This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Structured Bird's-Eye-View Traffic Scene Understanding from Onboard Images

Yigit Baran Can[1] Alexander Liniger[1] Danda Pani Paudel[1] Luc Van Gool[1,2]
[1]Computer Vision Lab, ETH Zurich [2]VISICS, ESAT/PSI, KU Leuven
{yigit.can, alex.liniger, paudel, vangool}@vision.ee.ethz.ch

Abstract

Autonomous navigation requires structured representation of the road network and instance-wise identification of the other traffic agents. Since the traffic scene is defined on the ground plane, this corresponds to scene understanding in the bird's-eye-view (BEV). However, the onboard cameras of autonomous cars are customarily mounted horizontally for a better view of the surrounding, making this task very challenging. In this work, we study the problem of extracting a directed graph representing the local road network in BEV coordinates, from a single onboard camera image. Moreover, we show that the method can be extended to detect dynamic objects on the BEV plane. The semantics, locations, and orientations of the detected objects together with the road graph facilitates a comprehensive understanding of the scene. Such understanding becomes fundamental for the downstream tasks, such as path planning and navigation. We validate our approach against powerful baselines and show that our network achieves superior performance. We also demonstrate the effects of various design choices through ablation studies. Code: https://github.com/ybarancan/STSU

1. Introduction

Road scene understanding is crucial for autonomous driving since it forms the interface between perception and planning. The fundamental task is to understand both the road network structure and the other traffic agents in the surrounding. Currently, the go-to solution is offline generated HD-maps combined with a modular perception stack [22, 41, 30, 36, 8]. For existing solutions to work, not only the precise localization in the HD-map but also understanding the dynamic parts of the scene is necessary [30, 44]. To achieve these requirements, most solutions use several sensors, including cameras and LIDAR. However, using expensive sensors and offline HD-maps limit the scalability of autonomous driving as they increase the cost of operation and limit self-driving cars to operate in geographically restricted areas.

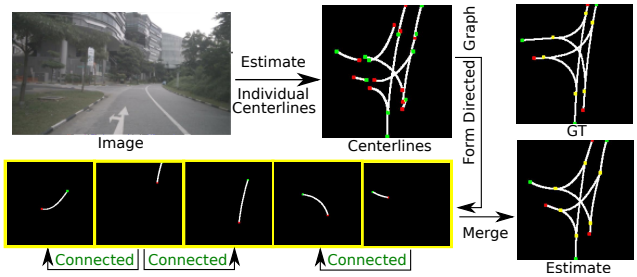

Figure 1. Our method extracts a directed graph that represents the local road network from a single frontal image. First, centerlines are estimated alongside a directed graph where vertices are centerlines, and edges show the connectivity. Then the existence and direction of the edges are estimates. Green dots indicate start points, and red dots indicate end points of centerlines. Traffic flows from green to red. This representation can be achieved thanks to the directional nature of the proposed method.

In this work we proposes an end-to-end vision method that performs lane-graph extraction and object detection given only a front-facing camera image. Our method directly estimates the graph structure of the road network and spline curves representing centerlines of individual lanes, as shown in Fig. 1. Besides estimating the road graph, our model can also detect objects such as cars, pedestrians, and others, directly on the BEV plane, as shown in Fig. 2. The output format of our method is ideal for downstream planning [2, 9] and prediction [12, 45, 20, 37] tasks, which require both the lane-graph and the location and class of objects. In fact, such a requirement can also be understood simply by observing the provided labels of existing datasets, such as [4], which provide the labels in a structured form. Often, existing approaches map the structured labels into other forms, such as semantic masks, to perform scene understanding [14]. The downstream tasks, however, require the structured form of these understandings [26, 34, 25, 19].

Understanding HD-maps is a challenging problem, mainly due to the complex topological changes. Recovering such topological structure coherently from a single image remains to be an unexplored problem. This work addresses this challenging problem for the first time while also detecting objects in the scene directly in the BEV coordinates.

Existing works either focus on (i) HD-map extraction from dense 3D points [19] or (ii) the detection of road lanes from a single image [21]. Other variants, such as BEV semantic understanding, also exist [40, 29, 32]. Note that the HD-map reconstruction of [19] is much more topologically challenging than the lane detection problem of [21]. Our work aims to achieve results similar to [19] using the image input setup of [21]. Additionally, we aim to detect objects using the same model as for structured HD-map predictions.

We represent the HD-map as a directed graph in BEV coordinates, whose edges are the road segments and the direction represents the traffic flow. We model each road segment using a Bezier curve, with starting and end points. The connections between the predicted segments are modeled using an assignment matrix. For the prediction, we make use of a transformer network, which is supervised by using the Hungarian algorithm at the output end. The predicted segments, along with their connectivity, defines a full lane graph HD-map. Our transformer network further predicts the parameters of 2D BEV objects. The object prediction branch is supervised, similar to the road segments. Two example outputs of our method for both lane graph HD-map and object estimation are shown in Fig. 2. To this end, our major contributions can be summarized as follows.

- We propose a unified method for structured BEV road network graph estimation and object detection from a single onboard monocular camera image.
- The results obtained by the proposed method are significantly superior to the compared baselines.

2. Related Works

Road network extraction: Early works on road network extraction use aerial images [13, 39]. Building upon the same setup, recent works [3, 42, 43] perform the network extraction more effectively. However, aerial imaging-based approaches only provide coarse road networks. Such predictions may be useful for routing, however, they are not accurate enough for action planning.

High definition maps: In the literature, HD-maps are often reconstructed offline using aggregated 2D and 3D visual information [25, 18, 26]. Although these works are the prime motivation behind our work, they require 3D point clouds for accurate HD-map reconstruction. More importantly, the offline methods recover the HD-maps in some canonical frame. Thus, using the recovered maps requires accurate localization, in many cases. In this regard, our work is similar to [17], where the lane boundaries are detected on highways in the form of polylines. An extension of this work [17] uses a recurrent neural network to generate initial boundary points in 3D point clouds. The initial points are then used as seeds for a Polygon-RNN [1] that predicts lane boundaries. Our method differs from [17] in two major aspects: (i) point clouds vs. single image input, (ii) highway lane boundaries vs. lane centerlines in an unrestricted setting.

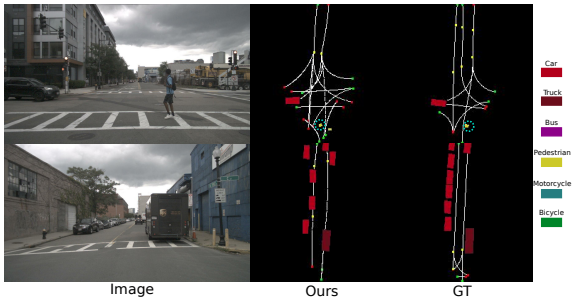

Figure 2. Our method can handle very complex cross roads scenes as well multiple object instances. Pedestrian is marked with circle.

BEV semantics understanding: Because of its practical use, scene understanding in BEV using images has recently gathered significant attention [40, 35, 5]. Some methods also combine images with LIDAR data [33, 16]. In this regard, methods developed in [40, 31] use a single image to understand the BEV HD-map semantics. Similarly, the method proposed in [5] uses video data for the same task. Methods in this category do not offer structured output suitable for many downstream tasks. These methods may be used for general scene understanding. However, their usage for the task of motion planning and navigation is not trivial. Furthermore, up to our knowledge, no existing method provides instance-level predictions on the BEV using single image input. Note that the method proposed in this paper predicts both HD-map and the road objects simultaneously, using one input image and a single neural network.

3. The Proposed Method

The core task of our model is to produce a directed graph that represents the road network in a BEV coordinate system, given only a single image from a front-facing camera mounted on a vehicle. For the complete traffic scene understanding, our model also outputs objects' instances in the form of BEV bounding boxes. Both these tasks require reasoning about the 3D space and projecting all the information on the BEV ground plane where the vehicle is moving.

In this section, we first introduce our trainable lane graph structure and describe the object representation. Given these building blocks, we introduce our transformer based model and explain how the neural network is trained.

3.1. Lane graph representation

In order to have a structured representation of the local road network, we build a directed graph of lane centerline segments, often called the lane graph. Let this directed graph be $G = (V, E)$ where V are the vertices of the graph (the centerlines) and the edges $E \subseteq \{(x, y) \mid (x, y) \in V^2\}$ represent the connectivity among those centerlines. The connectivity can be summarized by the incidence matrix I

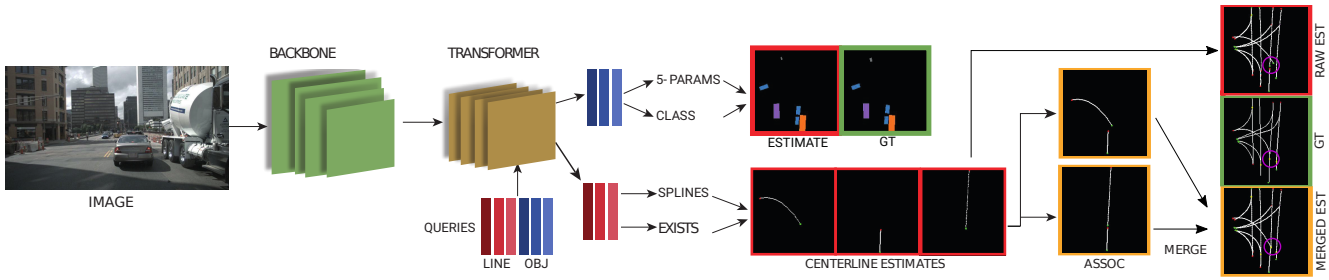

Figure 3. The core architecture of our neural network is a transformer [7] that processes learned centerline and object queries together. The processed line queries are used to output detection probability, control points, and centerline association features. The object queries are used to calculate the class probability and the oriented box parameters.

of the graph $G = (V, E)$. Thus, let us define when two centerlines x and y are connected; A centerline x is connected to another centerline y, i.e. $(x, y) \in E$ if and only if the centerline y's starting point is the same as the end point of the centerline x. Given this definition, an entry of the incidence matrix $I[x, y] = 1$ if the centerlines x and y are connected. Note that we do not apply a hard requirement to generate acyclical graphs, but cycles rarely occur due to our focus on a single image with a limited field-of-view (FOV). Thus, the incidence matrix often has the structure of an acyclical graph, where the main diagonal is zero and the sum of symmetric entries is at most one. Finally, the resulting incidence matrix also contains crucial information about the traffic flow directions, which is fundamental to understand a lane network.

With the graph established, we need to model each centerline (vertex of the graph) mathematically. In this work, we consider each centerline as a Bezier curve. A Bezier curve maps a scalar parameter $t \in [0, 1]$ to a point in $\mathbb{R}^{\Delta}$. We are interested in 2D curves for our lane graph, thus $\Delta = 2$. The curve can be written as the weighted sum of control points $P = \{P_0, P_1, ..., P_n\}$ where $P_i \in \mathbb{R}^2$. Given the control points, the curve B parameterized by t is defined as $B(t) = \sum_{k=0}^{n} \binom{n}{k}(1-t)^{n-k}t^k P_k$. A more compact matrix-based formulation is simply $B(t) = \mathbf{\Gamma}(t, n)\boldsymbol{P}$, where $\mathbf{\Gamma}_{ij}(t, n) = \binom{n}{j}(1-t_i)^{n-j}t_i^j$ represents the weight matrix, and $\boldsymbol{P} = [P_0, P_1, ..., P_n]$ is the vector of all control points. With this representation, finding the optimal control points given some observed points $\boldsymbol{Y} = [Y_0, Y_1, ..., Y_T]$ amounts to solving a least square problem, i.e., $\boldsymbol{P}^* = \arg\min_{\boldsymbol{P}} \|\mathbf{\Gamma}(t)\boldsymbol{P} - \boldsymbol{Y}\|$. Bezier curves are a good fit for centerlines since it allows us to model a curve of arbitrary length with a fixed number of 2D points. Thus, given our graph and centerline representation, the whole lane graph has a fixed-sized learnable representation, where the network can learn the centerlines in terms of Bezier control points, and the connectivity of the graph.

3.2. Object representation

Our method also produces object instance detections to complement the lane graph and give a complete traffic scene understanding. Different than semantic segmentation, instance outputs localize and identify individual objects. We represent each instance as a 2D box in normalized BEV coordinates. In order to fully specify such a box, one needs five parameters: location of the center point, short and long side length, and the heading angle. From these parameters, it is a simple conversion to the four corner point locations and vice versa. Apart from the localization and orientation of the instances, we also produce their semantics/object class using a one hot representation. Given this representation, an instance is fully identified.

3.3. Architecture

We have modeled each centerline and object instance as a fixed size vector. Thus, we can work within the framework of proposal generation and classification. This has been widely used in the fields of instance segmentation and object detection [38, 15, 7]. One crucial property of our formulation is the strong relationship among different centerlines as they form the lane graph. However, there is also a strong relation between centerlines and objects since, in traffic scenes, objects follow centerlines. In order to fully exploit this dependency, we adapt the transformer-based model proposed in [7], which allows us to train one joint model for lane graph and object understanding.

The transformer-based object detector proposed in [7] uses image backbone features and learnable query vectors to generate object proposals. We follow a similar approach, but we use two sets of learned query vectors $Q \in \mathbb{R}^C$: one set for centerlines and one for objects. The number of these vectors is higher than the maximum centerlines/objects that can occur in any scene. These query vectors are processed *jointly* by the transformer, which outputs a proposal vector for each query. These vectors encode all the information needed to fully identify a centerline or object. Each of these proposal vectors is further processed to generate an output. This processing is done in a separate lane and object branch, which output the lane graph and object detections. The overall architecture is given in Fig. 3.

3.3.1 Lane branch

The first of the two branches processing the output vectors of the transformer is the lane branch, which has four parts:
Detection head: The transformer output is processed by a multi-layer perceptron (MLP) with an output softmax layer. This output gives the probability that the centerline encoded by the corresponding query vector exists.
Control head: An MLP + sigmoid layer with $2 \times R$ output neurons, encoding the R Bezier curve control points.
Association head: An MLP that outputs a δ-dimensional association feature vector for each of the centerline vectors, where $\delta < C$. The classifier uses these association features to establish the connectivity of the estimated centerlines.
Association classifier: An MLP + sigmoid layer, which takes two δ dimensional association feature vectors corresponding to two centerlines as an input. This layer outputs the probability of the input centerline pair being associated.

As a first step of estimating the graph, we extract individual centerlines. This is done by the detection and control heads. These lines form the vertices V of the graph $G = (V, E)$. Given that N centerlines are selected, the feature vectors of the corresponding centerlines are processed by the association head to produce lower-dimensional association feature vectors $F \in R^{N \times \delta}$. Then, we obtain the association inputs $A \in \mathbb{R}^{N \times N \times 2\delta}$, where $A_{ij} = \text{concat}([F_i, F_j])$. This input encodes the directional nature of the graph. An MLP processes the matrix A to produce the incidence matrix probabilities. Note that the MLP has an input dimension of 2δ, and $N \times N$ is the batch size. Thus there is no constraint on the number of proposed centerlines.

During training, we first output centerline control points and detection probabilities and apply the Hungarian matching algorithm among the estimated and the ground truth (GT) centerlines. The association step is carried out on the matched estimates. During inference, we threshold the detection probability of the centerlines and carry out the association step on the active lines.

3.3.2 Object branch

The second branch that processes the transformer proposal vectors is the object branch. The branch consists of two modules and an optional post-processing network.
Detection head: The transformer output is processed by an MLP with a softmax output layer to produce class probability distribution, including a "no detection" class.
5-params head: An MLP + sigmoid layer that produces the normalized parameters of the oriented object boxes.
Refinement net: While the instance outputs are suitable for many tasks, it is also beneficial to produce semantic segmentation maps of the scene. This is especially true for small objects like pedestrians, and bikes, where the localization in the BEV from a single onboard camera is difficult. Therefore, we propose an optional post-processing network that converts the instance estimations to semantic segmentations using our refinement net. The structure of the refinement net is similar to the BEV decoder of [6], where a lower resolution input is upscaled to provide a fine-grained segmentation map. The network operates as follows: there are $C + 1$ classes including the background, and the region of interest is $H \times W$ dimensional. We first convert the 5-params output to a box and multiply this box with the class probabilities. This results in a matrix M of dimension $H \times W \times (C+1)$ where $\sum_i M_{h,w,i} = 1$ for all grid locations (h, w) that fall in the bounding box and 0 otherwise. Then we sum all these matrices and clip them to (0,1). The result is again of dimension $H \times W \times (C + 1)$. To inform refinement net about the visual cues in the image, we also include backbone features. Since the bounding box locations are in the BEV coordinates, we warp the backbone features to the BEV and use warped skip connections in the upsampling.

3.3.3 BEV positional embedding

Since transformers do not have a notion of position, positional embeddings (PE) are used to add spatial awareness. We use two different positional embeddings for the transformer. The first one encodes the image domain spatial information where similar to [7], we use sinusoidal functions on the normalized cumulative locations. The second set of positional embeddings encode the corresponding BEV location of a given pixel. For this purpose, we assume a flat surface where every real world-point has height $-C_h$, where C_h is the camera height. The resulting grid is very dense in image coordinates that correspond to real-world points close to the camera while it is sparse for further away positions. In order to provide more uniform location cues to the network, we use the logarithm of BEV locations. To generate the positional encoding, we take the cumulative sum, normalize, and convert it into a sinusoidal. We design the two positional embeddings (image and BEV) such that they are half of the channel size of the input feature map. Thus, we can add the image positional embeddings to half of the channels while BEV positional embeddings are added to the other half. The main reason for this design choice is that BEV coordinates are undefined for the upper half of the image, but they still hold important cues for the network about objects in this region. Simply adding these two positional embeddings would result in duplicates and inconsistent values in the lower half of the image. We dubbed this use of dual positional embeddings as *split positional encoding*.

3.3.4 Training

Since both the lane and object branches produce detection-like outputs, we use Hungarian matching on the estimations and the ground truth during training. The matching loss

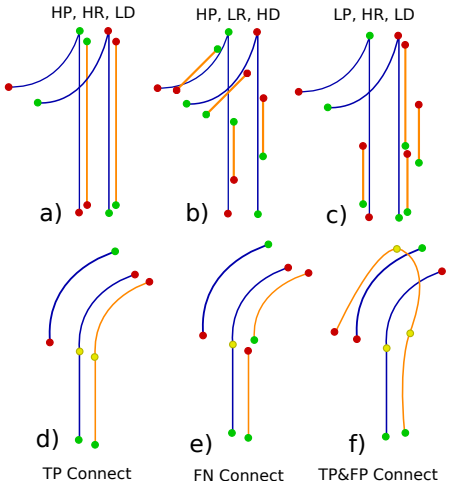

Figure 4. Some examples of precision-recall & detection (a,b,c) and connectivity metric (d,e,f). Blue lines are true centerlines and orange are estimations. Green/red dots represent starting and end points, respectively. Yellow points indicate a connection, which is only valid in the direction from green to red. H/L refers to High/Low while P=Precision, R=Recall and D=Detection. a) 2 out of 4 lines are missed but the matched lines are accurate. b) Matched true lines are longer than estimates creating false negatives. c) All estimates are matched to one true line (note the endpoint colors on the leftmost estimate), leaving no room for false negatives while creating false positives. d) and e) Show true positive and false negative connectivity, respectively. f) One connection is a true positive but the upmost connection is a false positive.

used for lanes and objects is similar. Both are of the form $L_m = L_{CE} + \lambda L_1$, where L_{CE} is the cross-entropy loss on the detection/class probability, and L_1 is the 1-norm loss on the Bezier control point locations/Box parameters for lines and objects, respectively. For both the centerline and object branch, the training detection loss is cross-entropy. Control point and object 5-params outputs are trained using L_1 loss, except for the angle of objects. Since 180° object flips are hard to distinguish, we predict only angles in the range $\alpha \in [0, \pi]$ and train it using a smooth sin/cos L_1 loss of the form $L_{angle} = |\cos(2\alpha) - \cos(2\phi)| + |\sin(2\alpha) - \sin(2\phi)|$, where ϕ is the GT angle. We train the refinement net independently using the cross-entropy loss, not backpropagating through the rest of the network.

4. Metrics

Since our problem setup is new, there is a lack of suitable metrics for performance evaluation. We wish to measure the performance in reproducing the real directed graph faithfully. For this purpose, we use three metrics that aim to highlight different aspects of the directed graph.

4.1. Precision-Recall

Following [24, 17], we calculate precision-recall on matched centerlines at different distance thresholds. We first match each estimation to the target with the minimum L_1 loss on Bezier coefficients. Thus, similar to [24], multiple estimations can be matched with the same target while each estimation can only be matched with one target. Then we interpolate the estimated coefficients to get dense centerlines. Note that using control points during matching is fundamentally different from using interpolated points because the control point based matching takes direction into account. Thus, two centerlines where only the order of control points is reversed (start and end points are swapped) are identical if interpolated points are matched. However, they are far apart in our control point matching approach. After matching based on the control points and then interpolation, a true positive is an estimated interpolation point within a threshold distance to the matched GT line and a false positive otherwise. A false negative is a point on a GT line that is not within the threshold distance of any of the matched estimated lines. Note that this metric does not penalize the missed centerlines, i.e., true centerlines that are not matched with any estimation. This is intentional since the focus of this metric is measuring how well the estimates fit the matched GT centerlines and how accurately the captured subgraph is represented.

4.2. Detection ratio

In order to measure the aforementioned issue of missed centerlines present in the precision-recall metric, we calculate the detection ratio. This is simply the number of unique GT centerlines that at least one estimated line is matched to over the total number of GT centerlines. High scores in precision-recall and a low detection score mean that the estimated centerlines are close to the matched true ones, but a substantial part of the GT centerlines is not detected. The inverse implies that the estimated centerlines cover the true road network but do not faithfully represent the structure. These two metrics summarize the performance on the vertices of the total graph G. However, we still lack a metric to evaluate performance on the edges of the graph, i.e., the connectivity.

4.3. Connectivity

In order to measure how well the estimated centerlines are associated, we propose a precision-recall-based metric, called connectivity metric. Let the estimated binary incidence matrix be E, and the GT incidence matrix be I. Let $M(i)$ be the index of the target that the ith estimation is matched to and $S(n)$ be the set of indices of estimations that are matched to target n. A positive entry E_{ij} is a true positive if $(M(i) == M(j)) \mid (I(M(i), M(j)) == 1)$, and a

false positive otherwise. On the other hand, a false negative is a positive entry of the incidence matrix $I_{m,n}$ where $\nexists\,(i,j) : ((i \in S(m))\ \&\ (j \in S(n))\ \&\ (E_{i,j} == 1))$.

This metric captures how close the connectivity pattern of the estimated graph is to the GT graph. With this metric, fragmenting a true centerline into multiple estimations is not a problem as long as they are associated. Some graphical illustrations of the three metrics are given in Fig. 4.

5. Experiments

5.1. Dataset

We use the NuScenes [4] dataset consisting of 1000 sequences recorded in Boston and Singapore. The sequences are annotated at 2Hz, and the dataset provides HD-Maps in the form of centerlines. The dataset also provides 3D bounding boxes of 23 object classes. For our experiments, we select the most frequent classes: car, truck, bus, pedestrian, bike, and motorcycle. We only use the front camera both for training and evaluations.

Given the set of real-world coordinates of a particular centerline, we first convert these coordinates to the camera coordinate system of the current reference frame. We resample these points with the target BEV map resolution and discard any point outside the region-of-interest (ROI). The points are then converted to normalized coordinates given the bounds of the ROI. This results in a set of points between [0,1], from which we extract the control points of the Bezier curve. The ground truth labels and the estimates are in normalized control point coordinates.

5.2. Implementation

We use images of size 448x800, and the target BEV area extends from -25 to 25m in x-direction and 1 to 50m in z-direction. BEV resolution is set to 25cm. Given the common structure of roads, the possible complexity of the curves that represent the centerlines segments is limited. Thus, we use three Bezier control points. We use two sets of 100 query vectors for centerlines and objects: one for right (Boston) and one for left-sided traffic (Singapore). The backbone network is Deeplab v3+ [10] pretrained on Cityscapes dataset [11]. The implementation is in Pytorch. The method runs with 11FPS without batching and including all association and refinement steps.

5.3. Baselines

Since there does not exist any method that deals with structured BEV road network estimation from a monocular image, we have generated two baselines. The first baseline is based on [17], where the authors generate lane boundaries from point clouds. We adapt their method to work with images and to output centerlines rather than lane boundaries.

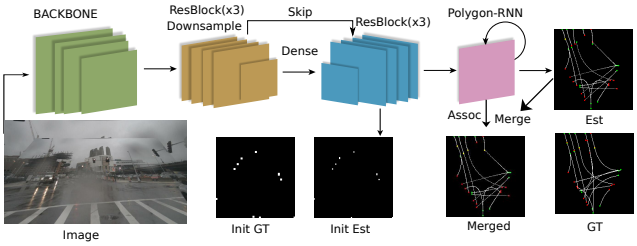

Figure 5. Polyline-RNN based method first extracts initial point estimations. Polygon-RNN uses the backbone features and initial point estimations to form the centerline curves.

To achieve this, we project Deeplabv3+ [10] backbone features of the image to the BEV with a GT projection matrix. We concatenate x-y grid locations with this backbone feature map, similar to [28]. A subnetwork with a fully connected layer at its core takes this input and outputs a grid of 49×50 points. We tested a convolutional RNN similar to the original work but did not achieve satisfactory results in our setting due to the restricted FOV. Note that the original task of finding lane boundaries on highways from aggregated LIDAR scans is significantly different from finding initial centerline points in urban traffic scenes. Moreover, the RNN required prohibitively many iterations, especially considering that one frame contains more than 40 centerlines. Therefore, we used a fully connected layer supported by several residual blocks, see Fig. 5. Given the initial locations and the backbone features, Polygon-RNN [1] produces the next control points of the centerline. We fix the number of iterations of Polygon-RNN to the number of spline coefficients used to encode centerlines. We use the focal loss [27] for the initial point estimation and an L_1 loss to supervise the control point estimation of Polygon-RNN. In training, Polygon-RNN uses GT initial points similar to [17]. For testing, we threshold the initial point estimations of the network and feed them to Polygon-RNN. To indicate the direction of traffic, we feed a binary indicator variable to the initial point estimator. The association estimation is done using the final feature map of Polygon-RNN using the same approach as in our network. For reference, we also report results with Polygon-RNN using GT initial points.

For our second baseline, we extract lane boundaries with the SOTA method of [23]. The extracted lane boundaries are projected onto the BEV grid using the GT transformation. Given these lane boundaries, we form the closest pairs and compute the centerlines using splines. Since this method does not give us a direction, we use the predicted and a flipped centerline version during evaluation, meaning PINET matching is directionless. In a sense, PINET estimates are manually assigned correct directions.

For the evaluation of our method's instance estimations, we compare against VED [29], VPN [32] and PON [40]. We use the same train/val split proposed in [40] for *both*

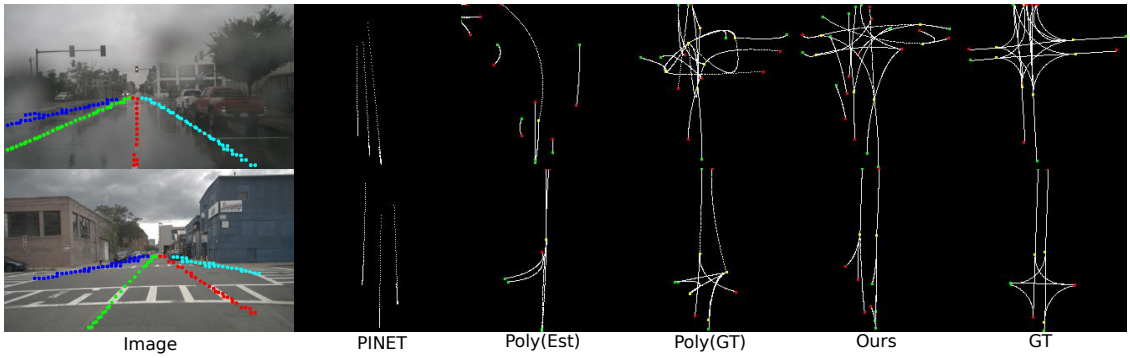

Figure 6. Sample centerline estimates. PINET boundary estimations are shown on the image. Our method produces the best lane graph representation. Statistical results for each sample are provided in the supplementary material.

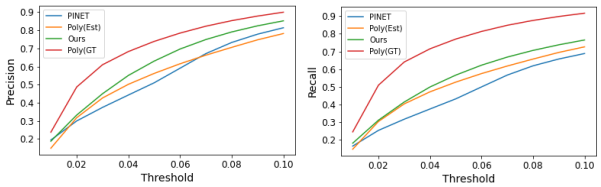

Figure 7. Precision/Recall vs thresholds. Thresholds are uniformly sampled in [0.01, 0.1] (normalized coordinates) with 0.01 increments. In our resolution, 0.01 corresponds to 50cm.

lane graph and object detection. We also follow [40] for the object label generation procedure. To compare results, we use the mIOU metric. Nevertheless, for future reference, we also present precision-recall results.

6. Results

Since our method produces a road network graph as well as dynamic object instance estimations, we divide the results into two subsections, studying those them individually.

6.1. Lane graph

The obtained results are provided in Tab. 1 and Fig. 7, where our method achieves the best results in all metrics when compared to the baselines. The performance of PINET is lower, as expected, since the centerlines are obtained through processing lane boundaries. From the Poly(Est) vs Poly(GT) results, it can be seen that the localization of initial points is very difficult. Our method produces better precision-recall than Poly(Est), and the difference in detection and connectivity scores are significant. It is not surprising that Poly(Est) suffers in the connectivity metric, particularly connectivity recall. This metric is closely related to detection score, and missed centerlines are penalized. Our method's performance in connectivity precision against Poly(GT) combined with the detection scores shows that our method produces much fewer false-positive associations in the detected sub-graph and more accurately estimates the graph. The superiority of Poly(GT) in precision-recall and detection metrics is expected. Since most centerlines are relatively short and divergence from the initial point is limited, knowing GT initial points provides a clear advantage. However, its performance validates the strength of the chosen baselines.

Method	M-Pre	M-Rec	Detect	C-Pre	C-Rec	C-IOU
PINET	54.1	45.6	19.2	-	-	-
Poly(Est)	54.7	51.2	40.5	58.4	16.3	14.6
Ours	**60.7**	**54.7**	**60.6**	**60.5**	**52.2**	**38.9**
Poly(GT)	70.0	72.3	76.4	53.8	52.0	36.0

Table 1. Lane graph results. M-Prec and M-Recall indicate mean of the sampled points of precision-threshold and recall-threshold curves, see Fig. 7. C-Prec and C-Rec refer to connectivity precision and recall, while C-IOU is connectivity TP/(TP + FP + FN).

Visual results for lane graphs are given in Fig. 6. Visual inspection shows that our method generally produces better results. In the last image, our method misses some centerlines. Overall, our method produces more faithful representations. On the other hand, Poly(GT) produces centerlines that are somewhat close, in the Euclidean sense, to the matched GT lines. However, the overall graph estimation is worse than ours. This shows the power of the connectivity metric where our method surpasses Poly(GT).

6.2. Objects

In Tab. 3, the refinement net outputs of our network are compared against SOTA methods. Other methods usually produce estimates for slightly more classes. However, considering that we produce structured instance outputs along with lane graphs, we chose the most common yet comprehensive set of classes. Our method surpasses PON in half of the classes and in the mean measure. Especially, the difference in the "car" category is rather significant.

The visual results for object estimates are given in Fig. 8. The competing methods tend to blob segmentation and

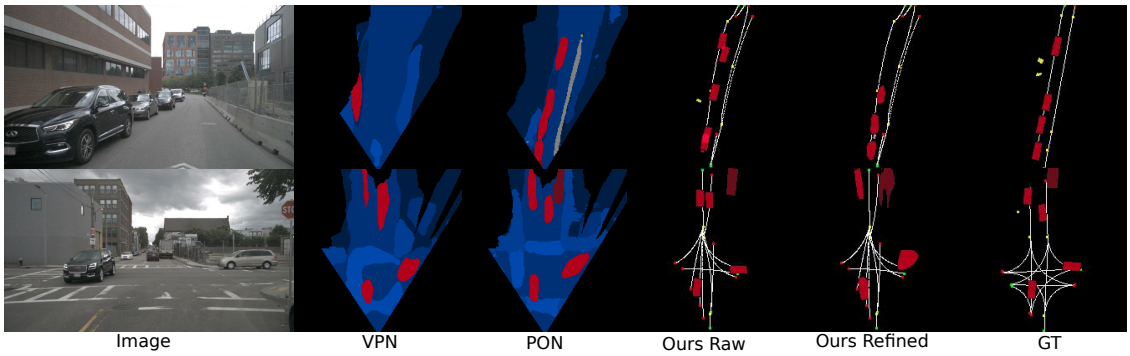

Figure 8. Visual results for object detection where we present the raw and refined estimates. We also show the road network estimates.

Method	M-pre	M-rec	detec	Con-IOU	car	truck	bus	ped	motor	bike	obj-mean
Large	57.2	53.9	58.8	41.0	20.0	**11.7**	13.9	1.9	2.2	1.4	8.5
Large + Split	59.9	**56.8**	52.8	40.8	20.0	10.1	16.8	1.9	2.8	0.8	8.7
Large + Split Log	**60.7**	54.7	60.6	38.9	21.8	11.0	14.5	2.1	3.8	2.1	9.2
Small	58.2	54.2	61.2	**41.9**	22.0	10.7	15.1	2.0	2.9	1.7	9.1
Small+Split	57.5	54.2	60.9	41.3	20.6	10.1	14.0	2.0	**4.1**	2.3	8.9
Small+Split Log	58.9	53.6	**61.5**	37.8	**22.6**	10.9	**17.6**	**2.4**	3.2	**2.9**	**9.9**

Table 2. Ablations are carried out on six models that test the performance contribution of the model size and positional embeddings. Object results are without refinement net and in the form of mIOU.

Method	car	truck	bus	ped	motor	bike	mean
VED	8.8	0.2	0.0	0.0	0.0	0.0	1.5
VPN	25.5	**17.3**	20.0	7.1	5.6	4.4	13.3
PON	24.7	16.8	20.8	**8.2**	7.0	**9.4**	14.5
Ours	**32.5**	15.7	**21.2**	6.2	**7.4**	6.4	**14.9**

Table 3. Object results in mIOU of different methods.

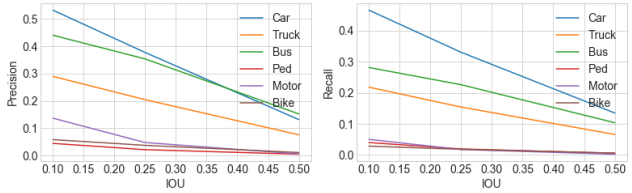

Figure 9. Precision/Recall vs IOU thresholds for object detection. We apply Hungarian matching with IOU to obtain corresponding estimate-GT pairs. If IOU is above the threshold, it is a true positive. Other GT objects count as false negatives, and the other estimates count as false positives.

making harder to separate instances. Our refinement net outputs also suffer from the same phenomenon compared to our raw estimates. Despite of which, our refined estimates strike a good trade-off between mIOU maximization and instance separation.

6.3. Ablation

We experimented with two transformer sizes. The small model has two encoder layers and tree decoder layers, while the large one has four encoder and four decoder layers. We tried using vanilla positional embeddings and our split embedding with and without taking the logarithm. The results are given in Tab. 2, where the object results are in mIOU *without* refinement net. We observe that our split embedding with log helps with objects, precision and detection scores while it causes a drop in connectivity IOU. Overall, the differences are rather low. Due to its good overall performance in object and lane results, we selected the "Large+Split Log" model as the final one. When the object results of the selected model are compared with and without refinement net, the difference is rather significant. Refinement net boosts the performance by 5.7 points in mIOU.

7. Conclusion

We proposed a novel learnable representation of local road networks based on directed graphs and Bezier curve centerlines. This representation is used to train a transformer-based neural network architecture that predicts a complete lane graph structure from a single onboard image. We also proposed a set of metrics that are suitable to evaluate the performance of the proposed graph representation based structured scene understanding. Additionally, along with the lane graph, our model also provides BEV object instances, thus offering a comprehensive understanding of the local traffic scene. Our extensive experimental comparisons with powerful baselines demonstrate the superior performance of the proposed method, in both lane graph and object detection tasks.

Acknowledgements: The authors gratefully acknowledge support by Toyota Motors Europe (TME).

References

[1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 859–868. IEEE Computer Society, 2018. 2, 6

[2] Mayank Bansal, Alex Krizhevsky, and Abhijit S. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In Antonio Bicchi, Hadas Kress-Gazit, and Seth Hutchinson, editors, *Robotics: Science and Systems XV, University of Freiburg, Freiburg im Breisgau, Germany, June 22-26, 2019*, 2019. 1

[3] Anil Batra, Suriya Singh, Guan Pang, Saikat Basu, CV Jawahar, and Manohar Paluri. Improved road connectivity by joint learning of orientation and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10385–10393, 2019. 2

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 1, 6

[5] Yigit Baran Can, Alexander Liniger, Ozan Unal, Danda Paudel, and Luc Van Gool. Understanding bird's-eye view semantic hd-maps using an onboard monocular camera. *arXiv preprint arXiv:2012.03040*, 2020. 2

[6] Yigit Baran Can, Alexander Liniger, Ozan Unal, Danda Pani Paudel, and Luc Van Gool. Understanding bird's-eye view semantic hd-maps using an onboard monocular camera. *CoRR*, abs/2012.03040, 2020. 4

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. 3, 4

[8] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. *arXiv preprint arXiv:2101.06806*, 2021. 1

[9] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on Robot Learning (CoRL)*, 2020. 1

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 833–851. Springer, 2018. 6

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[12] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *International Conference on Robotics and Automation (ICRA)*, pages 2090–2096, 2019. 1

[13] A Fortier, Djemel Ziou, Costas Armenakis, and S Wang. Survey of work on road extraction in aerial and satellite images. *Center for Topographic Information Geomatics, Ontario, Canada. Technical Report*, 241(3), 1999. 2

[14] Sourav Garg, Niko Sünderhauf, Feras Dayoub, Douglas Morrison, Akansel Cosgun, Gustavo Carneiro, Qi Wu, Tat-Jun Chin, Ian Reid, Stephen Gould, et al. Semantics for robotic mapping, perception and interaction: A survey. *arXiv preprint arXiv:2101.00443*, 2021. 1

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2017. cite arxiv:1703.06870Comment: open source; appendix on more results. 3

[16] Noureldin Hendy, Cooper Sloan, Feng Tian, Pengfei Duan, Nick Charchut, Yuesong Xie, Chuang Wang, and James Philbin. Fishing net: Future inference of semantic heatmaps in grids. *arXiv preprint arXiv:2006.09917*, 2020. 2

[17] Namdar Homayounfar, Wei-Chiu Ma, Shrinidhi Kowshika Lakshmikanth, and Raquel Urtasun. Hierarchical recurrent attention networks for structured online maps. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3417–3426. IEEE Computer Society, 2018. 2, 5, 6

[18] Namdar Homayounfar, Wei-Chiu Ma, Shrinidhi Kowshika Lakshmikanth, and Raquel Urtasun. Hierarchical recurrent attention networks for structured online maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3417–3426, 2018. 2

[19] 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, 2

[20] Joey Hong, Benjamin Sapp, and James Philbin. Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8454–8462, 2019. 1

[21] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1013–1021, 2019. 2

[22] Maximilian Jaritz. *2D-3D scene understanding for autonomous driving*. PhD thesis, PSL Research University, 2020. 1

[23] YeongMin Ko, Jiwon Jun, Donghwuy Ko, and Moongu Jeon. Key points estimation and point instance segmentation approach for lane detection. *CoRR*, abs/2002.06604, 2020. 6

[24] Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Shenlong Wang, and Raquel Urtasun. Convolutional recurrent network for road boundary extraction. *CoRR*, abs/2012.12160, 2020. 5

[25] Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Shenlong Wang, and Raquel Urtasun. Convolutional recurrent network for road boundary extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9512–9521, 2019. 1, 2

[26] Justin Liang and Raquel Urtasun. End-to-end deep structured models for drawing crosswalks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 396–412, 2018. 1, 2

[27] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327, 2020. 6

[28] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 9628–9639, 2018. 6

[29] Chenyang Lu, Marinus Jacobus Gerardus van de Molengraft, and Gijs Dubbelman. Monocular semantic occupancy grid mapping with convolutional variational encoder-decoder networks. *IEEE Robotics Autom. Lett.*, 4(2):445–452, 2019. 2, 6

[30] Wei-Chiu Ma, Ignacio Tartavull, Ioan Andrei Bârsan, Shenlong Wang, Min Bai, Gellert Mattyus, Namdar Homayounfar, Shrinidhi Kowshika Lakshmikanth, Andrei Pokrovsky, and Raquel Urtasun. Exploiting sparse semantic hd maps for self-driving vehicle localization. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5304–5311. IEEE, 2019. 1

[31] Kaustubh Mani, Swapnil Daga, Shubhika Garg, N. Sai Shankar, Krishna Murthy Jatavallabhula, and K. Madhava Krishna. Monolayout: Amodal scene layout from a single image. *CoRR*, abs/2002.08394, 2020. 2

[32] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics Autom. Lett.*, 5(3):4867–4873, 2020. 2, 6

[33] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 5(3):4867–4873, 2020. 2

[34] David Paz, Hengyuan Zhang, and Henrik I Christensen. Tridentnet: A conditional generative model for dynamic trajectory generation. *arXiv preprint arXiv:2101.06374*, 2021. 1

[35] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[36] B Ravi Kiran, Luis Roldao, Benat Irastorza, Renzo Verastegui, Sebastian Suss, Senthil Yogamani, Victor Talpaert, Alexandre Lepoutre, and Guillaume Trehard. Real-time dynamic object detection for autonomous driving using prior 3d-maps. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[37] Edoardo Mello Rella, Jan-Nico Zaech, Alexander Liniger, and Luc Van Gool. Decoder fusion rnn: Context and interaction aware decoders for trajectory prediction. *arXiv preprint arXiv:2108.05814*, 2021. 1

[38] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 3

[39] John Alan Richards and JA Richards. *Remote sensing digital image analysis*, volume 3. Springer, 1999. 2

[40] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11135–11144. IEEE, 2020. 2, 6, 7

[41] Heiko G Seif and Xiaolong Hu. Autonomous driving in the icity—hd maps as a key challenge of the automotive industry. *Engineering*, 2(2):159–162, 2016. 1

[42] Tao Sun, Zonglin Di, Pengyu Che, Chun Liu, and Yin Wang. Leveraging crowdsourced gps data for road extraction from aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7509–7518, 2019. 2

[43] Carles Ventura, Jordi Pont-Tuset, Sergi Caelles, Kevis-Kokitsi Maninis, and Luc Van Gool. Iterative deep learning for road topology extraction. *arXiv preprint arXiv:1808.09814*, 2018. 2

[44] Bin Yang, Ming Liang, and Raquel Urtasun. Hdnet: Exploiting hd maps for 3d object detection. In *Conference on Robot Learning*, pages 146–155. PMLR, 2018. 1

[45] Jan-Nico Zaech, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Action sequence predictions of vehicles in urban environments using map and social context. In *International Conference on Intelligent Robots and Systems (IROS)*, 2020. 1