes in 3D space are formed by a group of smooth parallel curves on the 3D road surface [7]. Instead of simply predicting discrete points of lane markings, an accurate reconstruction of 3D lanes should include the re-establishment of the reasonable geometry structure in 3D space. However, most of the existing DNN-based methods choose to achieve 3D lane detection in a data-driven manner with only point-wise supervision, which may not result in robust preservation of 3D lane geometry and would be vulnerable to outliers under extreme lane structures because of the absence of structural guidance. As a result, the geometry prior should be utilized to explicitly guide the learning of 3D lanes. We will first review the view projection systems proposed in previous literature and then analyze the structure prior under the existing projection system.

View projection. In 3D-LaneNet [12], a real top view projection is utilized for creating lane anchors on the flat ground. Real top view stands for the direct vertical projection from 3D space to the ground plane g. In this case, for a point $P_{3D}(x_{3D}, y_{3D}, z)$ in 3D space, the height dimension z is simply discarded, and the point is projected onto the ground plane position of $P_{2DR}(x_{3D}, y_{3D})$ under real top view projection with center at vehicle ego coordinate. However, as shown in the top row of Figure 3, such

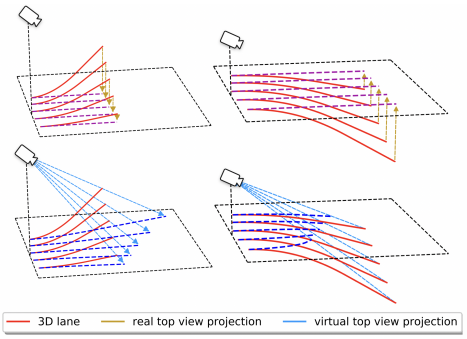

Figure 3. **Comparison of real and virtual top view.** Left: Uphill scenario, Right: Downhill scenario. The top row shows the 2D lane representation generated by real top view projection [13], while 2D lanes on the bottom row are generated by virtual top view projection [12]. The 2D lane representation generated by virtual top view projection [13] could reflect the changes of lane height in 3D space.

2D lane representation cannot properly reflect the change of lane height in 3D space, thus a heavy image feature encoder in [12] is necessary to estimate lane height from images.

Gen-LaneNet [13] then introduces the virtual top view projection. As shown in the bottom row of Figure 3, 3D lanes are projected onto the ground plane via virtual top view projection, which can be obtained from rays start from camera center C to the ground g. This is conceptually equivalent to 1) project the 3D lanes onto the image plane, and then 2) project image and lanes onto the ground plane by IPM. In this case, the 2D representation of lanes is no longer agnostic of height variance. Equation 1 shows the transformation of lane point coordinates from point P_{3D} in 3D to point P_{2DV} on the ground plane via virtual top view projection. With the increment of lane height z, the x and y of lane points on ground plane g would be projected away from the positions on real top view, causing the lane boundaries to be divergent in the uphill scenario.

$$P_{2DV} = \overrightarrow{CP_{3D}} \cap \Pi_g = \frac{h_{cam}}{h_{cam} - z} \cdot \begin{pmatrix} x_{3D} \\ y_{3D} \end{pmatrix} \quad (1)$$

Geometry prior. Parallel lane boundaries and constant lane width are basic assumptions for nearly all lane-based applications such as lane centering assistant (LCA) and lane keeping assistant (LKA). Instead of making a strong assumption that multiple lanes in one frame share a global lane width [17, 25], it is common to assume that a single lane would keep a relatively fixed width as it extends to infinity. For flat ground cases, the 2D lane representation projected via the virtual top view would be parallel and have constant lane width. In reality, not-flat ground cases are not rare, such as twisted lanes on a helicoidal surface. In this case, parallelism of the projected lane boundaries is not satisfied. Thus, the basic assumption of parallelism and constant width [7] can only be established in 3D space.

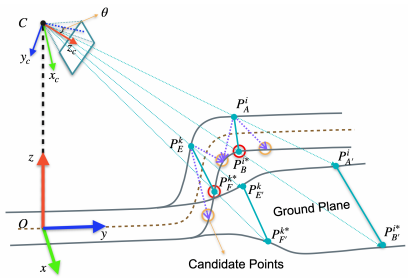

Figure 4. **Geometry prior of 3D lanes.**

Moreover, as shown in Equation 1, under virtual top view projection, lane boundaries in 3D space would be mapped onto the flat ground in different scales w.r.t. height information. Conceptually, the curvature of mapped lane boundaries is positively correlated to the lane slope in 3D space. That is to say, the width change in 2D projection could reflect the variance of lane height in 3D. This provides basic theoretical intuition of how the height information in 3D can be reconstructed from a monocular 2D image, and the geometry structure in the 2D lane mask can be the guidance for estimating the 3D lanes in the real world.

Geometry prior guided supervision. As illustrated above, the encoding of geometry structure information plays a vital role in accurately reconstructing lanes in 3D space. Specifically, we focus on the intra-lane and inter-lane properties between 3D lanes and 2D representation under the virtual top view projection. We propose the *geometry prior loss*, an auxiliary loss function to involve explicit supervision in the local-to-global preservation of 3D lane structure.

First, we build constraints upon the generalizable geometry prior of lane width in 3D space. The 3D lane width can be calculated as the 3D Euclidean distance D_{3D} between the closest point pair P_l^i on the left lane boundary and P_r^{i*} on the right as $D_{3D}(P_l^i, P_r^{i*}) = |\overrightarrow{P_l^i P_r^{i*}}|$, where i and $i*$ denote the corresponding anchor order in their y-anchor representation. Besides, we apply a greedy matching algorithm with sliding windows to find the corresponding lane width pairs (LWP) of points from the left lane boundary to the right. Formally: $LWP\{i, i*\} = \text{argmin}_j(D_{3D}(P_l^i, P_r^j)); j \in \mathbb{N} : \{i-1, i, i+1\}$

As shown in Figure 4, distance $D_{3D}(P_A^i, P_B^{i*})$ and $D_{3D}(P_E^k, P_F^{k*})$ should have the same magnitude. For a specific lane instance, we calculate D_{3D} for all point pairs on the lane boundaries with visibility $\hat{v}_l = 1$ as in Equation 2. The visibility flag $\hat{v}$ indicates whether a 3D lane point is visible when projected onto the Front view image.

$$D_{3D}(P_l^i, P_r^{i*}) = D_{3D}(P_l^{(i+1)}, P_r^{(i+1)*}); \quad i \in \mathbb{N} : \{0, 1, 2, ... \sum(\mathbb{1}\{\hat{v}_l = 1\}) - 1\} \quad (2)$$

Second, since projection method makes it achievable to bridge the 2D representation and 3D lanes in an explicit manner, we propose to make a joint optimization of offsets x, y, and height z by their relationship on the ground plane. As illustrated in Figure 4, point pairs $\{P'_A, P'_B\}$ and $\{P'_E, P'_F\}$ on ground plane are from the virtual top view projection of 3D point pairs $\{P_A, P_B\}$ and $\{P_E, P_F\}$. We define the 2D distance D_{2D} as the 2D Euclidean distance weighted by camera height h_{cam} and lane height. For the cases when z_l^i and z_r^{i*} are not the same, we use their mean value $\tilde{z}$ in approximation of the nonlinear function D_{2D}.

$$D_{2D}(P_l^i, P_r^{i*}) = \sqrt{(x_l^i - x_r^{i*})^2 + (y_l^i - y_r^{i*})^2} \cdot (h_{cam} - \tilde{z}); \quad \tilde{z} \sim (z_l^i + z_r^{i*})/2 \tag{3}$$

If and only if the point pairs $\{P'_A, P'_B\}$ and $\{P'_E, P'_F\}$ share the same height respectively, we have

$$D_{2D}(P_l^i, P_r^{i*}) = D_{3D}(P_l^i, P_r^{i*}) * h_{cam} = D_{2D}(P_l^{(i+1)}, P_r^{(i+1)*}) = D_{3D}(P_l^{(i+1)}, P_r^{(i+1)*}) * h_{cam} \tag{4}$$

in which the 2D distance equals to the product of 3D lane Euclidean distance and the camera height for $i \in \mathbb{N} : [0, \sum(\mathbb{1}\{\hat{v}_l = 1\}) - 1] \iff (z_l^i = z_r^{i*}) \cap (z_l^{(i+1)} = z_r^{(i+1)*})$

For twisted lanes, we relax the constraint in Equation 4. Under the heuristic that the curves of lane boundaries in 3D space have continuous derivatives in most cases, the height difference $\Delta(z_l - z_r)$ between point pairs is also locally smooth. As the D_{2D} is respect to the height difference of lane point pairs, specifically, we have Equation 5 for the point pair $i, i*$ and its neighbor pairs in the y-direction on the same lane.

$$D_{2D}(P_l^{(i-1)}, P_r^{(i-1)*}) \leq D_{2D}(P_l^i, P_r^{i*}) \leq D_{2D}(P_l^{(i+1)}, P_r^{(i+1)*}); i \in \mathbb{N} : [0, N-1] \Rightarrow \exists \frac{\partial D_{2D}}{\partial y} \tag{5}$$

Finally, we impose the dynamic constraint to the point pairs of each predicted lane with supervision from their neighboring pairs, as shown in Equation 6. For $d \in \{2D, 3D\}$, such supervision is applied on the predicted lane with probability p and visibility v equal to 1 in the corresponding ground truth (GT).

$$L_{geo} = \sum_d \sum_{i=1}^{N-1} p \cdot (||v^i \cdot ((D_d^{i-1} + D_d^{i+1}) - 2 * D_d^i)||_1) \tag{6}$$

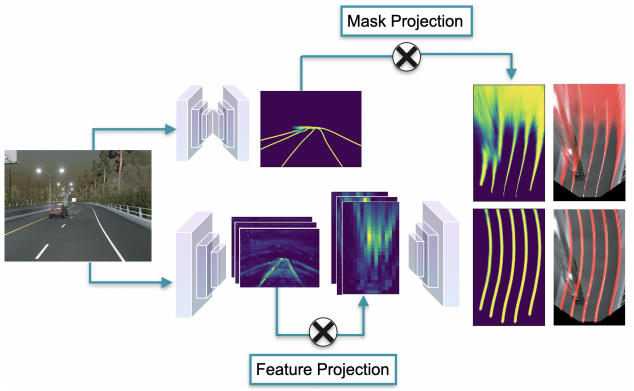

Figure 5. **Comparison between front view supervision with mask projection [13] and our top view supervision with feature projection.**

The definition of L_{geo} provides auxiliary supervision incorporating geometry prior into the training pipeline, which also leaves a margin for natural noises and gradual changes in lane width by weakening the geometry assumption of fixed lane width to adapt to both synthetic and real cases. The optimization of L_{geo} guides the model to maintain the lane structure from local to global.

3.2. Supervision from the Top View

As illustrated in section 3.1, the structure of 3D can be discerned from the geometry structure underlying the 2D lane mask. However, structure information like lane width variance is very subtle thus requiring accurate extraction of 2D lanes feature from the front view image.

In Gen-LaneNet [13], front view GT masks are used as supervision for the segmentation network. In the regression network, the mask is first projected onto the ground plane via planar IPM, and then fed into a lane regression network for estimating the 3D lanes. However, as shown in the top row of Figure 5, when the mask is projected onto the ground plane, the features on the far side are blurry, and different lane features are mixed together, which prevent the subsequent network from accurately decoding the structure information contained in the feature maps. We argue that the distant lane feature confusion is due to the misleading mask supervision conducted in the front view for the following reasons.

First, the IPM warps the front view features onto the ground plane via a non-linear transformation. And it is hard for the front view segmentation network to distinguish far lane instances in pixel level since the lanes are converging to the vanishing point and overlapped with each other. As shown on the top right of Figure 5, few far pixels on the front view will be mapped into a huge range on the top view, so the projected segmentation mask will be stretched drastically in the far end, resulting in feature confusion. Second, positive lane boundary pixels in the ground truth masks

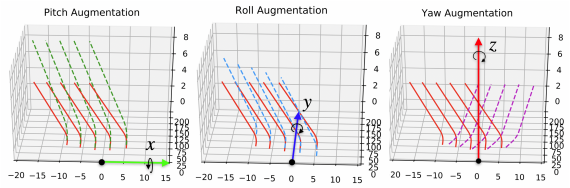

Figure 6. **Pitch, Roll and Yaw rotation augmentation on 3D lanes.**

of Gen-LaneNet [13] are generated on the front view with equal line thickness. However, due to the perspective effect, lane markings are always thicker in the close end and thinner in the far end. As a result, this kind of ground truth mask would not align properly with the shape of lane markings, which will lead to a fluctuation in extracting 2D lane structure and decoding 3D geometry information.

To address these issues, we propose a refined pipeline in which the ground truth mask is generated and used as supervision on the top view, as shown in the bottom row of Figure 5. We first project lane points from the front view to the virtual top view via the homography matrix H, then generate the GT mask of lane boundaries with equal thickness. A projective transformation network [12, 16] is used to project the output features of image encoder to the top view, and conduct segmentation supervision directly on the top view mask from decoder by optimizing the standard cross-entropy loss L_{seg}. Besides, we can also jointly estimate camera pitch θ and height h_{cam} from the front view images with $L_{cam} = ||\theta - \hat{\theta}||_1 + |h_{cam} - \hat{h}_{cam}||_1$. Such direct mask supervision in the top view will tremendously resolve the problem of feature confusion and preserve the global lane geometry structure information in the top view 2D mask.

3.3. Lane Augmentation in 3D Space

In both synthetic dataset and real world, the vast majority of ground planes are flat with almost zero slope, which severely dominates the distribution of training data. According to our statistics, in the Apollo synthetic dataset [13], 67.8% of lanes in training data have lane height within the range of [-0.1,0.1]m, while 44.4% of lanes fall into the range of [-0.01,0.01]m. Even though such a biased distribution is a close reflection of reality, it would dominate the direction of gradient optimization away from the long tail data, such as uphill and downhill cases.

Conventional augmentation methods, such as translation or center rotation, would violate the geometry prior that 3D lanes start from the center of ground plane. Thus, we propose a novel 3D lane rotation augmentation method to deal with the long tail distribution. Figure 6 shows the process of generating augmentation data by applying pitch, yaw and roll rotation on the 3D lanes. An example of applying pitch augmentation on 3D point P to get the target point P' under homogeneous transformation is shown in Equation 7. The rotated 3D points are projected onto front view to generate augmented input signals, which is equivalent to applying a perspective augmentation on the image and would be beneficial to the jointly understanding of semantic mask and camera pose.

$$P' = \mathbf{R}_x P; \mathbf{R}_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\theta & -sin\theta \\ 0 & sin\theta & cos\theta \end{bmatrix} \tag{7}$$

By applying pitch rotation, the ground plane would have greater fluctuation of slope while preserving the inner structure of lanes. By applying yaw rotation, the lane patterns are enriched without affecting the lane height. Lastly, by applying roll rotation, twisted 3D lanes are generated. Experimental results in Table 3 show the effectiveness of 3D lane augmentation over hard samples, while the results in section 4.3 shows its superiority on extra-long range data. Our proposed method can also be applied in both stages of mask prediction and anchor-based regression to improve robustness by building feature consistency under different perspective transformation [34, 35].

3.4. Training

Given an RGB image and the corresponding 3D lanes, a detailed pipeline for the framework is shown in Figure 2. For the segmentation head in feature extraction network, we optimize the cross-entropy loss L_{seg}. And for the optional camera pose regression head connected to the front-view image encoder, we apply the L_1-loss for supervision as in L_{cam}. The input of lane reconstruction network is the logits from the feature extraction network. Following the lane anchor settings in 3D-LaneNet [12] and Gen-LaneNet [13], lanes in 3D space are first projected onto the flat ground with virtual top view projection, then associated with closest anchor at Y_{ref}. For a predicted anchor X_A^i with probability p and the ground truth anchor $\hat{X_A^i} = \{(\hat{x_i}, \hat{z_i}, \hat{v_i}, \hat{p})\}$, each anchor point at a pre-defined y position y_i predicts the x offset x_i, the lane height z_i and the visibility v_i which means whether the 3D points are visible when projected onto front view. Both x-offset and y-position are defined on flat ground plane. We have the anchor loss written as

$$\begin{aligned} L_{anchor} = &- \sum_{i=1}^{N} (\hat{p}_i \log p_i + (1 - \hat{p}_i)(\log (1 - p_i)) \\ &+ \sum_{i=1}^{N} \hat{p}_i \cdot (||\hat{v}_i \cdot (x_i - \hat{x}_i)||_1 + ||\hat{v}_i \cdot (z_i - \hat{z}_i)||_1) \\ &+ \sum_{i=1}^{N} \hat{p}_i \cdot ||v_i - \hat{v}_i||_1 \end{aligned} \tag{8}$$

Table 1. Comprehensive evaluation on Apollo 3D synthetic dataset [13]. Parameters: Estimated network parameters to the best of our knowledge. GT mask: Use GT masks as the input of lane regression network to test the theoretical upper bound of corresponding method. The extra-long range refers to the data split of range 0-200m as described in section 4.3, while other splits refer to the default evaluation range of 0-100m.

		balanced scenes		rarely observed		visual variations		**extra-long range**	
Method	parameters	F-Score	AP	F-Score	AP	F-Score	AP	F-Score	AP
3D-LaneNet [12]	23.7M	86.4	89.3	72.0	74.6	72.5	74.9	60.1	63.2
Gen-LaneNet [13]	2.8M	88.1	90.1	78.0	79.0	85.3	87.2	68.5	69.2
Ours	2.8M	**91.9**	**93.8**	**83.7**	**85.2**	**89.9**	**92.1**	**83.6**	**85.3**
Gen-LaneNet (GT mask) [13]	-	91.8	93.8	84.7	86.6	90.2	92.3	80.7	82.5
Ours (GT mask)	-	**92.8**	**94.7**	**87.8**	**89.5**	**91.3**	**93.2**	**87.2**	**89.1**

Table 2. Results on offset metric and ablation study on balanced scenes in Apollo 3D synthetic dataset [13]. GS: Geometry supervision. TVS: Top view supervision. Aug: 3D lane data augmentation. GT mask: Use ground truth mask of 2D lanes instead of predicted mask for 3D lane regression. Joint: Evaluate on the intersection of matched samples across all methods in the table.

Method	F-Score	AP	x error near (m)	x error far (m)	**joint** x error far (m)	z error near (m)	z error far (m)	**joint** z error far (m)
3D-LaneNet [12]	86.4	89.3	0.068	0.477	0.448	0.015	0.202	0.183
Gen-LaneNet [13]**(baseline)**	88.1	90.1	0.061	0.496	0.470	0.012	0.214	0.192
Ours w/ GS	89.1	91.2	0.050	0.455	0.406	0.009	0.226	0.183
Ours w/ TVS	90.2	92.4	0.060	0.446	0.383	0.013	0.241	0.189
Ours w/ GS + TVS	91.2	93.2	0.065	0.415	0.365	0.009	0.220	0.179
Ours w/ GS + TVS + Aug	**91.9**	**93.8**	0.049	0.387	**0.332**	0.008	0.213	**0.171**
Gen-LaneNet (GT mask) [13]	91.8	93.8	0.054	0.412	0.353	0.011	0.226	0.177
Ours (GT mask)	**92.8**	**94.7**	0.044	0.360	**0.299**	0.007	0.219	**0.170**

Table 3. Comparison of augmentation on easy and hard samples

	all		easy		hard	
Method	F-Score	AP	F-Score	AP	F-Score	AP
Ours w/ GS	89.1	91.2	93.2	92.7	69.3	77.7
Ours w/ TVS	90.2	92.4	93.4	94.2	70.5	78.7
Ours (GS+TVS w/o Aug)	91.2	93.2	94.2	95.0	71.9	79.1
Ours (GS+TVS w/ Aug)	**91.9** (+0.7)	**93.8** (+0.6)	**94.6** (+0.4)	**95.2** (+0.2)	**75.4** (+3.5)	**83.7** (+4.6)

To sum up, the loss functions are composed of the loss of top view feature extraction network L_{fea} and the loss of 3D lane reconstruction network L_{rec}.

$$\begin{aligned} L_{fea} &= L_{seg} + \lambda_{cam} L_{cam} \\ L_{rec} &= L_{anchor} + \lambda_{geo} L_{geo} \end{aligned} \tag{9}$$

where the λ_{cam} and λ_{geo} are introduced to balance the relative importance between different loss items.

4. Experiment

4.1. Datasets and Implementation Details

To validate the performance of our proposed framework, we adopt the Apollo 3D synthetic dataset [13] for comparative experiments since other 3D lane datasets are not publicly available. This dataset contains 10500 discrete frames of monocular RGB images and the corresponding lane keypoint labels in 3D world coordinate. Follow the same setting in previous literature [13], the data is split into balanced, rarely observed and visual variation scenes. For evaluation metric, we adopt the minimum-cost flow algorithm [13] for matching between GT and predicted lanes, and report Average Precision (AP) and F-score for comprehensive evaluations. Besides, we also report horizontal offset on flat ground and the height distance in meters between matched pairs of GT and prediction lane on near (0-40m) and far (40-100m) range. We have λ_{cam}=1e^2 and λ_{geo}=1e^{-2} respectively.

For anchor representation, we adopt y-reference points in {5, 10, 15, 20, 30, 40, 50, 60, 80, 100}m, and set up Y_{ref} = 5 to associate each lane label with the closest anchor. For model training, we adopt the Adam [20] optimizer with initial learning rate of $5e^{-4}$, and no scheduler or weight decay is applied throughout the training. Our network can be trained in an end-to-end or two-stage way. We use batch size of 8 and train the entire network for 50 epochs on one GeForce GTX 1080Ti GPU. We adopt the same network architecture as ERFNet [29] and GeoNet in [13]. For lane augmentation, we randomly generate new lane data applying rotation for pitch within [-0.1,0.3] for p_{pitch} = 0.1, roll within [-3,3] for p_{roll} = 0.05, and yaw within [-3,3] for p_{yaw} = 0.2. We assume known camera pose in the training for fair comparison with previous work, even though the pose information can be obtained by network prediction as well.

4.2. Evaluation Result

The comprehensive experiment results are shown in Table 1, and the qualitative comparison is shown in Figure 7. Compared with the previous state-of-the-art works, we increase the F-score for more than 3.8% on all data splits without introducing extra network parameters, and our prediction results can closely achieve the theoretical results of Gen-LaneNet [13] with GT mask for the segmentation network. We also testify our methods on regression network with GT masks as input, and show that our framework can

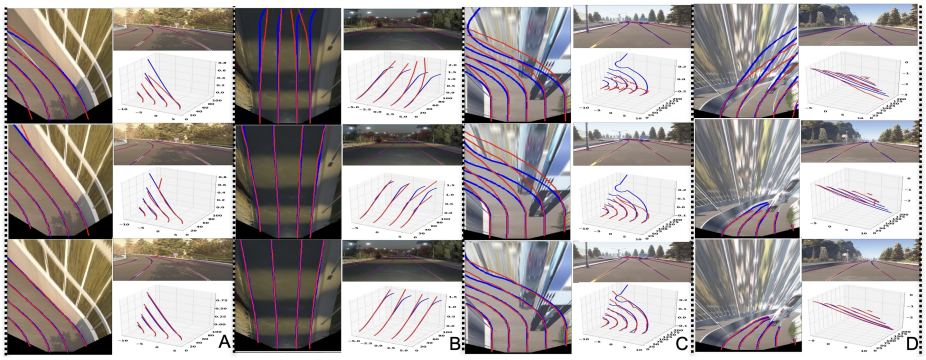

Figure 7. **Qualitative comparison results of proposed method.** First row: 3D-LaneNet [12]; Second row: Gen-LaneNet [13]; Third row: Our proposed method;different GTs are generated from different view projection; Column A & B: Results on default range (0-100m); Column C & D: Results on extra-long range (0-200m). GT and predicted lane boundaries are marked in blue and red respectively. Our proposed method could tremendously facilitate the reconstruction of 3D lanes, especially on the far side.

tremendously reduce the discrepancy between results under predicted 2D lane masks and GT masks with the same model architecture as [13], which proves the promising effect of our method in refining the top view features and preserving 3D lane structure.

4.3. Extra-long Range Comparison

The ability of detecting lanes at long distance is critical especially in the high speed scenario, thus we design an "extra-long range" data split to evaluate the ability of detecting 3D lanes within the range 0-200m, doubling the original evaluation range. We create this split by filtering lane data from the balanced scenes in [13] with y-position greater than 195m, and equally sample y-reference points from range [5, 200]m with a 5-meter interval. Experiment results are shown in Table 1. With the refinement of far-side features and the explicit geometry supervision, our proposed methods can increase both AP and F-score for roughly 15% over [13] and 25% over [12]. Specifically, our augmentation method makes an improvement on F-score from 79.7% to 83.6% and AP from 81.2% to 85.3%, which proves the superiority for generalizing into long-range detection.

4.4. Ablation Study

We conduct ablation study for evaluating the impact of each major part of our proposed method: Geometry supervision (GS), top view supervision (TVS) and 3D lane augmentation (Aug). We notice that in the evaluation of the offset metrics, the number of samples (i.e., positive matched lanes) evaluated in each method is not exactly the same, thus fairness cannot be assured since methods with higher F-score will consider more matched pairs of lanes which are often hard samples. To ensure fair comparison, we design a "joint" offset metric in which the offset errors are evaluated with the intersection of matched data samples from all methods in the table. As shown in Table 2, we use Gen-LaneNet [13] as the baseline. The GS and TVS can increase the performance of the baseline respectively. Compared with only applying GS on the baseline, the joint offset errors in the row 3-6 show that applying GS on the refined features with TVS can result in a much greater improvement in refining both "x error far" and "z error far", which reflects the joint benefit of GS and TVS on the preservation of geometry information in reconstructing lanes from 2D to 3D. Besides, to testify the ability to generalize on rare cases for our lane augmentation method, we split the balanced scenes of [13] into easy and hard cases. An image is labeled as hard case if the lane height of any lane within the image is greater than 1.78m, which is the value of the shared camera height in the dataset [13]. In general, 184 out of 1496 images are labeled as hard cases and the other 1312 images are labeled as easy. As shown in Table 3, the proposed augmentation method could tremendously improve the model performance on hard cases with 4.6% increment on AP and 3.5% on F-score, which proves the effectiveness on countering imbalanced data distribution.

5. Conclusion

In this paper[1], we propose a framework for monocular 3D lane detection with explicit geometry supervision, which outperforms previous methods without introducing extra parameters. We consider the task as a 2D to 3D lane reconstruction problem and prove that the structure prior could tremendously facilitate the reconstruction of 3D lanes by offering explicit guidance for lane geometry and refined lane features. On the Apollo 3D synthetic dataset, our proposed method can improve the performance of state-of-the-art methods by a large margin. Besides, our proposed approaches show the potential for extra-long distance 3D lane detection and achieve roughly 15% improvement in F-Score and AP over existing methods on the long-range data split.

[1]For the discussion of limitations and future work, please refer to the supplementary material.

References

[1] Mohamed Aly. Real time detection of lane markers in urban streets. In *2008 IEEE Intelligent Vehicles Symposium*, pages 7–12. IEEE, 2008. 1

[2] Min Bai, Gellert Mattyus, Namdar Homayounfar, Shenlong Wang, Shrinidhi Kowshika Lakshmikanth, and Raquel Urtasun. Deep multi-sensor lane detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3102–3109. IEEE, 2018. 2

[3] Massimo Bertozzi and Alberto Broggi. Gold: A parallel real-time stereo vision system for generic obstacle and lane detection. *IEEE transactions on image processing*, 7(1):62–81, 1998. 1

[4] Vadim A Butakov and Petros Ioannou. Personalized driver/vehicle lane change models for adas. *IEEE Transactions on Vehicular Technology*, 64(10):4422–4431, 2014. 1

[5] Zhilu Chen and Xinming Huang. End-to-end learning for lane keeping of self-driving cars. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1856–1860. IEEE, 2017. 1

[6] Pierre Coulombeau and Claude Laurgeau. Vehicle yaw, pitch, roll and 3d lane shape recovery by vision. In *Intelligent Vehicle Symposium, 2002. IEEE*, volume 2, pages 619–625. IEEE, 2002. 2

[7] Daniel DeMenthon. A zero-bank algorithm for inverse perspective of a road from a single image. In *Proceedings. 1987 IEEE International Conference on Robotics and Automation*, volume 4, pages 1444–1449. IEEE, 1987. 2, 3, 4

[8] Ernst D. Dickmanns and Birger D. Mysliwetz. Recursive 3-d road and relative ego-state recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 14(02):199–213, 1992. 2

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019. 2

[10] Netalee Efrat, Max Bluvstein, Shaul Oron, Dan Levi, Noa Garnett, and Bat El Shlomo. 3d-lanenet+: Anchor free lane detection using a semi-local representation. *arXiv preprint arXiv:2011.01535*, 2020. 2, 3

[11] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 2

[12] Noa Garnett, Rafi Cohen, Tomer Pe'er, Roee Lahav, and Dan Levi. 3d-lanenet: end-to-end 3d multiple lane detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2921–2930, 2019. 2, 3, 4, 6, 7, 8

[13] Yuliang Guo, Guang Chen, Peitao Zhao, Weide Zhang, Jinghao Miao, Jingao Wang, and Tae Eun Choe. Gen-lanenet: A generalized and scalable approach for 3d lane detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 666–681. Springer, 2020. 2, 3, 4, 5, 6, 7, 8

[14] Bei He, Rui Ai, Yang Yan, and Xianpeng Lang. Accurate and robust lane detection based on dual-view convolutional neutral network. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 1041–1046. IEEE, 2016. 2

[15] Namdar Homayounfar, Wei-Chiu Ma, Justin Liang, Xinyu Wu, Jack Fan, and Raquel Urtasun. Dagmapper: Learning to map by discovering lane topology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2920, 2019. 1

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015. 6

[17] Yan Jiang, Feng Gao, and Guoyan Xu. Computer vision-based multiple-lane detection on straight road and in a curve. In *2010 International Conference on Image Analysis and Signal Processing*, pages 114–117. IEEE, 2010. 1, 4

[18] Yujie Jin, Xiangxuan Ren, Fengxiang Chen, and Weidong Zhang. Robust monocular 3d lane detection with dual attention. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3348–3352. IEEE, 2021. 2, 3

[19] ZuWhan Kim. Robust lane detection and tracking in challenging scenarios. *IEEE Transactions on intelligent transportation systems*, 9(1):16–26, 2008. 1

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[22] Hanspeter A Mallot, Heinrich H Bülthoff, JJ Little, and Stefan Bohrer. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological cybernetics*, 64(3):177–185, 1991. 1

[23] Sergiu Nedevschi, Rolf Schmidt, Thorsten Graf, Radu Danescu, Dan Frentiu, Tiberiu Marita, Florin Oniga, and Ciprian Pocol. 3d lane detection system based on stereovision. In *Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No. 04TH8749)*, pages 161–166. IEEE, 2004. 2

[24] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 286–291. IEEE, 2018. 1

[25] Marcos Nieto, Luis Salgado, Fernando Jaureguizar, and Jon Arróspide. Robust multiple lane road modeling based on perspective analysis. In *2008 15th IEEE International Conference on Image Processing*, pages 2396–2399. IEEE, 2008. 2, 4

[26] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1

[27] Dean Pomerleau. Ralph: Rapidly adapting lateral position handler. In *Proceedings of the Intelligent Vehicles' 95. Symposium*, pages 506–511. IEEE, 1995. 1

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object de-

tection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[29] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2017. 7

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[31] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2

[32] Lu Xiong, Zhenwen Deng, Peizhi Zhang, and Zhiqiang Fu. A 3d estimation of structural road surface based on lane-line information. *IFAC-PapersOnLine*, 51(31):778–783, 2018. 2

[33] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[34] Yunhan Zhao, Shu Kong, and Charless Fowlkes. Camera pose matters: Improving depth prediction by mitigating pose distribution bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15759–15768, 2021. 6

[35] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7556–7566, 2021. 6