

## 3D-LaneNet: End-to-End 3D Multiple Lane Detection

Noa Garnett Rafi Cohen Tomer Pe'er Roei Lahav Dan Levi

General Motors Israel  
 HaMada St. 7, Herzlya, Israel

{noa.garnett, rafi.cohen, tomer.peer, roee.lahav, dan.levi}@gm.com

### Abstract

We introduce a network that directly predicts the 3D layout of lanes in a road scene from a single image. This work marks a first attempt to address this task with on-board sensing without assuming a known constant lane width or relying on pre-mapped environments. Our network architecture, 3D-LaneNet, applies two new concepts: intra-network inverse-perspective mapping (IPM) and anchor-based lane representation. The intra-network IPM projection facilitates a dual-representation information flow in both regular image-view and top-view. An anchor-per-column output representation enables our end-to-end approach which replaces common heuristics such as clustering and outlier rejection, casting lane estimation as an object detection problem. In addition, our approach explicitly handles complex situations such as lane merges and splits. Results are shown on two new 3D lane datasets, a synthetic and a real one. For comparison with existing methods, we test our approach on the image-only tuSimple lane detection benchmark, achieving performance competitive with state-of-the-art.

### 1. Introduction

3D lane detection, comprising of an accurate estimation of the 3D position of the drivable lanes relative to the host vehicle, is a crucial enabler for autonomous driving. Two complementary technological solutions exist: loading pre-mapped lanes generated off-line [33] and perception-based real-time lane detection [4]. The off-line solution is geometrically accurate given precise host localization (in map coordinates) but complex to deploy and maintain. The most common perception-based solution uses a monocular camera as the primary sensor for solving the task. Existing camera-based methods detect lanes in the image domain and then project them to the 3D world by assuming a flat ground [4], leading to inaccuracy not only in the elevation but also in the lane curvature when the assumption is violated. Inspired by recent success of convolutional neural networks (CNNs) in monocular depth estimation [20], we

propose instead to directly detect lanes in 3D. More formally, given a single image taken from a front-facing camera, the task is to output a set of 3D curves in camera coordinates, each describing either a lane delimiter or a lane centerline.

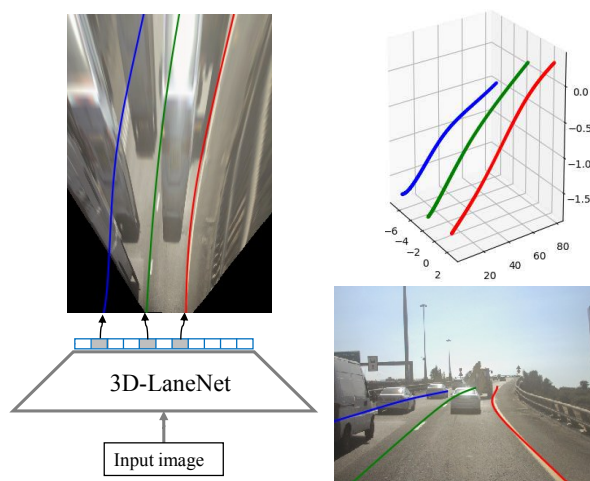


Figure 1. **End-to-end approach illustrated.** **Left:** Output represented in top view. **Top-right:** result visualized in 3D. **Bottom-right:** result projected to original input image

3D-LaneNet, our proposed solution, is a deep CNN that performs 3D lane detection. The network, trained end-to-end, outputs in each longitudinal road slice, the confidence that a lane passes through the slice and its 3D curve in camera coordinates. Our approach is schematically illustrated in Figure 1. Our direct, single-shot approach avoids post-processing used in existing methods such as clustering and outlier rejection. The network's backbone is based on a novel dual pathway architecture that uses several intra-network projections of feature maps to virtual bird-eye-view. This dual representation gives the network an enhanced ability to infer 3D in a road scene, and may possibly be used for other tasks requiring this ability (e.g. 3D car detection). The output is represented by a new column-

based anchor encoding which makes the network horizontally invariant and enables the end-to-end approach. Each output is associated to an anchor in analogy to single-shot, anchor-based object detection methods such as SSD [21] and YOLO [29]. Effectively, our approach casts the problem as an object detection one, in which each lane entity is an object, and its 3D curve model is estimated just like bounding box for an object.

We validate our approach on three different datasets. The primary dataset used to develop the approach is a new computer-graphics dataset, *synthetic-3D-lanes*<sup>1</sup> providing full access to the exact 3D position of each lane element. While several driving simulators exist [8, 30], they are not focused on the 3D lane detection task, and are limited in the variability of the relevant scene properties (e.g. lane curvature). Our main accomplishment in this domain, is the ability to randomly generate road segments with highly variable 3D shapes and lane topology. We therefore used it as the primary dataset for evaluation and ablation study. To validate our approach on real-world images we collected an additional dataset, *3D-lanes*, from a vehicle-mounted front camera. 3D Lane annotation was accomplished using a Lidar scanner in a semi-manual annotation scheme. Finally, to compare with state-of-the-art lane detection methods, which only operate in the image domain, we adapted our method to this task, and demonstrated end-to-end image-only lane detection. This image-only version is tested on the tuSimple dataset [1], reaching results competitive to state-of-the-art without the common post-processing techniques. To summarize, our main contributions are:

- Introduction of a novel problem: single-frame 3D lane detection without geometric assumptions, along with new evaluation metrics
- A novel dual-pathway architecture deploying intra-network feature map IPM projections
- An new anchor-based output representation of lanes enabling a direct, end-to-end trained network, for both 3D and image-based lane detection.
- A methodology for generating randomly synthetic examples with variation in lane topology (i.e. number of lanes, merges, splits) and 3D shape.

## 2. Related Work

Traditional lane detection systems (e.g. [10]) combine low-level operations such as directional filters with high-level heuristics such as the Hough transform to detect continuous lanes in the image. A common pipeline includes 4 stages: local lane-feature extraction (1), lane model fitting (2), image-to-world correspondence (3) and temporal aggregation (4). Bar-Hillel *et al.* [4] provide this modular

decomposition alongside a detailed overview of traditional systems. In recent years, the *local feature extraction stage* is performed by applying one or more CNNs to the image, but the overall pipeline remains very similar and the latter post processing stages remain.

Initially, CNNs were used to improve feature extraction by either enhancing the edge map (Kim and Lee [15]) or classifying candidate patches (He *et al.* [12]). Huval *et al.* [13] detects local lane line segments with an object detection CNN. VPGNet (Lee *et al.* [18]), follows a similar concept and additionally detects other road markings and the vanishing point to improve lane detection. Kim and Park [16] re-formulate the local-feature extraction stage as a semantic-segmentation problem, with two classes corresponding to the left and right lane delimiters, extending the reach of the network to perform clustering. However, a world-coordinate lane model must still be fitted to each cluster, and multiple lanes are not handled. Neven *et al.* [25] make an attempt towards end-to-end multi-lane detection, by training a CNN not only to create a binary lane pixel mask but also a feature embedding used for clustering lane points. Ghafoorian *et al.* [9] propose applying a generative adversarial network to make the semantic segmentation network output more realistic in the context of lane detection. Several works (e.g. Meyer *et al.* [23], Oliveira *et al.* [26]) are built on a similar approach in which the host and possibly adjacent lanes are the semantic classes (lane interior rather than the lane delimiters).

As opposed to all presented methods, 3D-LaneNet unifies the first three stages of the common pipeline by providing a full multi-lane representation in 3D world coordinates directly from the image in a single feed-forward pass. In addition, previous methods use the flat ground assumption for the image-to-world correspondence while our method fully estimates the parametrized 3D curves defining the lanes. Only a few methods directly address 3D lane estimation: [24], using stereo, and [34, 6] which follow a multi-view geometry approach and assume a known constant road / lane width to solve the depth ambiguity. Instead, we use a data driven approach and make no geometric assumptions.

Inverse perspective mapping (IPM) generates a virtual top-view (sometimes called bird-eye-view) of the scene from a camera-view as in the example in Figure 1. It was introduced in the context of obstacle detection by Malot *et al.* [22] and first used for lane detection by Pomerleau [28]. IPM has since been extensively used for lane detection (e.g. [5, 3]) since lanes are ordinarily parallel in this view and their curvature can be accurately fitted with low-order polynomials. In addition, removing the perspective effects causes lane markings to look similar (except for blurring effects) regardless of their distance from the camera. Most recently He *et al.* [12] introduced a “Dual-view CNN” which is composed of two separate sub-networks,

<sup>1</sup><https://sites.google.com/view/danlevi/3dlanes>

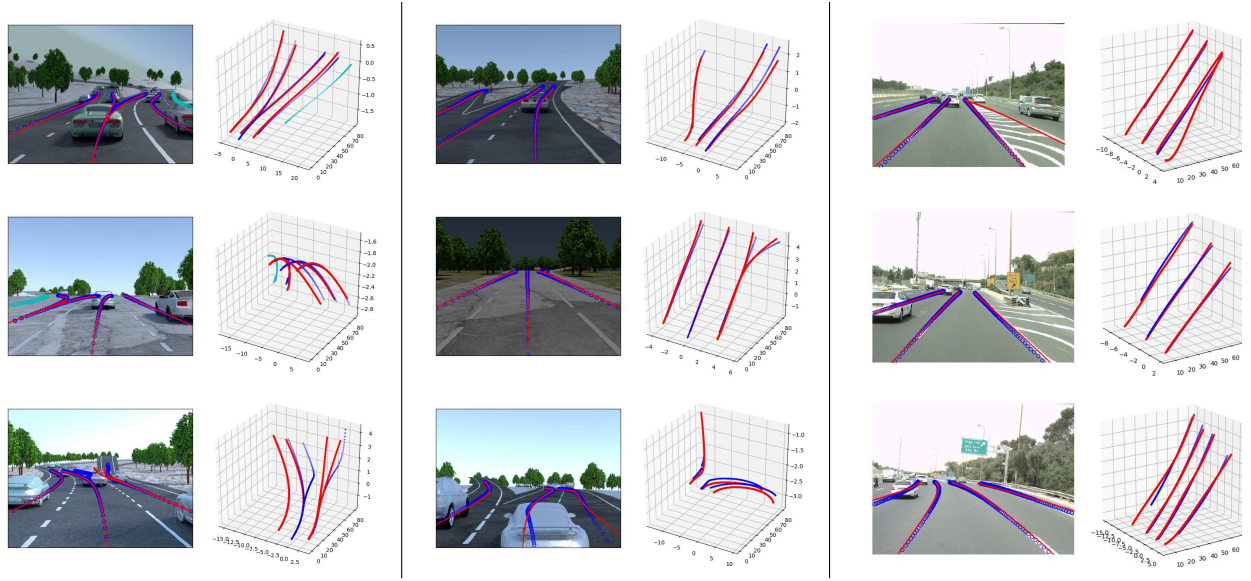


Figure 2. **Result visualization on test images. Centerline detection on *synthetic-3D-lanes* examples (Left and middle columns) and Delimiter detection on *3D-lanes* real image examples (Right column).** Detections with confidence  $> 0.5$  are shown. Ground truth (blue) and method result (red) shown in each image alongside a 3D visualization. Note that 3D axes are scene adaptive. Ignored lanes are marked in cyan. The leftmost bottom example shows a failure in correctly assigning a lane split, probably caused by occlusion.

each producing a descriptor (one per view) which are then concatenated and applied to candidate image locations. Li *et al.* [19] use a CNN to detect lane markings along with geometrical attributes, such as local position and orientation, directly on a top-view image which preserves invariance to these properties. In addition they deploy a second, recurrent network, that traverses the image to detect consistent lanes. Neven *et al.* [25] use the horizon, predicted in each image by a sub-network (“H-net”), to project the lanes to top-view for improved curve fitting. In contrast to previous work, we exploit both views in a synergistic single network approach.

More generally, we propose the first method that uses an end-to-end trained CNN to directly detect multiple lanes and estimate the 3D curvature for each such lane. We also show that our method is applicable both to centerlines and delimiters with an ability to handle splits and merges as well, without any further post-processing.

### 3. Method

Our method gets as input a single image taken from a front facing camera mounted on a vehicle as illustrated in Figure 3. We assume known intrinsic camera parameters  $\kappa$  (e.g. focal length, center of projection). We also assume that the camera is installed at zero degrees roll *relative to the local ground plane*. We do not assume a known camera height and pitch since these may change due to vehicle dynamics. The lanes in a road scene can be described both by the set of centerlines  $\{C_i\}_{i=1}^{N_C}$  of each lane and by the

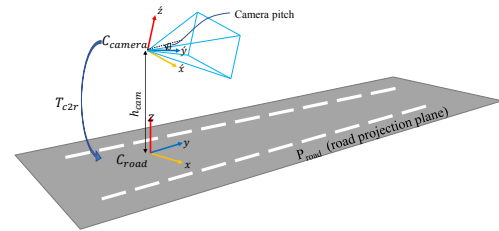


Figure 3. **Camera position and road projection plane**

set of lane delimiters  $\{D_i\}_{i=1}^{N_D}$  as illustrated in Fig. 5. Each such lane entity (centerline or delimiter) is a curve in 3D expressed in camera coordinates ( $C_{camera}$ ). The task is to detect either the set of lane centerlines and/or lane delimiters given the image.

#### 3.1. Top-view projection

We briefly review Inverse Perspective Mapping (IPM). In short, IPM is a homography that warps a front view image to a virtual top view image as depicted in the top-left image of Figure 1. It is equivalent to applying a camera rotation homography (view is rotated downwards) followed by an anisotropic scaling [11]. In our implementation we want to ensure that each pixel in the top view image corresponds to a predefined position on the road, *independent* of the camera intrinsics and its pose relative to the road.

See Figure 3 for an illustration of the following definitions. The camera coordinates  $C_{camera} = (\hat{x}, \hat{y}, \hat{z})$  are set

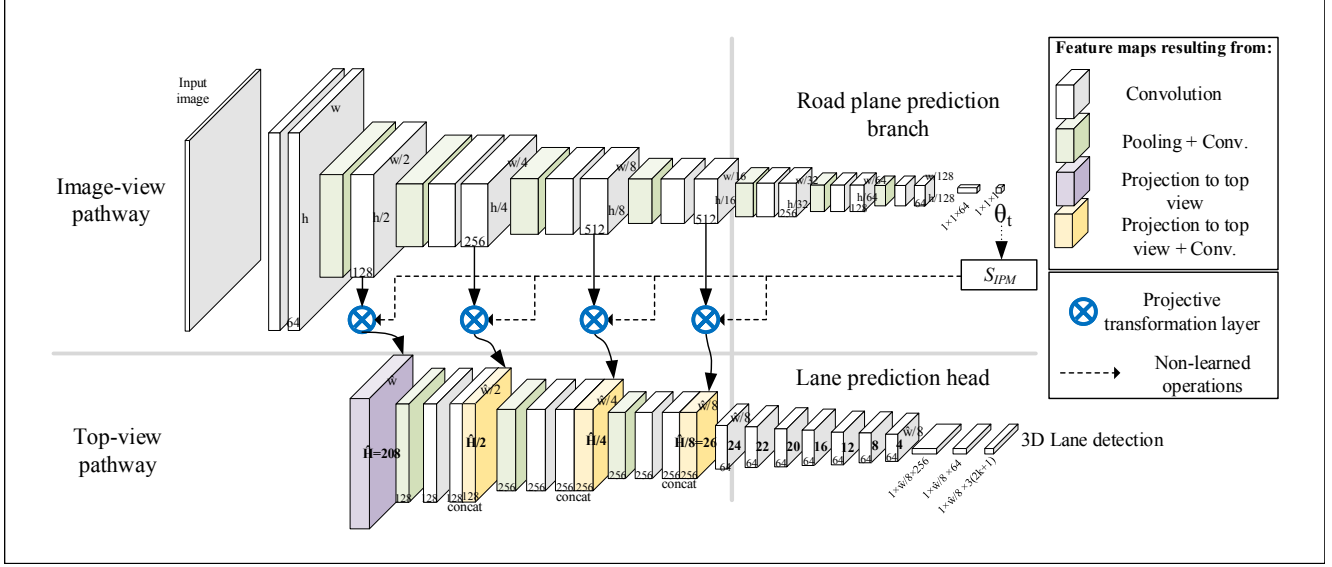


Figure 4. 3D-LaneNet network architecture.

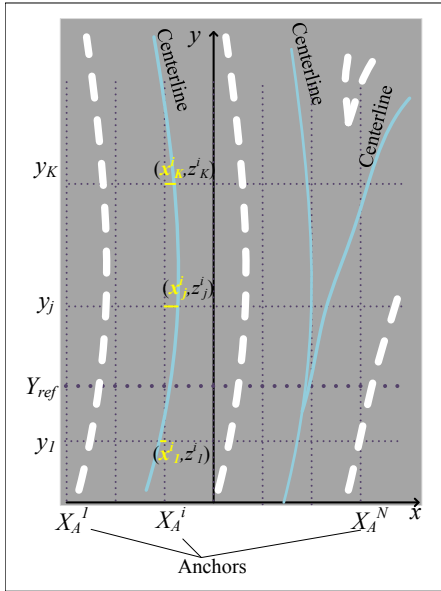


Figure 5. **Output representation.** Note that the number of anchors ( $N$ ) equals the output layer width (marked by  $\hat{w}/8$  in Fig. 4). The geometry representation for one lane *centerline* is shown. Lane *delimiters* (white dashed curves) are similarly represented.

such that  $\hat{y}$  is the camera viewing direction. Let  $P_{road}$  be the plane *tangent* to the local road surface. We define the road coordinates  $C_{road} = (x, y, z)$  as follows:  $z$  direction is the normal to  $P_{road}$ ,  $y$  is the projection of  $\hat{y}$  onto  $P_{road}$  and the origin is the projection of the camera center onto  $P_{road}$ . Let  $T_{c2r}$  be the 6-D.O.F. transformation between  $C_{camera}$  and  $C_{road}$  (3D translation and 3D rotation). Since we assume

zero camera roll,  $T_{c2r}$  is uniquely defined by the camera pitch angle  $\theta$  and its height above the ground  $h_{cam}$ . The homography  $H_{r2i} : \mathbb{P}^2 \mapsto \mathbb{P}^2$ , mapping each point on  $P_{road}$  to image plane coordinates, is determined by  $T_{c2r}$  and  $\kappa$  (See [11], Section 8.1.1). Finally, the IPM is obtained from  $H_{r2i}$  using a fixed set of parameters  $IPM_{Params}$  defining the top view region boundaries and an anisotropic scaling from meters to pixels. The top view image is generated using bilinear interpolation defined by a sampling grid  $S_{IPM}$ .

### 3.2. Network structure

An overview of the 3D-LaneNet is illustrated in Figure 4. Information is processed in two parallel streams or pathways: the image-view pathway and the top-view pathway. We call this the *dual-pathway backbone*. The image-view pathway processes and preserves information from the image while the top-view pathway provides the features with translation invariance and is used to predict the 3D lane detection output. The architecture for the image-view pathway is based on VGG16 [31] while the top-view pathway is similarly structured. Information flows to the top-view pathway through four *projective transformation* layers as follows.

### 3.3. The projective transformation layer

A main building block in our architecture is the *projective transformation layer* marked in blue in Fig. 4. This layer is a specific realization, with slight variations, of the *spatial transformer module* [14]. It performs a differentiable sampling of input feature maps corresponding spatially to the *image plane*, to output feature maps corresponding spatially to a *virtual top view* of the scene while preserving the # of channels. The differential sampling is achieved

through a grid generated as described in Sec. 3.1, using a IPM predicted by the *Road projection prediction branch* as described in the next section. The resulting projected feature maps, except for the first set, are concatenated to downstream feature maps from the top-view pathway. A subsequent neuron, operating on the concatenated feature maps combines the following two desirable properties for lane detection. First, translational invariance in the top-view plane. This is valuable since in the top view lanes have similar appearance and geometry across the space. Second, preservation of a dual information context - in both image and top view. The additional image-view context encodes information which is not present in the top view such as fences, skyline and trees which are crucial for deciphering the 3D structure of the scene. Particularly, in the far range, the image-view context is much richer in visual information, and represents a much larger actual area compared to the top view.

### 3.3.1 Road projection prediction branch

The first intermediate output of the image-view pathway network is an estimation of the “road projection plane”  $P_{road}$ . Essentially, this branch predicts  $T_{c2r}$ , the camera ( $\mathcal{C}_{camera}$ ) to road ( $\mathcal{C}_{road}$ ) transformation. It is trained in a supervised manner.  $T_{c2r}$  determines the top-view homography  $H_{r2i}$  and sampling grid  $S_{IPM}$  as explained in Section 3.1, and is therefore needed for the feed-forward step of the top-view pathway. At inference time, it is used also to translate the network output which is expressed in  $\mathcal{C}_{road}$ , back to  $\mathcal{C}_{camera}$ . As described in Section 3.1,  $T_{c2r}$  is defined in our case by the camera height  $h_{cam}$  and pitch  $\theta$ , and therefore these are the two outputs of this branch.

### 3.3.2 Lane prediction head

At the heart of our end-to-end approach lies the **anchor-based lane representation**. Inspired by object detection, we use anchors to define lane candidates and a refined geometric representation to describe the precise 3D lane shape for each anchor. The output coordinate system is the estimation of  $\mathcal{C}_{road}$  determined by  $h_{cam}$ ,  $\theta$ . Our anchors correspond to longitudinal lines in this coordinate system and the refined lane geometry to 3D points relative to the respective anchor. As illustrated in Figure 5, we define the anchors by equally spaced vertical (longitudinal) lines in  $x$ -positions  $\{X_A^i\}_{i=1}^N$ . Per anchor  $X_A^i$ , a 3D lane is represented by  $2 \cdot K$  output neurons activation  $(\mathbf{x}^i, \mathbf{z}^i) = \{(x_j^i, z_j^i)\}_{j=1}^K$ , which together with a fixed vector of  $K$  predefined  $y$  positions ( $\mathbf{y} = \{y_j\}_{j=1}^K$ ) define a set of 3D lane points. The values  $x_j^i$  are horizontal offsets *relative to the anchor position*  $X_A^i$ . Meaning, the output  $(x_j^i, z_j^i)$  represents the point  $(x_j^i + X_A^i, y_j, z_j^i) \in \mathbb{R}^3$ , in  $\mathcal{C}_{road}$  coordinates. In addition,

for each anchor  $i$ , we output the confidence  $p^i$  that there is a lane associated with the anchor. We use a predefined longitudinal coordinate  $Y_{ref}$  for the association. The anchor  $X_A^i$  associated to a lane is the one closest to the  $x$ -coordinate of the lane at  $y = Y_{ref}$ .

Per anchor, the network outputs up to *three* types ( $t$ ) of lane descriptors (confidence and geometry), the first two ( $c_1, c_2$ ) represent lane centerlines and the third type ( $d$ ) a lane delimiter. Assigning two possible centerlines per anchor yields the network support for merges and splits which may often result in having the centerlines of two lanes coincide at  $Y_{ref}$  and separating at different road positions as in the rightmost example in Figure 5. The topology of lane *delimiters* is generally more complicated compared to centerlines and our representation cannot capture all situations (for example the lane delimiters not crossing  $y = Y_{ref}$  in Fig. 5). The prediction head of the 3D-LaneNet is designed to produce the described output. Through a series of convolutions with no padding in the  $y$  dimension, the feature maps are reduced, and finally the prediction layer size is  $3 \cdot (2 \cdot K + 1) \times 1 \times N$  s.t. each column  $i \in \{1 \dots N\}$  corresponds to a single anchor  $X_A^i$ . Per anchor,  $X_A^i$  and type  $t \in \{c_1, c_2, d\}$  the network output is denoted by  $(\mathbf{x}_t^i, \mathbf{z}_t^i, p_t^i)$ . The final prediction performs a 1D non-maximal suppression as common in object detection: only lanes which are locally maximal in confidence (compared to the left and right neighbor anchors) are kept. Each remaining lane, represented by a small number ( $K$ ) of 3D points, is translated to a smooth curve using spline interpolation.

### 3.4. Training and ground truth association

Given an image example and its corresponding 3D lane curves,  $\{C_i\}_{i=1}^{N_C}$  (centerlines) and  $\{D_i\}_{i=1}^{N_D}$  (delimiters), the training proceeds as follows. First, the ground truth (GT) coordinate system  $\mathcal{C}_{road}$  is defined for the local road tangent plane as described in Sec. 3.1 using the known pitch ( $\hat{\theta}$ ) and camera height ( $\hat{h}_{cam}$ ). Next, each lane curve, projected to the  $x - y$  plane of  $\mathcal{C}_{road}$ , is associated with the closest anchor at  $Y_{ref}$ . The leftmost lane delimiter and leftmost centerline associated with an anchor are assigned to the  $c_1$  and  $d$  output types for that anchor. If an additional centerline is associated to the same anchor it is assigned to output type  $c_2$ . This assignment defines the GT per example in the same format as the output: per anchor  $X_A^i$  and type  $t$  the associated GT is denoted by  $(\hat{\mathbf{x}}_t^i, \hat{\mathbf{z}}_t^i, \hat{p}_t^i)$ , where  $\hat{p}_t^i$  is an anchor/type assignment indicator, and the coords in  $\mathcal{C}_{road}$ .

Both in training time and in the evaluation, entire lanes are ignored if they do not cross  $Y_{ref}$  inside valid top-view image boundaries, and lane points are ignored if occluded by the terrain (i.e. beyond a hill top). The overall loss function of the network is given in Eq. 1. It combines three equally weighed loss terms: lane detection (Cross-entropy-loss), lane geometry and road plane estimation ( $L_1$ -loss).



$$\begin{aligned}
\mathcal{L} = & - \sum_{t \in \{c_1, c_2, d\}} \sum_{i=1}^N (\hat{p}_t^i \log p_t^i + (1 - \hat{p}_t^i) \log (1 - p_t^i)) \\
& + \sum_{t \in \{c_1, c_2, d\}} \sum_{i=1}^N \hat{p}_t^i \cdot (\|\mathbf{x}_t^i - \hat{\mathbf{x}}_t^i\|_1 + \|\mathbf{z}_t^i - \hat{\mathbf{z}}_t^i\|_1) \\
& + |\theta - \hat{\theta}| + |h_{cam} - \hat{h}_{cam}|
\end{aligned} \tag{1}$$

## 4. Experiments

Our experimental work is presented as follows. We first present the methodology used for generating a new synthetic dataset *synthetic-3D-lanes*, which is used to derive most of this study conclusions. Next, we introduce the *3D-lanes* dataset generated for validation on real-world imagery. Using a newly proposed evaluation method for 3D lane detection, we then present results on both datasets, including an ablation study carefully examining the contribution of each concept in our overall approach. Finally, we compare an image-only version of 3D-LaneNet to existing state-of-the-art methods on the tuSimple benchmark [1].

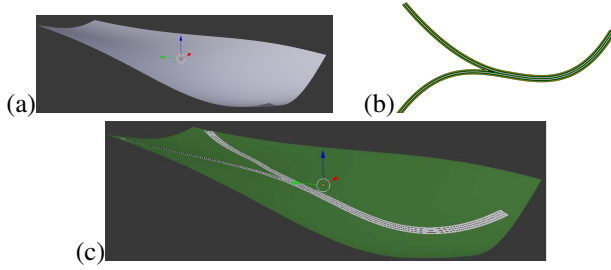


Figure 6. **Synthetic scene generation example.** (a) Surface (b) Road topology and curvature (c) Road on surface

### 4.1. Synthetic 3D lane dataset

We generated the *synthetic-3D-lanes* dataset using the open source graphics engine *blender* [2]. Our programmatic approach allows us to randomize each of the modeled elements, from the 3D geometry of the scene to object types as illustrated in Figure 6. The process of generating *each scene* is composed of the following steps:

**Terrain 3D.** The terrain is modeled by a Mixture of Gaussians distribution with the number of Gaussians and their parameters randomized. Figure 6(a) shows an example of such a terrain.

**Lane topology.** Number of lanes on the main road is selected. Then we choose if there is a secondary road and the number of lanes in it. Depending on the later direction of the camera in the scene the junction of the secondary road is viewed as a merge or a split.

**Lane top view geometry.** The geometry of the main road in top view is modeled by a 4<sup>th</sup> degree polynomial producing mild to extreme curvatures. The junction point for merges / splits as well as the lane width are chosen. This results in a top view lane-level map as shown in Fig. 6(b).

**Lane 3D.** The top-view lane map is placed on the terrain, and secondary roads are lifted to simulate common road topography. Fig. 6(c) shows the result of this stage.

**Terrain and road appearance.** The texture of the road and the terrain are chosen from a set of textures. The type and color of the lane markings is also randomized.

**Objects.** Cars and trees selected from a set of models are placed in the scene, on and off the road respectively.

**Scene rendering.** The host vehicle camera is positioned on the main road by choosing its lane and a lateral offset around lane center. The camera height is set randomly between 140cm and 190cm and a downward pitch between 0 and 5 degrees is selected. Finally, illumination is set and the scene is rendered from the camera view. The 3D points of each lane centerline and delimiter are translated to camera coordinates to generate the ground truth.

Each generated example consists of an image ( $360 \times 480$  pixels) and its associated ground truth: 3D lanes, camera height and pitch. Figure 2(Left and middle columns) presents several examples showing the resulting diversity and complexity. The exact parameters used in the random generation process are listed in Appendix I. The generated dataset contains 300K train and 5K test examples. An additional 1K validation set was used for learning rate scheduling and choosing the best performing snapshot.

### 4.2. Real-world 3D lane dataset

Acquiring ground truth labeled data with 3D for the task is an endeavor that requires a complex multiple-sensor setup and possibly also expensive HD maps. To this end we introduce a new such dataset, *3D-lanes*, created using a multi-sensor setup including a forward-looking camera, a Velodyne HDL32 lidar scanner and a high-precision IMU all synchronized and accurately aligned. The data was collected in 6 drives each on a different road segment totaling nearly 2 hours of driving. Using the Lidar and IMU we generate aggregated lidar top view images as in [33], which are then used together with a semi-manual annotation tool for generating ground truth. In total, 85K images were annotated, out of which 1K, consisting of a separate drive, were used as the test set and the remaining as the train set. The lidar information is additionally used to provide the full 3D curve of each lane. A disadvantage of this approach is that lanes not sufficiently visible to the lidar, due to occlusions or limited resolution at distance, are missing from the ground truth. Therefore, the labeling is somewhat noisy as can be observed in Fig. 2(Right column). In addition, the dataset variability in terms of geometry and topology is modest

compared to the *synthetic-3D-lanes* dataset. We therefore used the synthetic data which has perfect ground truth to develop the method and conduct ablation studies while the real-world dataset is used for validating transferability of the approach to real data and qualitative analysis.

#### 4.2.1 Evaluation results

**Evaluation metrics.** We propose an evaluation of 3D lane detection that separates the detection accuracy from the geometric estimation accuracy. *Detection accuracy* is computed via the standard average precision (AP) measure of the precision-recall curve. We first compute a curve to curve distance between a GT and a detected lane as a weighted sum of point-wise Euclidean distances. We measure distances on a set of predefined y-values along the curves, every 80cm in the range 0-80 meters. Weight is decreased for farther away points. We then perform a one-to-one (curve) matching by choosing pairs in decreasing similarity. A matching is considered correct if the weighted distance is below a certain, rather permissive, threshold (1.5 meters). Iterating over lane confidence thresholds we generate the precision-recall curve.

For matched detections we assess the *geometric estimation accuracy* by measuring the distribution of the error (point-wise Euclidean distance) over the same points used to measure the curve to curve distance. We further divide the entire dataset to lane points in the **near range** (0-30m) and **far range** (30-80m) due to the differences in the magnitude of errors. We then compute per range the  $1\sigma$  error, as the 68 error percentile and the  $2\sigma$  error as the 95 percentile. Lane centerline and delimiter detection are separately evaluated using this methodology. Irrelevant lane points are ignored in the evaluation as in the training phase.

**Implementation details.** 3D-LaneNet and all the variants brought in this section were initialized and trained using an identical protocol. The image-view pathway is initialized from VGG16 trained on imagenet [7]. We train with Adam optimization [17] and with initial learning rate  $5 \cdot 10^{-4}$ . We use a variation on the cyclic learning rate regime described in [32] with a minimal learning rate of  $10^{-6}$ . The y-range of the top view representation is 80 meters and the x-range is 20 meters. The IPM scale is different in x and y: in the first top-view feature map each pixel corresponds to 16cm laterally (x) and 38.4cm longitudinally (y). The last top-view feature map is  $\times 8$  smaller and since there is an anchor per column the distance between anchors is  $16 \times 8 = 128\text{cm}$ . We set the  $K (= 6)$  vertical reference points to be  $y = \{5, 20, 40, 60, 80, 100\}$  and  $Y_{ref} = 20\text{m}$ .

**Results on synthetic-3D-lanes dataset.** Typical network results on the test set are shown in Figure 2, with ground truth marked. The first row in Table 1 shows the quantitative results of 3D-LaneNet for centerline detection. A valid

concern with synthetic datasets is that its variability is too limited and that the learned network memorizes the entire example space instead of learning to generalize. A positive indication that this is not the case is that test AP (0.952) is well below train AP (0.966) as are the geometric evaluation measures. All networks trained in the ablation tests presented here were initialized from VGG16 just as the 3D-LaneNet was and were trained with the same training parameters and number of epochs.

We first examine the role of the *dual-pathway architecture* by comparing it to alternative architectures. The **image-view** only version connects the image-view pathway directly to the lane detection head which outputs the representation in  $\mathcal{C}_{road}$  exactly as 3D-LaneNet does. The anchor positions  $X_A$  in this case are determined by the columns in the last feature map: for each column we pick a pixel at a predefined image y-coordinate and project it to top-view to determine the anchor corresponding to the column. The **top-view** only version first projects the *image* itself to top view and continues the same computation as the top-view pathway. In addition, we tested two versions which include a limited version of the dual-pathway. The **early IPM** includes a single dual context module (the first amongst the four in the full network). The **late IPM** similarly contains only the last dual context module out of the four. The results, summarized in table 1, show that the full dual-pathway architecture has superior performance compared to all other variants. In particular, the worst result is delivered by the **image-view** only version, stressing the importance of the top-view processing pathway. Note that the **late stage IPM**, consisting of a trimmed version of the dual pathway, delivers the second best accuracy, but with a reduced computational cost, making it a good candidate for real-time implementations.

We also tried alternative *definitions of the road projection plane*. One approach takes into consideration the entire scene when fitting the road plane and not only the local road normal. To test it we devised a ground truth generation algorithm which takes the farthest visible road point and connects it to the local road position to determine the pitch. This method, is termed **horizon** in Table 1 since it resembles horizon estimation methods. Evidently, it performed slightly worse in general, although we observed consistently cases in which the scene topography favors this definition. We also tried assuming a **fixed position** of the camera, in which the average pitch ( $2.5^\circ$ ) and camera height (165cm) were used to define  $T_{c2r}$ . Finally, we note that learning to predict the best road projection plane per scene *without explicit supervision*, as proposed in [25], failed to produce satisfying results for our task.

The last row in Table 1 (**flat ground**) is brought to stress the importance of full 3D lane estimation compared to the current existing approach: image-only detection and

*image-to-world* translation using the flat-ground assumption. Image-only detection is obtained by projecting the 3D-LaneNet results to the image plane. For the image-to-world stage we need to choose the plane to project the image result to. We tried two options, both computed using the ground truth: the road plane  $P_{road}$  and the plane defined by **horizon** as described in the previous experiment. As one may expect, the **horizon** based method, which essentially uses the best planar fit for the entire scene, produced the better results, which are still inferior to those of 3D-LaneNet which performs full 3D estimation.

The *delimiter detection* performance obtained by 3D-LaneNet is 0.971 AP (positional errors: 12.9cm@1 $\sigma$ , 33cm@2 $\sigma$  near range; 30cm@1 $\sigma$ , 106cm@2 $\sigma$  far range). These metrics show a slightly better performance compared to centerline detection. A possible explanation is that delimiters are clearly marked on the road while centerlines are indirectly inferred. Since output is transformed from road to camera coordinates using an estimated  $T_{c2r}$  we also measured the quality of this estimation and its effect on the results. The median values of the absolute errors for pitch ( $\theta$ ) and camera height ( $h_{cam}$ ) are 0.09° and 2.4cm respectively. To eliminate the contribution of this error we evaluated performance in road coordinates  $C_{road}$  by taking the raw network output (before transforming to  $C_{camera}$ ) and got a negligible difference in measured performance.

**Results on 3D-lanes dataset.** For operation on real-world data we trained the 3D-LaneNet on the train part of the *3D-lanes* dataset. Result examples from the respective test set are shown in Fig. 2 (Right column). Note that since the camera is mounted with a downward pitch the 3D lanes are detected as rising upward. Evaluation metrics are presented in Table 2. As in the synthetic data, using the **flat ground** assumption on the real data degrades performance, achieving a 4 times larger error in the far range.

Table 1. Centerline detection results on *synthetic-3D-lanes* dataset

	AP	Error <b>near</b> (cm)		Error <b>far</b> (cm)	
		1 $\sigma$	2 $\sigma$	1 $\sigma$	2 $\sigma$
<b>3D-LaneNet</b>	<b>0.952</b>	<b>13.3</b>	<b>34.4</b>	<b>33.1</b>	<b>122</b>
<b>image-view</b>	0.819	20.3	50	74.7	241
<b>top-view</b>	0.929	17.5	39.6	49.5	208
<b>early IPM</b>	0.934	13.7	35.5	43.5	189
<b>late IPM</b>	0.948	14.5	37.2	37.4	139
<b>horizon</b>	0.949	14.8	40.4	36.7	132
<b>fixed position</b>	0.948	13.6	37.3	35.4	139
<b>flat ground</b>	0.566	46.9	114	99	289

#### 4.3. Evaluation of image-only lane detection

The purpose of this experiment is to compare our approach to the current state of the art, which exists for image-

Table 2. Delimiter detection results on *3D-lanes* dataset

	AP	Error <b>near</b> (cm)		Error <b>far</b> (cm)	
		1 $\sigma$	2 $\sigma$	1 $\sigma$	2 $\sigma$
<b>3D-LaneNet</b>	<b>0.918</b>	<b>7.5</b>	<b>19.6</b>	<b>12.4</b>	<b>33</b>
<b>flat ground</b>	0.894	19.1	37.4	64.1	137

only lane detection. The tuSimple lane dataset [1] consists of 3626 training and 2782 test images. Unfortunately, today there is no access to the labels for the test images. We therefore divide the original training set to our own train/validation sets (90% train and 10% validation). While we are aware that there may be deviations between our evaluation (obtained on the validation set) and the one on the test set, we can expect a similar performance and reach the same conclusion qualitatively. Since this dataset does not contain 3D information, we train a variation of 3D-LaneNet, which detects the lanes in the image domain. Instead of a 3D representation, the network output was reduced to 2D points on the road projection plane by eliminating the elevation ( $z_t^i$ ) component. Only the delimiter output type is maintained ( $t = d$ ) since the marked entities in the dataset are lane delimiters. A fixed homography,  $H_{tuSimple}$ , between image plane and road projection plane, was manually selected, such that straight lanes become parallel in top view. The lanes directly predicted by the network are transformed to lanes in the image view using  $H_{tuSimple}$ . Since  $H_{tuSimple}$  is fixed, the road projection plane prediction branch is not used. Other than the aforementioned, the network is identical to the 3D-LaneNet as configured for the synthetic-3D-lanes dataset. The tuSimple main evaluation metric (*acc*) [1] is the average ratio of detected ground truth points per image. Using our end-to-end approach on our validation set we reached an accuracy of 0.951 which is competitive with the one achieved by the tuSimple 2017 competition winning method [27], (0.965). This result is encouraging and somewhat surprising given that our entire approach was designed towards the 3D estimation task. In particular, our geometric loss (Eq. 1) is computed in top view coordinates, giving in practice a much higher weight to distant lane points while in the tuSimple *acc* metric all points equally contribute.

## 5. Conclusions

We presented a novel problem, 3D multiple lane detection, along with an end-to-end learning-based solution, 3D-LaneNet. The approach has been developed using a newly introduced synthetic dataset and validated on real data as well. The approach is applicable in principle to all driving scenarios except for complex urban intersections. Finally, we believe that the *dual-pathway* architecture can facilitate additional on-road important 3D estimation tasks such as 3D vehicle detection.



## References

- [1] <http://benchmark.tusimple.ai>, lane challenge.
- [2] <https://www.blender.org/>.
- [3] Mohamed Aly. Real time detection of lane markers in urban streets. In *IVS*, pages 7–12, 2008.
- [4] Aharon Bar Hillel, Ronen Lerner, Dan Levi, and Guy Raz. Recent progress in road and lane detection: a survey. *Machine Vision and Applications*, 25(3):727–745, Apr 2014.
- [5] Amol Borkar, Monson Hayes, and Mark T. Smith. Robust lane detection and tracking with ransac and kalman filter. In *ICIP*, pages 3261–3264, Nov 2009.
- [6] Pierre Coulombe 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 vol.2, 2002.
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.
- [8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.
- [9] Mohsen Ghafoorian, Cedric Nugteren, Nóra Baka, Olaf Booi, and Michael Hofmann. El-gan: Embedding loss driven generative adversarial networks for lane detection. *CoRR*, abs/1806.05525, 2018.
- [10] R. Gopalan, T. Hong, M. Shneier, and R. Chellappa. A learning approach towards detection and tracking of lane markings. *IEEE Transactions on Intelligent Transportation Systems*, 13(3):1088–1098, Sept 2012.
- [11] Richard I. Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.
- [12] Bei He, Rui Ai, Yang Yan, and Xianpeng Lang. Accurate and robust lane detection based on dual-view convolutional neural network. In *IVS*, pages 1041–1046, June 2016.
- [13] Brody Huval, Tao Wang, Sameep Tandon, Jeff Kiske, Will Song, Joel Pazhayampallil, Mykhaylo Andriluka, Pranav Rajpurkar, Toki Migimatsu, Royce Cheng-Yue, Fernando Mujica, Adam Coates, and Andrew Y. Ng. An empirical evaluation of deep learning on highway driving. *CoRR*, abs/1504.01716, 2015.
- [14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015.
- [15] Jihun Kim and Minho Lee. Robust lane detection based on convolutional neural network and random sample consensus. In *Neural Information Processing*, pages 454–461, 2014.
- [16] Jiman Kim and Chanjong Park. End-to-end ego lane estimation based on sequential transfer learning for self-driving cars. In *CVPR Workshops*, pages 1194–1202, July 2017.
- [17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [18] Seokju Lee, Junsik Kim, Jae Shin Yoon, Seunghak Shin, Oleksandr Bailo, Namil Kim, Tae-Hee Lee, Hyun Seok Hong, Seung-Hoon Han, and In So Kweon. Vpnet: Vanishing point guided network for lane and road marking detection and recognition. In *CVPR*, pages 1947–1955, 2017.
- [19] Jun Li, Xue Mei, Danil Prokhorov, and Dacheng Tao. Deep neural network for structural prediction and lane detection in traffic scene. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):690–703, March 2017.
- [20] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, pages 5162–5170, 2015.
- [21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Max Welling, and Nicu Sebe, editors, *ECCV*, pages 21–37, Germany, 1 2016. Springer.
- [22] Hanspeter A. Mallot, Heinrich H. Bülthoff, James Little, and Stefan Bohrer. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological Cybernetics*, 64(3):177–185, Jan 1991.
- [23] A. Meyer, N. Salscheider, P. Orzechowski, and C. Stiller. Deep semantic lane segmentation for mapless driving. In *IROS*, Oct 2018.
- [24] Sergiu Nedevschi, Rolf. Schmidt, Thorsten Graf, Radu Danescu, Dan Frentiu, Tiberiu Marita, Florin Oniga, and Ciprian Pocol. 3d lane detection system based on stereo-vision. In *IEEE Conference on Intelligent Transportation Systems*, pages 161–166, Oct 2004.
- [25] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. *CoRR*, 1802.05591, 2018.
- [26] Gabriel L. Oliveira, Wolfram Burgard, and Thomas Brox. Efficient deep models for monocular road segmentation. In *IROS*, pages 4885–4891, Oct 2016.
- [27] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial CNN for traffic scene understanding. In *AAAI*, pages 7276–7283, 2018.
- [28] Dean Pomerleau. Ralph: rapidly adapting lateral position handler. In *IVS*, pages 506–511, Sept 1995.
- [29] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.
- [30] Shital Shah, Debadepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017.
- [31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [32] Leslie N. Smith. Cyclical learning rates for training neural networks. In *WACV*, pages 464–472, March 2017.
- [33] Chris Urmson, Joshua Anhalt, Drew Bagnell, Christopher R. Baker, Robert Bittner, M. N. Clark, John M. Dolan, Dave Duggins, Tugrul Galatali, Christopher Geyer, Michele Gittelman, Sam Harbaugh, Martial Hebert, Thomas M. Howard, Sascha Kolski, Alonzo Kelly, Maxim Likhachev, Matthew

McNaughton, Nick Miller, Kevin M. Peterson, Brian Pilnick, Raj Rajkumar, Paul E. Rybski, Bryan Salesky, Young-Woo Seo, Sanjiv Singh, Jarrod M. Snider, William Whittaker, Ziv Wolkowicki, Jason Ziglar, Hong Bae, Thomas Brown, Daniel Demitrish, Bakhtiar Litkouhi, Jim Nickolaou, Varsha Sadekar, Wende Zhang, Joshua Struble, Michael Taylor, Michael Darms, and Dave Ferguson. Autonomous driving in urban environments: Boss and the urban challenge. *Journal of Field Robotics*, 25(8):425–466, 2008.

- [34] 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.