SparseLaneSTP: Leveraging Spatio-Temporal Priors with Sparse Transformers for 3D Lane Detection

Maximilian Pittner[1,2], Joel Janai[1], Mario Faigle[*1,3], Alexandru Paul Condurache[1,2]

[1]Bosch Mobility Solutions, Robert Bosch GmbH
[2]Institute of Neuro- and Bioinformatics, University of Lübeck
[3]Institute for Signal Processing and System Theory, University of Stuttgart

{Maximilian.Pittner, Joel.Janai, Mario.Faigle, AlexandruPaul.Condurache}@de.bosch.com

Abstract

3D lane detection has emerged as a critical challenge in autonomous driving, encompassing identification and localization of lane markings and the 3D road surface. Conventional 3D methods detect lanes from dense birds-eye-viewed (BEV) features, though erroneous transformations often result in a poor feature representation misaligned with the true 3D road surface. While recent sparse lane detectors have surpassed dense BEV approaches, they completely disregard valuable lane-specific priors. Furthermore, existing methods fail to utilize historic lane observations, which yield the potential to resolve ambiguities in situations of poor visibility. To address these challenges, we present SparseLaneSTP, a novel method that integrates both geometric properties of the lane structure and temporal information into a sparse lane transformer. It introduces a new lane-specific spatio-temporal attention mechanism, a continuous lane representation tailored for sparse architectures as well as temporal regularization.

Identifying weaknesses of existing 3D lane datasets, we also introduce a precise and consistent 3D lane dataset using a simple yet effective auto-labeling strategy. Our experimental section proves the benefits of our contributions and demonstrates state-of-the-art performance across all detection and error metrics on existing 3D lane detection benchmarks as well as on our novel dataset.

1. Introduction

Accurate and robust 3D lane detection forms a pivotal task in autonomous driving addressing safe and reliable identification and localization of the road surface and lane markings. While significant progress has been made in 2D lane detection [10, 13, 17–19, 21, 35, 36, 40, 41, 46, 48, 55], the prediction output is only provided in image space lacking

---

*Work done in context of master's thesis at Bosch.

depth information, which is crucial for autonomous navigation in the 3D world. In contrast, 3D lane detection jointly estimates lane markings and the road surface, directly producing 3D lanes in a vehicle-centered coordinate system. Conventional methods operate on dense feature maps that are transformed from the front-viewed (FV) to the birds-eye-viewed (BEV) perspective using inverse perspective mapping (IPM) [9, 11, 38] or learned mappings [6, 39, 49]. Although this step yields a suitable intermediate representation to capture the road, potential errors in the transformation cause misalignment between the resulting BEV and the true road surface that can hardly be compensated in the subsequent lane estimation step.

As opposed to this two-stage approach, sparse detection methods have gained popularity in object detection [34] and have recently been adapted to 3D lane detection [2, 31]. These methods model 3D lane points as queries and associate them with uncorrupted FV image features - avoiding the necessity of error-prone BEV representations. Queries composed of learned context embedding vectors are integrated into transformer architectures [3, 54] leveraging their ability to learn global context. However, current approaches completely ignore valuable well-known priors about lane and road geometry. Spatial priors have already been utilized in [39]. Based on lane continuity and other properties like line parallelism, a continuous lane representation was introduced together with an efficient regularization. Despite notable advances brought by this approach, the architecture is based on dense BEV representations and the proposed priors have not been successfully customized for sparse architectures. Another class of prior knowledge is grounded in the history of observations. Among various temporal fusion paradigms, object-centric query propagation [50] has recently been proposed for 3D object detection in sparse architectures. In lane detection, temporal information has the potential to resolve ambiguities, particularly in scenarios with limited visibility or occlusions. Though, it has not

been applied effectively so far.

We therefore present a new method SparseLaneSTP that combines spatial and temporal priors with a sparse 3D lane transformer architecture. Our model features a transformer-tailored continuous lane representation providing smooth curves directly. This, combined with robust regularization using additional spatial and temporal objectives, enables the model to learn regression and visibility estimation more accurately. Additionally, spatial and temporal knowledge is integrated into the attention mechanism in a novel way. Instead of standard global self-attention, we design the layers to focus on relevant relations based on lane-structure. Finally, lane queries are temporally propagated and incorporated into the attention to leverage valuable historic keys.

On top of that, we provide a novel 3D lane dataset using a simple yet efficient auto-labeling pipeline. Existing datasets often yield inaccurate labels containing noise and outliers, particularly in the far-range, since they rely on Li-DAR to recover 3D information. Our auto-labeling instead uses temporal aggregation along video-sequences resulting in consistent and accurate 3D lane labels up to 250 meters, making a valuable contribution to the research community. Our contributions can be summarized as follows:

- We present a novel 3D lane detector that integrates spatial and temporal knowledge into a sparse transformer.
- We propose a new attention mechanism that focuses on learning lane-structure and successfully leverages temporal keys from past lane observations.
- We formulate a continuous lane representation tailored for sparse architectures and a temporal regularization.
- Our method achieves state-of-the-art performance on all 3D lane benchmarks.
- We provide a new dataset for 3D lane detection providing accurate and consistent long-range labels.

2. Related work

Dense BEV vs. sparse query-based detection. A common strategy in 3D lane detection involves operating on an intermediate dense BEV representation generated from the FV. Early methods, such as 3D-LaneNet [9], GenLaneNet [11] and 3D-SpLineNet [38], applied IPM to project FV features onto a flat ground plane. However, the assumption of a perfectly flat road is often violated, leading to degraded BEV features. Alternatively, methods from related fields like 3D object detection [14, 15, 23] and BEV segmentation [37] proposed to learn the BEV representation. These concepts were applied to 3D lane detection with PersFormer [6] leveraging deformable cross-attention (DCA) [54], whereas BEV-LaneDet [49] relies on simple multi-layer perceptrons (MLPs). LaneCPP [39] incorporates surface priors to learn 3D features using depth classification inspired by Lift-Splat-Shoot (LSS) [37]. Despite these advancements, all methods remain constrained by an intermediate error-prone BEV representation.

In contrast, sparse methods based on DETR [3, 54] avoid BEV representation by modeling 3D objects as queries that exchange information via attention mechanisms. In 3D object detection, 3D position-aware object queries were introduced that are directly associated with the perspective view through DCA [26, 34]. Similarly, in 3D lane detection, CurveFormer [2] adapts the sparse query design to lanes using polynomial-based curve queries, though these lack the flexibility to capture diverse lane structures in real-world scenarios. LATR [31] refines this approach by representing lanes as individual points but fails to incorporate crucial priors in both the attention mechanism and lane representation.

Priors in lane detection. An indisputable property about lanes lies in their smoothness and continuity. Several methods have proposed continuous lane representations in both 2D [8, 29, 47] and 3D [2, 28, 38, 39] lane detection. In contrast to discrete approaches [6, 9, 11, 16, 49], continuous methods directly provide smooth curves, requiring almost no post-processing and exploit the entire available dense ground truth during training [38, 39]. While CLGO [28] and CurveFormer [2] use simple polynomials, 3D-SpLineNet [38] and LaneCPP [39] use a B-Spline [7] representation, which offers local control over curve segments enabling it to model complex shapes. Additionally, lane structures exhibit inherent properties, such as parallelism between lanes. E.g. SGNet [44] introduces a penalty term to enforce a fixed lane width, but assumes a flat ground plane. GP [20] employs a parallelism loss to maintain a constant local distance between neighboring points, though its effectiveness is influenced by the number of anchor points used. In contrast, LaneCPP [39] demonstrates a more elegant way to encourage parallelism as well as other spatial properties like surface smoothness and curvature restriction using regularization losses. Motivated by LaneCPP, we propose a more suitable Catmull-Rom spline [4] representation tailored for the transformer's attention with sparse queries. We further present more sophisticated loss formulations and integrate prior knowledge into the transformer attention mechanism enhancing the focus on relevant relations. While a concurrent work also uses priors in attention [5], our approach focuses on explicit modeling of structural lane relations and even effectively incorporates temporal knowledge in lane detection for the first time.

Temporal modeling. Another source of prior knowledge is grounded in the history of observations. For video-based 3D object detection various approaches exist to model temporal interactions. BEV-based methods [23] transform past BEV features according to the ego-motion and fuse them with the current representation. More advanced perspective view (PV) methods [26, 30, 32] instead follow the sparse query design but require storing dense historic PV feature maps to interact with object queries. Recently,
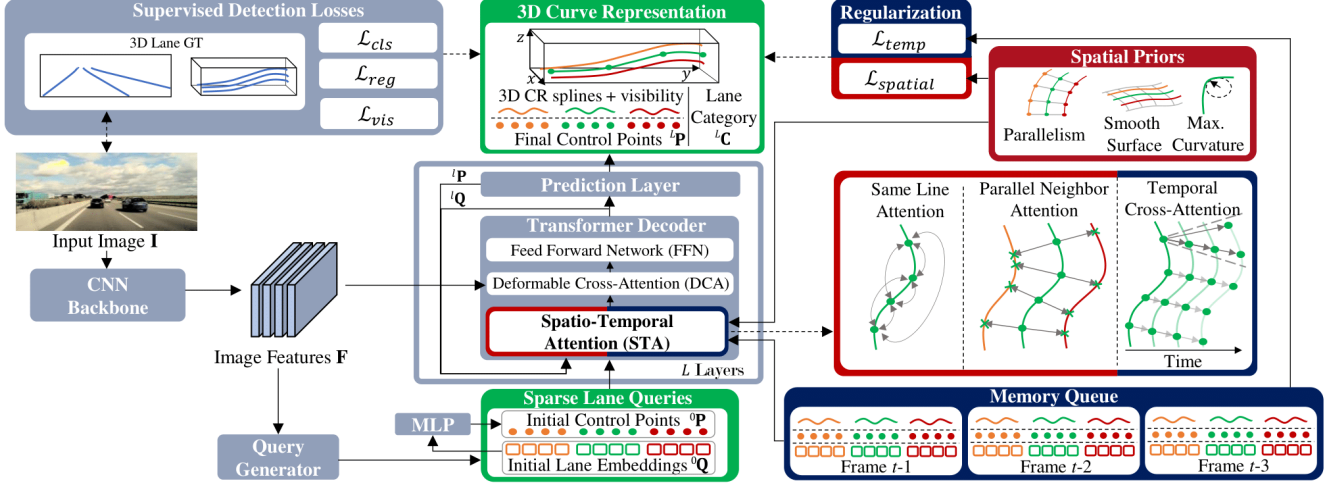
Figure 1. Overview of our method SparseLaneSTP. Sparse lane queries are processed by a transformer integrating temporal knowledge and spatial lane structure priors in a novel spatio-temporal attention. Based on these priors, we formulate spatial and temporally consistent regularization. Finally, our network predicts control points defining our new continuous 3D lane representation.

object-centric temporal modeling [27, 50] has been proposed, which only stores sparse historic object queries and propagates them according to ego- and object-motion in 3D space. In particular, Sparse4Dv2 [27] and StreamPETR [50] leverage the sparse query design and model object queries as hidden states of a transformer, where the relation between current and historic queries is learned via attention.

In 3D lane detection, only few approaches exist that use temporal modeling. STLane3D [51] follows a BEV-based method, PETRv2 [30] fuses historic PV features together with propagated 3D positional embeddings. On the other hand, Anchor3DLane-T [16] leverages temporal information by associating 3D anchors from the current frame with features from previous frames. In contrast, our sparse query-based architecture is more suitable for the effective memory-efficient object-centric paradigm. Consequently, we propose a temporal attention mechanism inspired by [27, 50], which exploits lane-specific properties by reinforcing most relevant query interactions. The static nature of lane markings also allows us to formulate regularization to encourage temporally consistent perception behavior.

3. Methodology

In the following, we describe our 3D lane detection method illustrated in Fig. 1. Given an image $\mathbf{I}$, the model predicts a set of N lanes as continuous curves in 3D space, parameterized by our novel tailored spline representation with M control points $\mathbf{P}$ (see Section 3.1) and lane categories $\mathbf{C}$. The core component of our architecture is the transformer decoder, where each control point, inspired by sparse query design, is paired with a learned context embedding and is iteratively processed through L decoder layers.
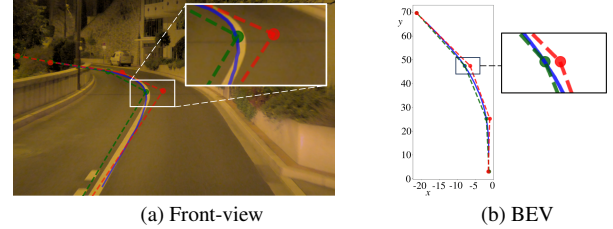

(a) Front-view (b) BEV

Figure 2. CR spline and B-Spline control points in comparison, both fit to the ground truth. B-Spline control points do not align with the curve and are therefore not suiting the sparse query design, whereas CR control points exactly match curve geometry.

In contrast to previous work [2, 24, 31] that simply adapt the DETR [3, 54] framework, our transformer architecture introduces a novel spatio-temporal attention layer to capture spatial intra- and inter-lane relations, while a temporal cross-attention mechanism integrates past observations via propagated lane queries from a memory queue. Finally, besides standard detection losses, our model is cautiously regularized by spatial and temporal objectives to enhance robust and consistent detection behavior.

3.1. 3D lane representation

For the 3D lane representation, we take inspiration from prior work [38, 39] that demonstrated significant benefits of continuous curves. In these frameworks, the model output serves as control points of B-Spline curves that do not necessarily align with the curve geometry. However, using the sparse query design, the 3D position of control points serve as the model's internal state and should therefore correspond to the exact 3D position of lane points. Thus, spline

control points should directly lie on the curve, which is not the case for B-Splines (see Fig. 2). Therefore, we choose Catmull-Rom (CR) splines [4], a class of piece-wise defined smooth third-order polynomial splines, for which the curve inherently passes through its control points.

We adapt this spline-model to define a curve representation $\mathbf{f}_i(s) \in \mathbb{R}^4$ representing the entire i^{th} line proposal parameterized by M control points with 3 spatial (x_{ij}, y_{ij}, z_{ij}) and 1 visibility (v_{ij}) dimension as

$$\mathbf{f}_i(s) = \begin{bmatrix} s^3 & s^2 & s & 1 \end{bmatrix} \cdot \mathbf{M}_{\text{CR}} \cdot \mathbf{P}_i \quad \text{with} \qquad (1)$$

$$\mathbf{P}_i = [\mathbf{P}_{\text{3D},i} | \mathbf{P}_{v,i}] = \begin{bmatrix} x_{i1} & y_{i1} & z_{i1} & v_{i1} \\ & & \cdots & \\ x_{ij} & y_{ij} & z_{ij} & v_{ij} \\ & & \cdots & \\ x_{iM} & y_{iM} & z_{iM} & v_{iM} \end{bmatrix}, \quad (2)$$

with curve argument $s \in [0, 1]$. $\mathbf{P}_i \in \mathbb{R}^{M \times 4}$ denotes the control point matrix composed of $\mathbf{P}_{\text{3D},i} \in \mathbb{R}^{M \times 3}$ defining the curve 3D shape and visibility $\mathbf{P}_{v,i} \in \mathbb{R}^{M \times 1}$. $[\cdot | \cdot]$ is the concatenation operator and $\mathbf{M}_{\text{CR}}$ the CR coefficient matrix. Note that we can eventually express the curve computation as a simple matrix multiplication between the pre-computed spline arguments and the entire predicted control points $\mathbf{P}$.

3.2. Overall architecture

Backbone and query initialization. The RGB input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ is first processed by a CNN backbone (e.g. ResNet [12]) to extract image features of $\mathbf{F} \in \mathbb{R}^{H_{\mathbf{F}} \times W_{\mathbf{F}} \times C}$, where C is the channel size. Similar to [31], we then feed $\mathbf{F}$ through a lane instance segmentation branch to obtain initial query embedding vectors $^0\mathbf{Q} \in \mathbb{R}^{N \times M \times C}$, where N denotes the number of line proposals and M the number of spline control points. An MLP then predicts initial control points $^0\mathbf{P} \in \mathbb{R}^{N \times M \times 4}$ from $^0\mathbf{Q}$.

Transformer decoder. In each transformer decoder layer l, queries are processed by two attention layers

$$^l\mathbf{Q}_{\text{STA}} = \text{STA}\left({}^{l-1}\mathbf{Q}, {}^{l-1}\mathbf{P}, \mathbf{Q}_{\text{Mem}}, \mathbf{P}_{\text{Mem}} \right) \quad (3)$$

$$^l\mathbf{Q}_{\text{DCA}} = \text{DCA}\left({}^l\mathbf{Q}_{\text{STA}}, \mathbf{F}, {}^{l-1}\mathbf{P}_{\text{2D}} \right), \quad (4)$$

$\text{STA}(\cdot)$ denotes our proposed spatio-temporal attention, which is explained in Section 3.3, and $\text{DCA}(\cdot)$ the standard deformable cross-attention presented in [54]. $\mathbf{P}_{\text{2D}}$ are the 2D spline control points in image coordinates resulting from re-projecting $\mathbf{P}_{\text{3D}}$. Finally, the query embeddings are fed through a feed-forward network $^l\mathbf{Q} = \text{FFN}({}^l\mathbf{Q}_{\text{DCA}})$ composed of an MLP and layer normalization.

Prediction layer. After each transformer attention layer, the emerging query embeddings $^l\mathbf{Q}$ are fed through a prediction layer that outputs the spline control points $^l\mathbf{P}$ as well as classification probabilities $\mathbf{C} \in \mathbb{R}^{N \times (K+1)}$ for K lane categories and one background class per line. Particularly,
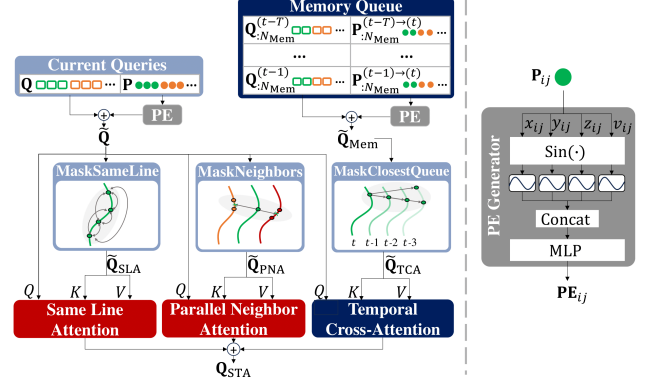

Figure 3. Overview of our spatio-temporal attention module.

for each layer l and query $^l\mathbf{Q}_{ij}$ three MLPs with shared weights among all layers are used to predict the x-, z- and v-component of the control point matrix

$$^l\mathbf{P}_{x,ij} = \sigma(\text{MLP}_x(^l\mathbf{Q}_{ij})) \cdot (x_e - x_s) + x_s \quad (5)$$

$$^l\mathbf{P}_{z,ij} = \sigma(\text{MLP}_z(^l\mathbf{Q}_{ij})) \cdot (z_e - z_s) + z_s \quad (6)$$

$$^l\mathbf{P}_{v,ij} = \sigma(\text{MLP}_v(^l\mathbf{Q}_{ij})), \quad (7)$$

where the sigmoid $\sigma(\cdot)$ normalizes the MLP output to $[0, 1]$. The start-/end- of x- and z-range $x_{s/e}$, $z_{s/e}$ scales up the normalized output to the desired range. The longitudinal y-component $^l\mathbf{P}_{y,ij}$ is pre-defined and uniformly distributed between y_s and y_e to avoid over-parameterization and potential overfitting as argued in [39]. For classification, all queries belonging to the same line proposal are averaged and fed into an MLP followed by a softmax layer.

Finally, the formulation of our supervised losses is inspired by [39], with L1 loss for regression of x and z, binary cross-entropy for visibility v and focal loss [25] for classification.

3.3. Integrating spatial and temporal knowledge

Memory queue. To facilitate attention to past observations, queries are stored within a memory queue. Following a similar approach to [50], this queue is recursively updated after each prediction step and retains queries from the past T frames. Therefore, embedding vectors and their associated control points are recursively pushed into the queue using a first-in, first-out (FIFO) strategy, which yields

$$\mathbf{Q}_{\text{Mem}} = \left[\mathbf{Q}_{:N_{\text{Mem}}}^{(t-1)} \,\middle|\, \mathbf{Q}_{:N_{\text{Mem}}}^{(t-2)} \,\middle|\, \ldots \,\middle|\, \mathbf{Q}_{:N_{\text{Mem}}}^{(t-T)} \right] \quad \text{and} \quad (8)$$

$$\mathbf{P}_{\text{Mem}} = \left[\mathbf{P}_{:N_{\text{Mem}}}^{(t-1)} \,\middle|\, \mathbf{P}_{:N_{\text{Mem}}}^{(t-2)} \,\middle|\, \ldots \,\middle|\, \mathbf{P}_{:N_{\text{Mem}}}^{(t-T)} \right], \quad (9)$$

where $\mathbf{Q}_{:N_{\text{Mem}}}^{(t)} \in \mathbb{R}^{(N_{\text{Mem}} \times M \times C)}$ denotes the embedding vectors from the N_{Mem} most confident predictions of frame t and $\mathbf{Q}_{\text{Mem}} \in \mathbb{R}^{((T \cdot N_{\text{Mem}}) \times M \times C)}$ the full memory queue.

Temporal propagation. Since the model is supposed to learn spatio-temporal context, it must incorporate the positional information of 3D control points from past frames.

However, due to the vehicle's ego-motion, lane positions shift relative to the vehicle. Therefore, the 3D coordinates of past lane control points stored in the memory queue must be transformed according to the ego-pose $\mathbf{E}^{(t)}$ of frame t, which is represented as a homogeneous transformation matrix. Consequently, a control point $\mathbf{P}_{ij}^{(t-k)\to(t)}$ propagated from frame $t-k$ to present frame t is given as

$$\mathbf{P}_{ij}^{(t-k)\to(t)} = \left[\, \mathbf{E}_{inv}^{(t)} \cdot \mathbf{E}^{(t-k)} \cdot \mathbf{P}_{3D,ij}^{(t-k)} \,\middle|\, \mathbf{P}_{v,ij}^{(t-k)} \,\right], \quad (10)$$

where $\mathbf{E}_{inv}$ denotes the inverted matrix $\mathbf{E}$. Obviously, only the geometry component is affected by the propagation, whereas the visibility component remains unchanged. To incorporate spatio-temporal information from propagated control points into queries, we apply a positional encoding (PE) based on 3D geometry and visibility (see Fig. 3). This process yields position-informed embedding vectors

$$\tilde{\mathbf{Q}}_{ij}^{(t-k)\to(t)} = \mathbf{Q}_{ij}^{(t-k)\to(t)} + \mathrm{PE}\big(\mathbf{P}_{ij}^{(t-k)\to(t)}\big), \quad (11)$$

which are processed by the spatio-temporal attention layer. **Spatio-temporal attention.** Most methods [2, 31] rely solely on global self-attention, which contains a majority of redundant lane query interactions that can distract the model from learning relevant context. To address this, we leverage spatial and temporal priors, distinguishing three key relation types: intra-lane (between adjacent points on the same lane), inter-lane (between parallel lanes), and historic (between current and past queries in the memory queue).

We therefore introduce same line attention (SLA), parallel neighbor attention (PNA) and temporal cross-attention (TCA). Applying specific masking, SLA restricts interactions to queries $\tilde{\mathbf{Q}}_{\mathrm{SLA}}$ within the same line, PNA facilitates interactions with queries $\tilde{\mathbf{Q}}_{\mathrm{PNA}}$ from neighboring lines, and TCA enables interactions with closest past queries $\tilde{\mathbf{Q}}_{\mathrm{TCA}}$ in the memory queue (see Fig. 3). Such restriction of query interactions reduces redundancy, encourages the model to focus on critical lane-specific spatial context and integrates valuable temporal information from past observations.
Regularization. Based on spatial and temporal priors we can formulate regularization techniques that promote robust and consistent detection behavior. For the spatial regularization $\mathcal{L}_{spatial}$ we adopt the scheme proposed in [39] encouraging lane parallelism, surface smoothness and suppressing excessive curvature. Besides, temporal consistency can be achieved by optimizing a smoothness loss

$$\mathcal{L}_{temp} = \frac{1}{N} \sum_{i}^{N} \int_{s} \bar{\mathbf{f}}_{v,i}^{(t)}(s) \cdot \big|\big| \mathbf{f}_{3D,i}(s) - \bar{\mathbf{f}}_{3D,i}^{(t)}(s)\big|\big|_{1}, \quad (12)$$

where $\bar{\mathbf{f}}_{3D,i}^{(t)}(s)$ or $\bar{\mathbf{f}}_{v,i}^{(t)}(s)$ is an exponentially moving average of the past predictions

$$\bar{\mathbf{f}}_i^{(t)} = \alpha \cdot \bar{\mathbf{f}}_i^{(t)} + (1-\alpha) \cdot \bar{\mathbf{f}}_i^{(t-1)} \quad (13)$$

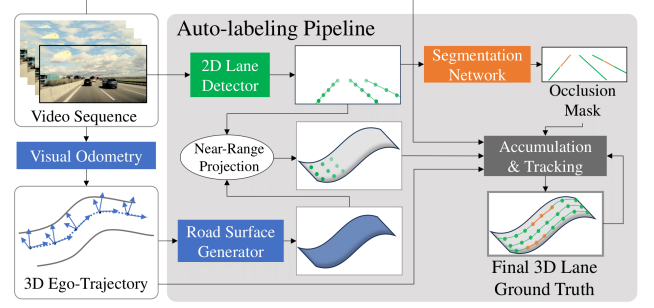with smoothing factor $\alpha \in [0, 1]$.

Figure 4. Our auto-labeling strategy.

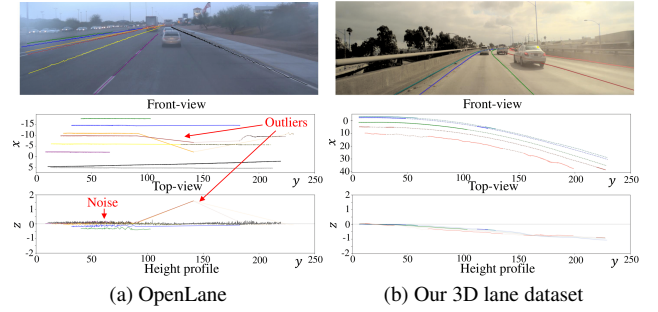(a) OpenLane   (b) Our 3D lane dataset

Figure 5. Examples from OpenLane compared to ours.

4. Auto-labeling and our 3D lane dataset

In this section, we describe our auto-labeling pipeline and provide details about our new 3D lane dataset.

4.1. Auto-labeling strategy

While 3D lane datasets like OpenLane [6] and ONCE-3DLanes [52] have contributed significantly to the field, they possess specific shortcomings that our auto-labeling strategy addresses. As illustrated in Fig. 5a the LiDAR-based annotation approach, which recovers 3D information of 2D lane labels, introduces noise and outliers, particularly in the far range due to sparse scan points, while also posing calibration and synchronization challenges. To overcome these limitations, we propose an automated labeling approach (see Fig. 4) that leverages confident near-range lane point estimates from a 2D detector, which are temporally accumulated in 3D space along an accurate ego-trajectory. More precisely, common visual odometry methods [1, 42, 43] are applied to recover the vehicle ego-trajectory from the input video sequence. Our road surface generator then fits a spline surface to the vehicle's trajectory and orientations. On the other hand, lane pseudo-labels in image space are obtained from a state-of-the-art 2D lane detector [21, 46], providing particularly reliable detection results in in the near-range. Lane and surface information are then combined in 3D by projecting near-range detection results to the road surface and accumulating line points frame
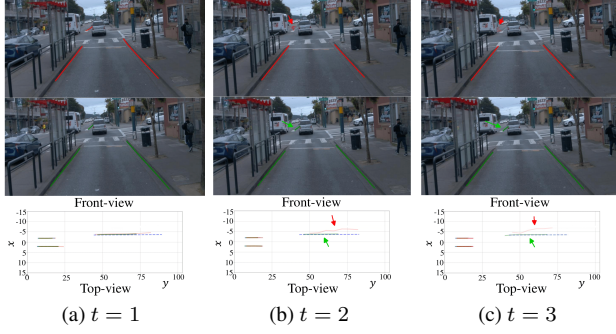
(a) $t = 1$ (b) $t = 2$ (c) $t = 3$

Figure 6. Qualitative comparison of two models, one without temporal and one with temporal for three consecutive frames.

by frame along the ego-trajectory of the sequence. Finally, we apply an accurate semantic segmentation network to automatically assign occlusion labels to each lane point.

4.2. Dataset properties

Using the approach described in 4.1, we are able to automatically generate temporally consistent, high-quality 3D lane labels in a range up to $250\,\mathrm{m}$ (see Fig. 5b). Besides 3D lane geometry, we provide camera intrinsics and extrinsics, lane categories, lane visibility information and consistent track IDs. We further supply relational information about lanes (like line parallelism), local 3D ego-trajectory and global pose information for each frame. Important to mention is also that in contrast to former datasets, our visibility information contains occlusions caused by other traffic participants, enabling models to consistently learn lane structure independent of interruptions and lane occlusions. We refer to the supplementary material for additional information.

5. Experiments

In this section, we first outline our experimental setup, followed by a comprehensive analysis of our approach on two public and our own novel 3D lane dataset.

5.1. Datasets and evaluation metrics

OpenLane [6] based on the Waymo Open dataset [45] remains the most prominent real-world 3D lane dataset to date. It is comprised of 200K images from 1000 sequences recorded in several cities in the USA.
ONCE-3DLanes [52] is another real-world dataset derived from the ONCE dataset [33]. It consists of 211K images collected from multiple cities in China.
Our 3D lane dataset, which was introduced in Section 4, was gathered from multiple regions and countries across the globe. It consists of 511K images with a $90\,\%$ / $10\,\%$ train/test split.
Evaluation metrics. The OpenLane evaluation scheme [6, 11] uses the euclidean distance at uniformly distributed

Continuous Rep.	STA	Regularization	F1(%)↑	Gain(%)
			61.8	(baseline)
✓			62.9	+1.1
✓	✓		65.0	+2.1
✓	✓	✓	65.3	+0.3

Table 1. Performance gain for different contributions on Open-Lane using our novel **Continuous Lane Representation**, **Spatio-Temporal Attention (STA)** and **Regularization**.

Attention	Global	SLA	PNA	SLA + PNA	SLA + PNA + TCA
F1-Score	62.9	62.8	63.5	63.8	**65.0**

Table 2. Analysis of spatio-temporal attention. Same line attention (SLA), parallel neighbor attention (PNA) and temporal cross-attention (TCA) and combinations compared to global attention.

Num. Frames (T)	0	1	2	**3**	4	5
F1(%)↑	63.8	64.1	64.6	**65.0**	64.5	63.9

Table 3. Analysis of number of frames in the memory for STA.

points and a range-IoU to compute the F1-Score. Moreover, x- and z-errors in near- (0-$40\,\mathrm{m}$) and far-range (40-$100\,\mathrm{m}$) evaluate the geometric accuracy. ONCE-3DLanes instead uses the uni-lateral chamfer distance (CD) and computes Precision (P), Recall (R) and F1-Score based on it.

For our dataset, we extend the well-established Open-Lane evaluation framework [6, 11]. Due to the long-range labels, we add x-/z-error intervals for the ranges 100-$150\,\mathrm{m}$ and 150-$200\,\mathrm{m}$. Since our dataset offers visibility labels for occlusions, we also add a visibility Intersection-over-Union metric (Vis-IoU). This score helps to gain understanding about the model's capability of distinguishing visible lane segments from parts occluded by other traffic participants and provides a measure for start-end range detection.

5.2. Implementation details

We use an image input size 720×960 and ResNet-50 [12] as our standard backbone. The query generator is trained on the auxiliary task of lane instance segmentation similar to [31]. The transformer decoder applied in the large model used for the evaluation in Section 5.4 consists of $L = 6$ layers, whereas the lighter model used for ablation studies in Sec. 5.3 has 2 layers. The lane representation uses $M = 20$ control points on OpenLane and ONCE-3DLanes and 30 control points on our dataset due to the larger range. The number of lane proposals is set to $N = 40$ for OpenLane and ours and 10 for ONCE-3DLanes, which is sufficient for the small number of occurring lanes per frame. Regarding temporal fusion, our best model uses propagated queries from $T = 3$ historic frames.

Method	Backbone	Resolution	F1-Score(%)↑	X-error near(m)↓	X-error far(m)↓	Z-error near(m)↓	Z-error far(m)↓	F1-Score(%) per Scenario ↑					
								U&D	C	EW	N	I	M&S
3D-LaneNet [9]	VGG-16	360 x 480	44.1	0.479	0.572	0.367	0.443	40.8	46.5	47.5	41.5	32.1	41.7
Gen-LaneNet [11]	ERFNet	360 x 480	32.3	0.591	0.684	0.411	0.521	25.4	33.5	28.1	18.7	21.4	31.0
PersFormer [6]	EfficientNet-B7	360 x 480	50.5	0.485	0.553	0.364	0.431	42.4	55.6	48.6	46.6	40.0	50.7
CurveFormer [2]	EfficientNet-B7	360 x 480	50.5	0.340	0.772	0.207	0.651	45.2	56.6	49.7	49.1	42.9	45.4
BEV-LaneDet [49]	ResNet-34	360 x 480	58.4	0.309	0.659	0.244	0.631	48.7	63.1	53.4	53.4	50.3	53.7
Anchor3DLane-T [16]	ResNet-18	360 x 480	54.3	0.275	0.310	0.105	0.135	47.2	58.0	52.7	48.7	45.8	51.7
PETRv2 [30]	VoV-99	360 x 480	61.2	0.400	0.573	0.265	0.413	–	–	–	–	–	–
LATR [31]	ResNet-50	720 x 960	61.9	0.219	0.259	0.075	0.104	55.2	68.2	57.1	55.4	52.3	61.5
LaneCPP [39]	EfficientNet-B7	360 x 480	60.3	0.264	0.310	0.077	0.117	53.6	64.4	56.7	54.9	52.0	58.7
PVALane [53]	Swin-B	720 x 960	63.4	0.226	0.257	0.093	0.119	56.1	67.7	64.0	58.6	53.6	60.8
GroupLane [22]	ConvNext-Base	not specified	64.1	0.320	0.441	0.233	0.402	–	–	–	–	–	–
SparseLaneSTP (Ours)	ResNet-50	720 x 960	66.1	0.203	0.240	0.066	0.092	57.3	73.0	60.1	58.3	58.2	66.5

Table 4. Quantitative comparison on OpenLane [6]. **Best performance** and <u>second best</u> are highlighted.

Method	F1-Score(%)↑	X-error / Z-error (m)↓				Vis-IoU(%)↑
		0 m - 40 m	40 m - 100 m	100 m - 150 m	150 m - 200 m	
PersFormer [6]	59.2	0.191 / 0.072	0.344 / 0.145	0.671 / 0.220	0.845 / 0.379	70.1
LaneCPP [39]	63.1	0.144 / 0.040	0.298 / 0.093	0.494 / 0.188	0.761 / 0.293	77.5
Baseline (LATR) [31]	65.1	0.122 / 0.038	0.263 / 0.098	0.468 / 0.185	0.700 / 0.264	78.3
SparseLaneSTP (ours)	**68.2**	**0.109 / 0.033**	**0.233 / 0.092**	**0.443 / 0.182**	**0.646 / 0.263**	**81.4**

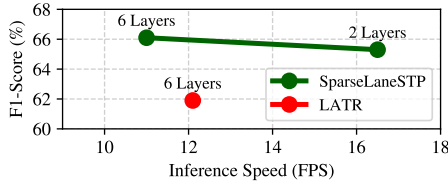Table 5. Quantitative comparison on our 3D lane dataset. **Best performance** and <u>second best</u> are highlighted.

Figure 7. Efficiency analysis of SparseLaneSTP.

Method	F1(%)↑	P(%)↑	R(%)↑	CD(m)↓
3D-LaneNet [9]	44.73	61.46	35.16	0.127
Gen-LaneNet [11]	45.59	63.95	35.42	0.121
SALAD [52]	64.07	75.90	55.42	0.098
PersFormer [6]	74.33	80.30	69.18	0.074
Anchor3DLane [16]	74.87	80.85	69.71	0.060
LATR [31]	80.59	86.12	75.73	0.052
PVALane [53]	76.35	82.81	70.83	0.059
GroupLane [22]	80.73	82.56	78.90	0.053
SparseLaneSTP (ours)	**82.75**	**86.47**	**79.33**	**0.048**

Table 6. Quantitative comparison on ONCE-3DLanes [52].

5.3. Ablation studies

We perform a comprehensive analysis validating the effectiveness of our design choices in SparseLaneSTP. For these experiments, we used a smaller model consisting of two decoder layers evaluated on OpenLane.

Table 1 summarizes the impact of our contributions. Our CR spline based lane representation improves the F1-Score by over 1 percentage point compared to the discrete one (baseline). The most significant gain in F1-Score is achieved by the novel STA component, which encourages focus on relevant relations based on lane structure and, more importantly, on related past observations by integrating temporal cross-attention. Eventually, regularization enhances robustness and consistency, leading to improved generalization, as evidenced by the additional F1-Score gain.

As shown in Table 2, global self-attention apparently fails to capture important relations of lane structures. The combination of our proposed SLA and PNA demonstrates the benefits of enhancing the focus on intra- and inter-lane relations by an F1-Score increase $+0.9\,\%$. Integrating knowledge about past observations via our suggested TCA yields the greatest improvement of additional $+1.2\,\%$. As indicated by Table 3, $T = 3$ frames in the memory produces the best results. Using too few frames provides insufficient temporal context, whereas too many frames apparently introduce redundancy and distract the model from relevant information. We further qualitatively demonstrate the beneficial impact of our primary contribution on temporal fusion in Fig. 6. The non-temporal model fails to retain the detection from frame $t = 1$ in subsequent frames as lane visibility deteriorates. In contrast, the temporally-aware model maintains the detection despite partial occlusions, demonstrating improved robustness. Regularization further preserves temporal consistency mitigating the gradual disappearing and drifting in the non-temporal model.

Finally, Fig. 7 reveals efficiency benefits of our approach. SparseLaneSTP with six decoder layers achieves 11.0 FPS, comparable to LATR at 12.1 FPS. While our proposed temporal integration (memory queue and STA) in-
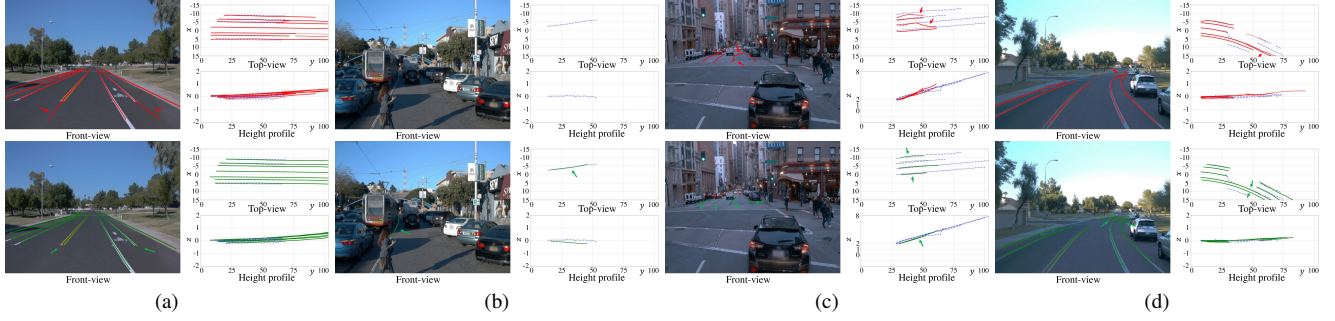
Figure 8. Qualitative comparison of LATR and SparseLaneSTP on OpenLane with ground truth. Main differences are highlighted.

troduces a modest 9% overhead, it also enables a smaller model with only two layers to reach 65.3% F1-Score ($+3.4\%$ over six-layer LATR) at a significantly higher speed of 16.5 FPS.

5.4. Main results

In this section, we compare our method to other 3D lane detectors on public benchmarks as well as on our own dataset. **Results on OpenLane.** Table 4 presents a quantitative comparison demonstrating the superiority of our approach. SparseLaneSTP reaches the highest F1-Score, exceeding 66%, significantly surpassing the second-best model despite its larger backbone and even achieves the lowest geometric errors.

Fig. 8 illustrates the advantages of our approach qualitatively across multiple aspects. We compare our Sparse-LaneSTP to LATR [31], which is most related to our baseline. As shown in Fig. 8a, our continuous lane representation enables precise visibility estimation with accurate detection of start/end of markings, whereas LATR - constrained by fixed discrete anchor points - produces erroneous estimates. In Fig. 8b, our temporal fusion preserves the line despite partial occlusions, whereas non-temporal LATR fails. It also struggles to estimate faint lane markings on the far side of the junction shown in Fig. 8c, while with STA our model captures lane structures more effectively. Eventually, in Fig. 8d our model exhibits a stronger reliance on priors resulting in robust and consistent behavior, most likely due to the incorporated regularization. **Results on ONCE-3DLanes.** As shown in Table 6, our method achieves the highest detection scores and demonstrates accurate regression with a significantly lower CD, despite the inaccurate camera parameters provided by the dataset. Notably, even though ONCE-3DLanes has a lower frame rate, our temporal approach performs effectively. **Results on our 3D lane dataset.** Initial results on our novel dataset are presented in Table 5. Our proposed extension of the evaluation metrics provides a more comprehensive assessment of overall performance, particularly in the far-range and for occlusion detection. To facilitate a fair
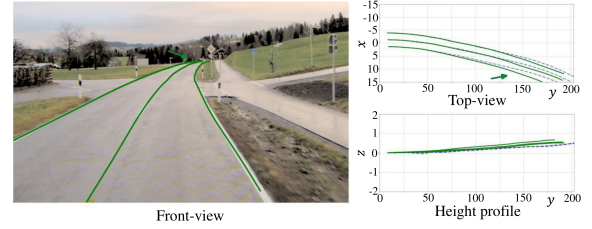

Figure 9. Qualitative example of SparseLaneSTP on our data.

comparison, we train the baseline and other methods with specific modifications to range and capacity. Our method excels in F1-Score, error metrics and notably in the Vis-IoU, highlighting the superior performance of our temporal model in detecting occlusions. Fig. 9 shows a qualitative result on our dataset, showcasing accurate long-range lane estimation ($200\,\mathrm{m}$). For more qualitative results with comparisons to the baseline we refer to the supplementary.

6. Conclusion

In this work, we presented a novel 3D lane detection approach that combines sparse lane transformers with spatial and temporal priors. We proposed a continuous lane representation, tailored for the sparse query design, thereby overcoming limitations of discrete representations. Additionally, we developed a novel attention layer that effectively captures intra- and inter-lane relations while integrating temporal information through propagated lane queries. Furthermore, we introduced a dedicated regularization scheme enhancing temporal consistency. Our experiments validate the effective integration of temporal information and spatial priors, as well as the benefits of our lane representation, leading to state-of-the-art performance across all evaluated 3D lane datasets. Moreover, we presented a new 3D lane dataset, constructed using a simple yet effective auto-labeling pipeline, providing geometrically accurate labels that span long distances and explicitly annotate occlusions, fostering further advancements in the field. In future, we will explore an extension to 3D lane tracking to leverage the full potential of temporal information.

References

[1] Lucas R Agostinho, Nuno M Ricardo, Maria I Pereira, Antoine Hiolle, and Andry M Pinto. A practical survey on visual odometry for autonomous driving in challenging scenarios and conditions. *IEEE Access*, 2022. 5

[2] Yifeng Bai, Zhirong Chen, Zhangjie Fu, Lang Peng, Pengpeng Liang, and Erkang Cheng. Curveformer: 3d lane detection by curve propagation with curve queries and attention. In *Proc. IEEE International Conf. on Robotics and Automation (ICRA)*, 2023. 1, 2, 3, 5, 7

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 1, 2, 3

[4] Edwin Catmull and Raphael Rom. A class of local interpolating splines. *Computer aided geometric design*, 1972. 2, 4

[5] Yifan Chang, Junjie Huang, Xiaofeng Wang, Yun Ye, Zhujin Liang, Yi Shan, Dalong Du, and Xingang Wang. Rethinking lanes and points in complex scenarios for monocular 3d lane detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2

[6] Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, et al. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 1, 2, 5, 6, 7

[7] Carl de Boor. On calculating with b-splines. *Journal of Approximation Theory*, 1972. 2

[8] Zhengyang Feng, Shaohua Guo, Xin Tan, Ke Xu, Min Wang, and Lizhuang Ma. Rethinking efficient lane detection via curve modeling. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[9] Noa Garnett, Rafi Cohen, Tomer Pe'er, Roee Lahav, and Dan Levi. 3d-lanenet: End-to-end 3d multiple lane detection. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 1, 2, 7

[10] Mohsen Ghafoorian, Cedric Nugteren, Nóra Baka, Olaf Booij, and Michael Hofmann. EL-GAN: embedding loss driven generative adversarial networks for lane detection. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018. 1

[11] Yuliang Guo, Guang Chen, Peitao Zhao, Weide Zhang, Jinghao Miao, Jingao Wang, and Tae Eun Choe. Gen-lanenet: A generalized and scalable approach for 3d lane detection. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 1, 2, 6, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 6

[13] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 1

[14] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *CoRR*, abs/2203.17054, 2022. 2

[15] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *CoRR*, abs/2112.11790, 2021. 2

[16] Shaofei Huang, Zhenwei Shen, Zehao Huang, Zi han Ding, Jiao Dai, Jizhong Han, Naiyan Wang, and Si Liu. Anchor3dlane: Learning to regress 3d anchors for monocular 3d lane detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 7

[17] Brody Huval, Tao Wang, Sameep Tandon, Jeff Kiske, Will Song, Joel Pazhayampallil, Mykhaylo Andriluka, Pranav Rajpurkar, Toki Migimatsu, Royce Cheng-Yue, Fernando A. Mujica, Adam Coates, and Andrew Y. Ng. An empirical evaluation of deep learning on highway driving. *arXiv/1504.01716*, 2015. 1

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

[19] 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. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2017. 1

[20] Chenguang Li, Jia Shi, Ya Wang, and Guangliang Cheng. Reconstruct from top view: A 3d lane detection approach based on geometry structure prior. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[21] Xiang Li, Jun Li, Xiaolin Hu, and Jian Yang. Line-cnn: End-to-end traffic line detection with line proposal unit. *IEEE Trans. on Intelligent Transportation Systems (T-ITS)*, 2020. 1, 5

[22] Zhuoling Li, Chunrui Han, Zheng Ge, Jinrong Yang, En Yu, Haoqian Wang, Xiangyu Zhang, and Hengshuang Zhao. Grouplane: End-to-end 3d lane detection with channel-wise grouping. *IEEE Robotics and Automation Letters (RA-L)*, 2024. 7

[23] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2

[24] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized HD map construction. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023. 3

[25] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2017. 4

[26] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *CoRR*, abs/2211.10581, 2022. 2

[27] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d v2: Recurrent temporal fusion with sparse model. *CoRR*, abs/2305.14018, 2023. 3

[28] Ruijin Liu, Dapeng Chen, Tie Liu, Zhiliang Xiong, and Ze-

jian Yuan. Learning to predict 3d lane shape and camera pose from a single image via geometry constraints. In *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2022. 2

[29] Ruijin Liu, Zejian Yuan, Tie Liu, and Zhiliang Xiong. End-to-end lane shape prediction with transformers. In *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2

[30] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 2, 3, 7

[31] Yueru Luo, Chaoda Zheng, Xu Yan, Tang Kun, Chao Zheng, Shuguang Cui, and Zhen Li. Latr: 3d lane detection from monocular images with transformer. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[32] Zhipeng Luo, Changqing Zhou, Gongjie Zhang, and Shijian Lu. Detr4d: Direct multi-view 3d object detection with sparse attention. *arXiv preprint arXiv:2212.07849*, 2022. 2

[33] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Chunjing Xu, and Hang Xu. One million scenes for autonomous driving: ONCE dataset. In *NeurIPS Datasets and Benchmarks*, 2021. 6

[34] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 1, 2

[35] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. In *Proc. IEEE Intelligent Vehicles Symposium (IV)*, 2018. 1

[36] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial CNN for traffic scene understanding. In *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2018. 1

[37] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2

[38] Maximilian Pittner, Alexandru Condurache, and Joel Janai. 3d-splinenet: 3d traffic line detection using parametric spline representations. In *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023. 1, 2, 3

[39] Maximilian Pittner, Joel Janai, and Alexandru P Condurache. Lanecpp: Continuous 3d lane detection using physical priors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 4, 5, 7

[40] Fabio Pizzati, Marco Allodi, Alejandro Barrera, and Fernando García. Lane detection and classification using cascaded cnns. In *Proc. of the International Conf. on Computer Aided Systems Theory (EUROCAST)*, 2019. 1

[41] Zhan Qu, Huan Jin, Yang Zhou, Zhen Yang, and Wei Zhang. Focus on local: Detecting lane marker from bottom up via key point. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[42] Davide Scaramuzza, Friedrich Fraundorfer, and Roland Siegwart. Real-time monocular visual odometry for on-road vehicles with 1-point ransac. In *Proc. IEEE International Conf. on Robotics and Automation (ICRA)*, 2009. 5

[43] Frank Steinbrücker, Jürgen Sturm, and Daniel Cremers. Real-time visual odometry from dense rgb-d images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV) Workshops*, 2011. 5

[44] Jinming Su, Chao Chen, Ke Zhang, Junfeng Luo, Xiaoming Wei, and Xiaolin Wei. Structure guided lane detection. In *Proc. of the International Joint Conf. on Artificial Intelligence (IJCAI)*, 2021. 2

[45] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[46] Lucas Tabelini, Rodrigo Berriel, Thiago M Paixao, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Keep your eyes on the lane: Real-time attention-guided lane detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 5

[47] Lucas Tabelini Torres, Rodrigo Ferreira Berriel, Thiago M. Paixão, Claudine Badue, Alberto F. De Souza, and Thiago Oliveira-Santos. Polylanenet: Lane estimation via deep polynomial regression. In *Proc. of the International Conf. on Pattern Recognition (ICPR)*, 2020. 2

[48] Jinsheng Wang, Yinchao Ma, Shaofei Huang, Tianrui Hui, Fei Wang, Chen Qian, and Tianzhu Zhang. A keypoint-based global association network for lane detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[49] Ruihao Wang, Jian Qin, Kaiying Li, Yaochen Li, Dong Cao, and Jintao Xu. Bev-lanedet: An efficient 3d lane detection based on virtual camera via key-points. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 7

[50] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 1, 3, 4

[51] Yin Wang, Qiuyi Guo, Peiwen Lin, Guangliang Cheng, and Jian Wu. Spatio-temporal fusion-based monocular 3d lane detection. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2022. 3

[52] Fan Yan, Ming Nie, Xinyue Cai, Jianhua Han, Hang Xu, Zhen Yang, Chaoqiang Ye, Yanwei Fu, Michael Bi Mi, and Li Zhang. Once-3dlanes: Building monocular 3d lane detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5, 6, 7

[53] Zewen Zheng, Xuemin Zhang, Yongqiang Mou, Xiang Gao, Chengxin Li, Guoheng Huang, Chi-Man Pun, and Xiaochen Yuan. Pvalane: Prior-guided 3d lane detection with view-agnostic feature alignment. In *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2024. 7

[54] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021. 1, 2, 3, 4

[55] Qin Zou, Hanwen Jiang, Qiyu Dai, Yuanhao Yue, Long Chen, and Qian Wang. Robust lane detection from continuous driving scenes using deep neural networks. *IEEE Trans. on Vehicular Technology (VTC)*, 2020. 1