THU-AM-093

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Anchor3DLane: Learning to Regress 3D Anchors for Monocular 3D Lane Detection

Shaofei Huang[1,2] Zhenwei Shen[3*] Zehao Huang[3] Zi-han Ding[4,5]
Jiao Dai[1,2] Jizhong Han[1,2] Naiyan Wang[3] Si Liu[4,5]

([*]Work done while at TuSimple)

[1]IIE, CAS [2]UCAS [3]TuSimple [4]IAI, BUAA [5]HII, BUAA

Code available at: https://github.com/tusen-ai/Anchor3DLane

中国科学院 信息工程研究所 INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学 University of Chinese Academy of Sciences

北京航空航天大學

tu 图森未来

Preview

Anchor3DLane: Learning to Regress 3D Anchors for Monocular 3D Lane Detection

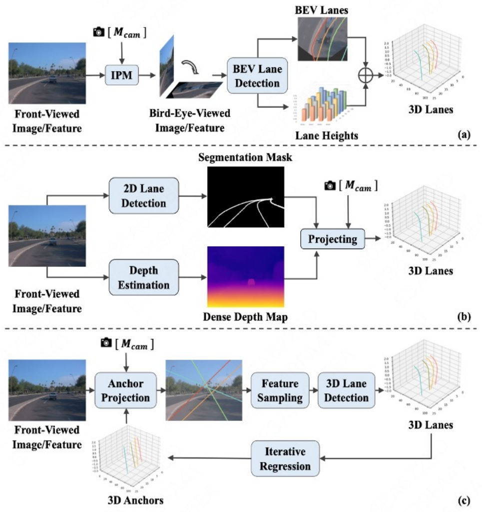
Shaofei Huang[1,2] Zhenwei Shen[3*] Zehao Huang[3] Zi-han Ding[4,5] Jiao Dai[1,2] Jizhong Han[1,2] Naiyan Wang[3] Si Liu[4,5]

[1]IIE, CAS [2]SCS, UCAS [3]TuSimple [4]IAI, BUAA [5]HII, BUAA (*work done in TuSimple)

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

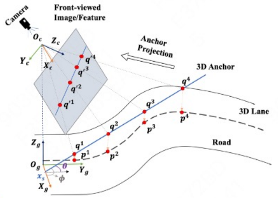
Motivation

- **BEV-based methods (a)** warp FV images/features into BEV space with IPM, which relies on the strict assumption of flat ground. Useful height information and context information are also lost inevitably in BEV representations.
- **Non-BEV method (b)** decomposes 3D lane detection task into 2D lane segmentation and dense depth estimation tasks and lacks structured representations of 3D lanes.
- Our **Anchor3DLane (c)** directly defines anchors in 3D space and regresses 3D lanes directly from FV without introducing BEV. 3D lane anchors are projected to the FV features to extract their features which contain both good structural and context information.



Representations of 3D Lane Anchor

- A 3D lane / anchor is described by 3D points with N uniformly sampled y-coordinates $\{y^k\}, k \in [1, N]$.
- Lane point of the i-th lane: $\mathbf{p}_i^k = (x_i^k, y^k, z_i^k, vis_i^k), k \in [1, N]$;
- Anchor point of the j-th anchor: $\mathbf{q}_j^k = (x_j^k, y^k, z_j^k), k \in [1, N]$.

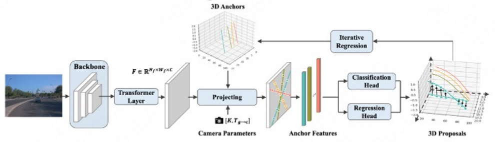


Anchor3DLane

- To obtain features of 3D anchors, we first project them into the plane of FV feature using camera parameters:

$$\begin{bmatrix}\tilde{u}^k \\ \tilde{v}^k \\ d^k\end{bmatrix} = \mathbf{K}\mathbf{T}_{g\to c}\begin{bmatrix}x^k \\ y^k \\ z^k \\ 1\end{bmatrix}, \quad u^k = W_f / W \cdot \frac{\tilde{u}^k}{d^k}, \quad v^k = H_f / H \cdot \frac{\tilde{v}^k}{d^k},$$

- The feature of each anchor is obtained through bilinear interpolation around the projected points on the FV feature.
- A classification head and a regression head are appended after each anchor features to predict its category, x/z offset and visibilities for each anchor point.
- The lane predictions can also serve as new 3D anchors for iterative regression.



Temporal Context Modeling

- Anchor3DLane can be further extended to multi-frame 3D lane detection to incorporate temporal context for larger perception range.
- 3D points in the t-th frame's ground coordinate system can be transformed into the t'-th frame's ground coordinate system by relative transformation matrix to gather anchor features from previous frame:

$$\begin{bmatrix}x_{t'} \\ y_{t'} \\ z_{t'}\end{bmatrix} = \mathbf{T}_{g(t)\to g(t')}\begin{bmatrix}x_t \\ y_t \\ z_t \\ 1\end{bmatrix}$$

- Cross-frame attention is adopted for anchor feature fusion.

Equal-Width Constraint

- we leverage the geometry property of 3D lanes, i.e., lanes in 3D space are nearly parallel with each other, and formulate it as an equal-width constraint to adjust the x-coordinates of lane predictions.
- Given $\{x_j^k\}_{k=1}^N$ and $\{x_{j'}^k\}_{k=1}^N$ as x-coordinates of two lanes and $\tilde{\Delta}\mathbf{x}^k$ as adjustment to x^k, the objective function is formulated as:

$$\min_{\{\tilde{\Delta}\mathbf{x}_j\}_{j\in[1,Q]}} \frac{1}{Q(Q-1)}\sum_{j=1}^{Q}\sum_{j'=1, j'\neq j}^{Q}\mathcal{L}(\mathbf{w}_{j,j'}) + \alpha\frac{1}{Q}\sum_{j=1}^{Q}\|\tilde{\Delta}\mathbf{x}_j\|_2,$$

$$w_{j,j'}^k = |\cos\theta_j^k(x_j^k + \tilde{\Delta}x_j^k - x_{j'}^k - \tilde{\Delta}x_{j'}^k)|,$$

$$\mathcal{L}(\mathbf{w}_{j,j'}) = \sum_{k=1}^{N}|w_{j,j'}^k - \frac{1}{N}\sum_{k'=1}^{N}w_{j,j'}^{k'}|.$$

Experiments

- Extensive ablation studies have shown the effectiveness of each components of our method.

Comparison with BEV feature sampling

Input	Feat	F1(%)	x err/C(m)	x err/F(m)	z err/C(m)	z err/F(m)
BEV	BEV	47.6	0.466	0.421	0.119	0.170
FV	BEV	47.6	0.443	0.446	0.118	0.160
FV	FV	**53.1**	**0.300**	**0.31**	**0.103**	**0.139**

Equal-Width Constraint (EWC)

Method	F1(%)	x err/C(m)	x err/F(m)
w/o EWC	54.8	**0.318**	0.349
w/ EWC	**55.0**	**0.318**	**0.337**

Iterative steps

Iter	F1(%)	x err/C(m)	x err/F(m)	z err/C(m)	z err/F(m)
1	54.8	0.318	0.349	**0.101**	**0.147**
2	56.3	**0.287**	0.335	0.103	0.152
3	**57.0**	**0.287**	**0.327**	0.104	0.148

Training Frames

Frame Range	F1(%)	x err/C(m)	x err/F(m)	z err/C(m)	z err/F(m)
3 frames	55.0	**0.306**	**0.326**	**0.099**	0.148
5 frames	55.2	0.308	0.330	**0.099**	**0.145**
7 frames	**56.1**	0.312	0.335	0.101	0.150

- Quantitative and qualitative results on popular benchmarks show the superiority of our method.

ApolloSim

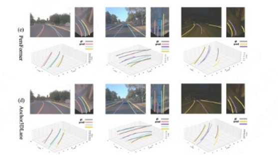
Scene	Method	AP(%)↑	F1(%)↑	x err/C(m)↓	x err/F(m)↓	z err/C(m)↓	z err/F(m)↓
Balanced Scene	3DLaneNet [7]	89.3	86.4	0.068	0.477	0.015	**0.202**
	Gen-LaneNet [8]	90.1	88.1	0.061	0.496	0.012	0.214
	CLGo [20]	94.2	91.9	0.061	0.361	0.029	0.250
	PersFormer [5]	-	92.9	0.054	0.356	0.010	0.234
	GP [16]	93.8	91.9	0.049	0.387	**0.008**	0.213
	Anchor3DLane (Ours)	**97.2**	**95.6**	0.052	0.306	0.015	0.223
	Anchor3DLane†(Ours)	97.1	95.4	**0.045**	**0.300**	0.016	0.223
Rare Subset	3DLaneNet [7]	74.6	72.0	0.166	0.855	0.039	**0.521**
	Gen-LaneNet [8]	79.0	78.0	0.139	0.903	0.030	0.539
	CLGo [20]	88.3	86.1	0.147	0.735	0.071	0.609
	PersFormer [5]	-	87.5	0.107	0.782	0.024	0.602
	GP [16]	85.2	83.7	0.126	0.903	**0.023**	0.625
	Anchor3DLane (Ours)	**96.9**	**94.4**	0.094	**0.693**	0.027	0.579
	Anchor3DLane† (Ours)	95.9	**94.4**	**0.082**	0.699	0.030	0.580
Visual Variations	3D-LaneNet [7]	74.9	72.5	0.115	0.601	0.032	0.230
	Gen-LaneNet [8]	87.2	85.3	0.074	0.538	0.015	0.232
	CLGo [20]	89.2	87.3	0.084	0.464	0.045	0.312
	PersFormer [5]	-	89.6	0.074	0.430	0.015	0.266
	GP [16]	92.1	89.9	0.060	0.446	**0.011**	0.235
	Anchor3DLane (Ours)	**93.6**	91.4	0.068	0.367	0.020	0.232
	Anchor3DLane† (Ours)	92.5	**91.8**	**0.047**	**0.327**	0.019	**0.219**

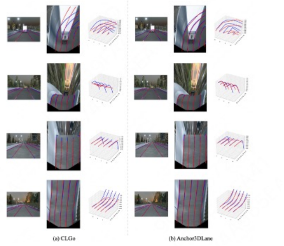
OpenLane

Method	F1(%)↑	Cate Acc(%)↑	x err/C(m)↓	x err/F(m)↓	z err/C(m)↓	z err/F(m)↓
3D-LaneNet [7]	44.1	-	0.479	0.572	0.367	0.443
GenLaneNet [8]	32.3	-	0.591	0.684	0.411	0.521
PersFormer [5]	50.5	**92.3**	0.485	0.553	0.364	0.431
Anchor3DLane (Ours)	53.1	90.0	0.300	0.311	**0.103**	0.139
Anchor3DLane† (Ours)	53.7	90.9	0.276	0.311	0.107	0.138
Anchor3DLane-T† (Ours)	**54.3**	90.7	**0.275**	**0.310**	0.105	**0.135**

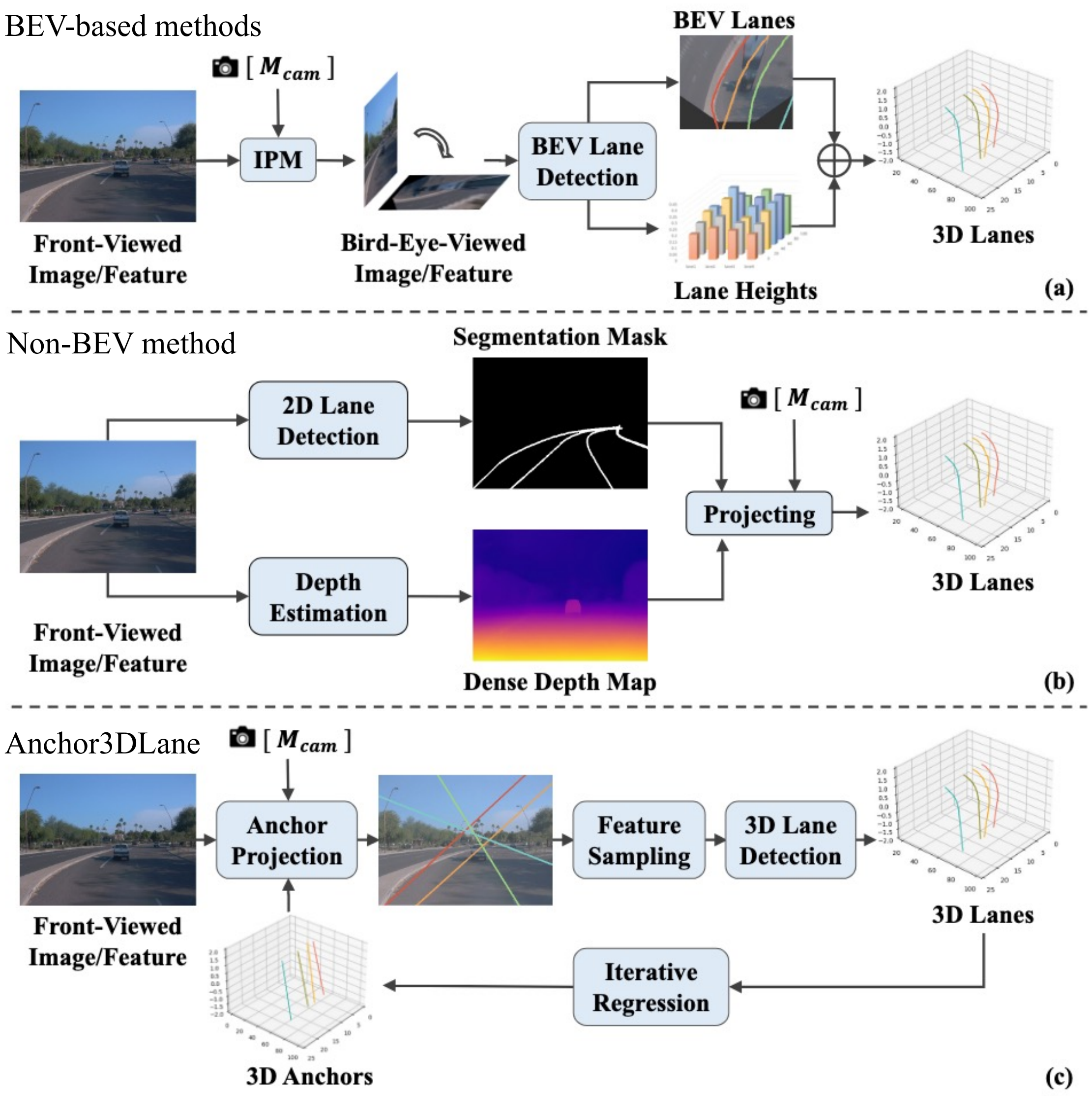
ONCE-3DLane

Method	F1(%)↑	P(%)↑	R(%)↑	CD Error(m)↓
3D-LaneNet [7]	44.73	61.46	35.16	0.127
Gen-LaneNet [8]	45.59	63.95	35.42	0.121
SALAD [36]	64.07	75.90	55.42	0.098
PersFormer [5]	74.33	80.30	69.18	0.074
Anchor3DLane (Ours)	74.44	80.50	69.23	0.064
Anchor3DLane† (Ours)	**74.87**	**80.85**	**69.71**	**0.060**





Motivation



BEV-based methods

- ➢ IPM relies on the assumption of flat ground, which does not always hold true
- ➢ Useful height and context information above the road surface are lost after IPM

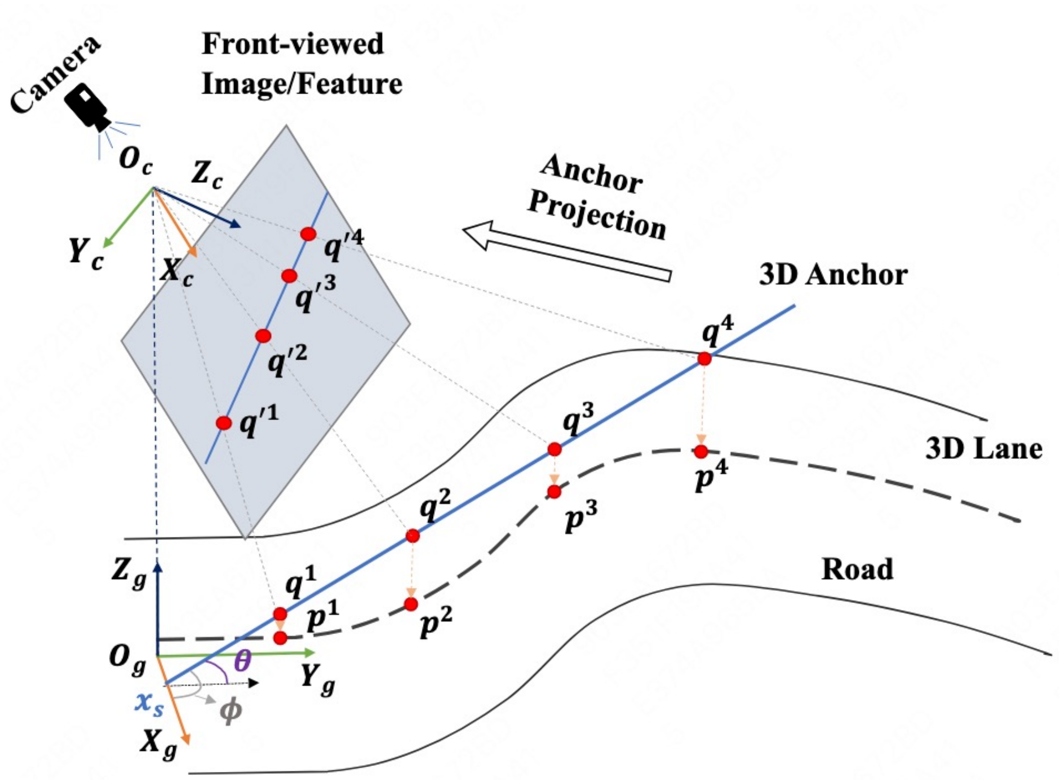
Non-BEV method

- ➢ Lacks structured representations of 3D lanes
- ➢ Performances lag behind BEV-based methods

Our Anchor3DLane

- ➢ Defines 3D lane anchors for structural representation of 3D lanes
- ➢ Retains context information by projecting 3D anchors and sampling anchor features from original FV features
- ➢ Easily extended to iterative regression and multi-frame settings

Representations of 3D Lanes and 3D Anchors



A 3D lane / anchor is described by 3D points with N uniformly sampled y-coordinates $\{y^k\}$, $k \in [1, N]$.

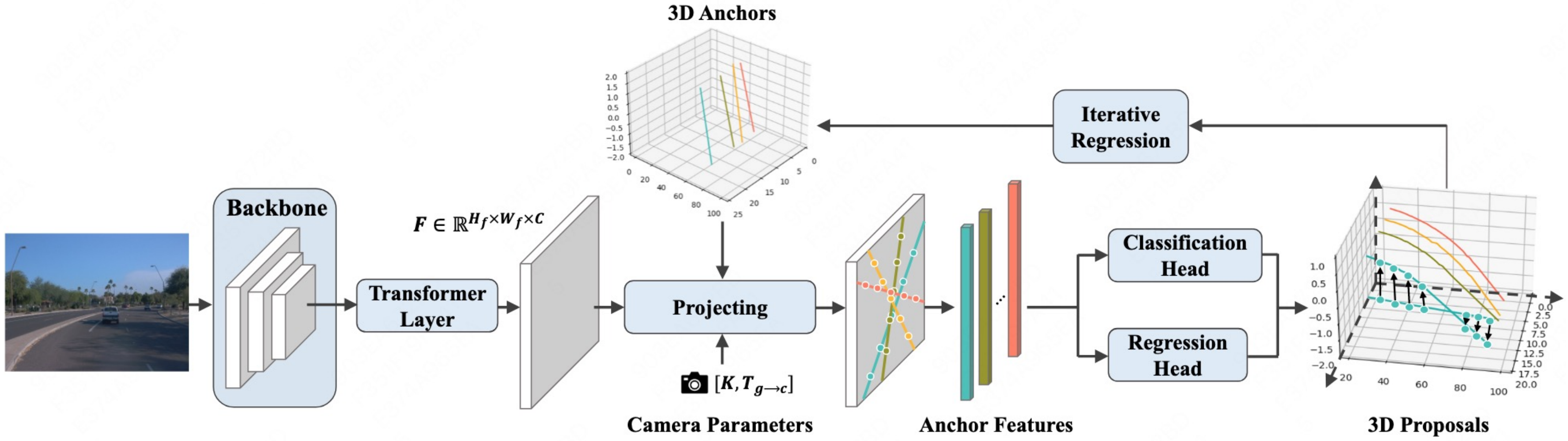
Representation of 3D Lanes

- The i-th lane $\boldsymbol{G}_i = \{\mathbf{p}_\mathrm{i}^k\}, k \in [1, N]$
- The k-th point of $\boldsymbol{G}_i$: $\mathbf{p}_i^k = (x_i^k, y^k, z_i^k, vis_i^k)$

Representation of 3D Anchors

- Starting from $(x_s, 0, 0)$, pitch θ, yaw ϕ
- The j-th anchor $\boldsymbol{A}_j = \{\boldsymbol{q}_\mathrm{i}^k\}, k \in [1, N]$
- The j-th point of $\boldsymbol{A}_j$: $\mathbf{q}_j^k = (x_j^k, y^k, z_j^k)$

Anchor3DLane



Anchor Projection and Feature Sampling

- Projecting $\boldsymbol{q}^k = (\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)$ to $\boldsymbol{q}'^k = (\boldsymbol{u}^k, \boldsymbol{v}^k)$

$$\begin{bmatrix} \tilde{u}^k \\ \tilde{v}^k \\ d^k \end{bmatrix} = \mathbf{K}\mathbf{T}_{g\to c} \begin{bmatrix} x^k \\ y^k \\ z^k \\ 1 \end{bmatrix},$$

$$u^k = W_f / W \cdot \frac{\tilde{u}^k}{d^k},$$

$$v^k = H_f / H \cdot \frac{\tilde{v}^k}{d^k},$$

- Sampling anchor feature as $\left\{\boldsymbol{F}_{(u^k, v^k)}\right\}_{k=1}^{N}$

3D Lane Prediction

- For each anchor, we have:
  - Classification probabilities $\boldsymbol{c}_j \in \mathbb{R}^L$
  - Offsets $(\Delta\boldsymbol{x}_j \in \mathbb{R}^N, \Delta\boldsymbol{z}_j \in \mathbb{R}^N) = \{(\Delta x_j^k, \Delta z_j^k)\}_{k=1}^{N}$
  - Visibility $\boldsymbol{vis}_j = \{vis_j^k\}_{k=1}^{N}$
- The j-th 3D proposal is generated as $\boldsymbol{P_j} = (\boldsymbol{c}_j, \boldsymbol{x}_j + \Delta\boldsymbol{x}_j, \boldsymbol{y}, \boldsymbol{z}_j + \Delta\boldsymbol{z}_j, \boldsymbol{vis}_j)$

- The generated 3D proposals can also be used as curve anchors for **iterative regression**

Loss functions of Anchor3DLane

- Positive samples are selected by distance metric between ground truth and anchors:

$$D(\mathbf{G}_i, \mathbf{A}_j) = \frac{\sum_{k=1}^{N} vis_i^k \cdot \sqrt{(x_i^k - x_j^k)^2 + (z_i^k - z_j^k)^2}}{\sum_{k=1}^{N} vis_i^k}.$$

- Overall loss function:

$$\mathcal{L}_{cls} = -\sum_{j=1}^{M}\sum_{l=1}^{L} \alpha^l (1 - c_j^l)^\gamma \log c_j^l,$$

$$\begin{aligned}\mathcal{L}_{reg} = &\sum_{i=1}^{M_p}\sum_{k=1}^{N} (\|\hat{vis}_i^k \cdot (x_i^k + \Delta x_i^k - \hat{x}_i^k)\|_1 \\ &+ \sum_{i=1}^{M_p}\sum_{k=1}^{N} \|\hat{vis}_i^k \cdot (z_i^k + \Delta z_i^k - \hat{z}_i^k)\|_1) \\ &+ \sum_{i=1}^{M_p}\sum_{k=1}^{N} \|\hat{vis}_i^k - vis_i^k\|_1.\end{aligned}$$

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{reg}\mathcal{L}_{reg}.$$

Temporal Context Modeling

- Anchor3DLane can be further extended to multi-frame 3D lane detection to incorporate temporal context for larger perception range

- 3D point (x_t, y_t, z_t) for the t-th frame's can be transformed into the t'-th frame's ground coordinate system by relative transformation matrix to gather anchor features from previous frame $\boldsymbol{T}_{g(t)\rightarrow g(t')}$:

$$\begin{bmatrix} x_{t'} \\ y_{t'} \\ z_{t'} \end{bmatrix} = \mathbf{T}_{g(t)\rightarrow g(t')} \begin{bmatrix} x_t \\ y_t \\ z_t \\ 1 \end{bmatrix},$$

- Cross-frame attention is then adopted to fuse sampled anchor features from the t-th frame and t'-th frame.

Equal-Width Constraint

- The geometry property of 3D lanes, i.e., lanes in 3D space are nearly parallel with each other can be formulated as an equal-width constraint to adjust the x-coordinates of lane predictions.

- Given $\{\mathrm{x}_j^k\}_{k=1}^N$ and $\{\mathrm{x}_{j'}^k\}_{k=1}^N$ as x-coordinates of two lanes and $\tilde{\Delta}\mathbf{x}^k$ as adjustment to x^k, the objective function is formulated as:

$$w_{j,j'}^k = |\cos\theta_j^k(x_j^k + \tilde{\Delta}x_j^k - x_{j'}^k - \tilde{\Delta}x_{j'}^k)|,$$

$$\mathcal{L}(\mathbf{w}_{j,j'}) = \sum_{k=1}^{N} |w_{j,j'}^k - \frac{1}{N}\sum_{k'=1}^{N} w_{j,j'}^{k'}|.$$

$$\min_{\{\tilde{\Delta}\mathbf{x}_j\}_{j\in[1,Q]}} \frac{1}{Q(Q-1)} \sum_{j=1}^{Q} \sum_{j'=1, j'\neq j}^{Q} \mathcal{L}(\mathbf{w}_{j,j'})$$

$$+ \alpha \frac{1}{Q} \sum_{j=1}^{Q} \|\tilde{\Delta}\mathbf{x}_j\|_2,$$

- The first term restricts the width to be consistent and the second term serves as a regularization

Ablation Study

Table 1: Comparison with BEV feature sampling

Input	Feat	F1(%)	x err/C(m)	x err/F(m)	z err/C(m)	z err/F(m)
BEV	BEV	47.6	0.466	0.421	0.119	0.170
FV	BEV	47.6	0.443	0.446	0.118	0.160
FV	FV	**53.1**	**0.300**	**0.31**	**0.103**	**0.139**

Table 2: Iterative regression

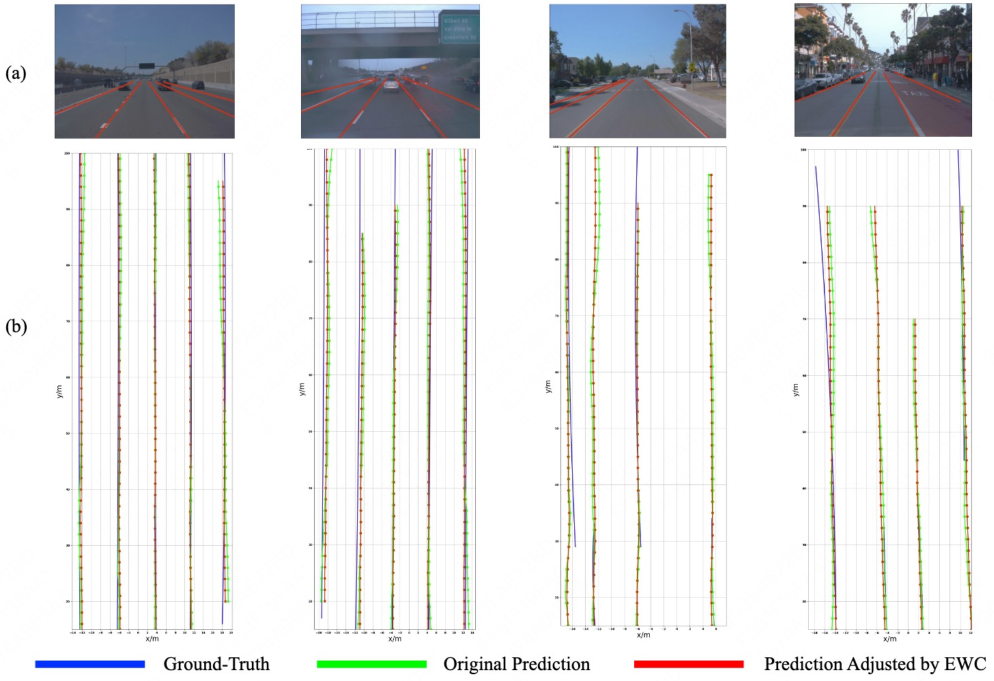
Iter	F1(%)	x err/C(m)	x err/F(m)	z err/C(m)	z err/F(m)
1	54.8	0.318	0.349	**0.101**	**0.147**
2	56.3	**0.287**	0.335	0.103	0.152
3	**57.0**	**0.287**	**0.327**	0.104	0.148

Table 3: Temporal integration method

Method	F1(%)	x err/C(m)	x err/F(m)	z err/C(m)	z err/F(m)
w/o Temporal	54.8	0.318	0.349	0.101	0.147
Linear Fusion	54.9	0.322	0.343	0.102	0.148
Weighted Sum	**55.8**	0.320	0.346	0.101	0.150
Attention	55.2	**0.308**	**0.330**	**0.099**	**0.145**

Table 4: Equal-Width Constraint (EWC)

Method	F1(%)	x err/C(m)	x err/F(m)
w/o EWC	54.8	**0.318**	0.349
w/ EWC	**55.0**	**0.318**	**0.337**



Quantitative Results

ApolloSim

Scene	Method	AP(%)↑	F1(%)↑	x err/C(m) ↓	x err/F(m) ↓	z err/C(m) ↓	z err/F(m) ↓
Balanced Scene	3DLaneNet [7]	89.3	86.4	0.068	0.477	0.015	**0.202**
	Gen-LaneNet [8]	90.1	88.1	0.061	0.496	0.012	0.214
	CLGo [20]	94.2	91.9	0.061	0.361	0.029	0.250
	PersFormer [5]	-	92.9	0.054	0.356	0.010	0.234
	GP [16]	93.8	91.9	0.049	0.387	**0.008**	0.213
	Anchor3DLane (Ours)	**97.2**	**95.6**	0.052	0.306	0.015	0.223
	Anchor3DLane†(Ours)	97.1	95.4	**0.045**	**0.300**	0.016	0.223
Rare Subset	3DLaneNet [7]	74.6	72.0	0.166	0.855	0.039	**0.521**
	Gen-LaneNet [8]	79.0	78.0	0.139	0.903	0.030	0.539
	CLGo [20]	88.3	86.1	0.147	0.735	0.071	0.609
	PersFormer [5]	-	87.5	0.107	0.782	0.024	0.602
	GP [16]	85.2	83.7	0.126	0.903	**0.023**	0.625
	Anchor3DLane (Ours)	**96.9**	**94.4**	0.094	**0.693**	0.027	0.579
	Anchor3DLane† (Ours)	95.9	**94.4**	**0.082**	0.699	0.030	0.580
Visual Variations	3D-LaneNet [7]	74.9	72.5	0.115	0.601	0.032	0.230
	Gen-LaneNet [8]	87.2	85.3	0.074	0.538	0.015	0.232
	CLGo [20]	89.2	87.3	0.084	0.464	0.045	0.312
	PersFormer [5]	-	89.6	0.074	0.430	0.015	0.266
	GP [16]	92.1	89.9	0.060	0.446	**0.011**	0.235
	Anchor3DLane (Ours)	**93.6**	91.4	0.068	0.367	0.020	0.232
	Anchor3DLane† (Ours)	92.5	**91.8**	**0.047**	**0.327**	0.019	**0.219**

Table 1. Comparison with state-of-the-art methods on ApolloSim dataset with three different split settings. "C" and "F" are short for close and far respectively. † denotes iterative regression.

ONCE-3DLane

Method	F1(%)↑	P(%)↑	R(%)↑	CD Error(m)↓
3D-LaneNet [7]	44.73	61.46	35.16	0.127
Gen-LaneNet [8]	45.59	63.95	35.42	0.121
SALAD [39]	64.07	75.90	55.42	0.098
PersFormer [5]	74.33	80.30	69.18	0.074
Anchor3DLane (Ours)	74.44	80.50	69.23	0.064
Anchor3DLane† (Ours)	**74.87**	**80.85**	**69.71**	**0.060**

Table 4. Comparison with state-of-the-art methods on ONCE-3DLanes validation set. Results under $\tau_{CD} = 0.3$ are displayed here. † denotes iterative regression. "P" and "R" are short for precision and recall respectively.

OpenLane

Method	F1(%)↑	Cate Acc(%)↑	x err/C(m) ↓	x err/F(m) ↓	z err/C(m) ↓	z err/F(m) ↓
3D-LaneNet [7]	44.1	-	0.479	0.572	0.367	0.443
GenLaneNet [8]	32.3	-	0.591	0.684	0.411	0.521
PersFormer [5]	50.5	**92.3**	0.485	0.553	0.364	0.431
Anchor3DLane (Ours)	53.1	90.0	0.300	0.311	**0.103**	0.139
Anchor3DLane† (Ours)	53.7	90.9	0.276	0.311	0.107	0.138
Anchor3DLane-T† (Ours)	**54.3**	90.7	**0.275**	**0.310**	0.105	**0.135**

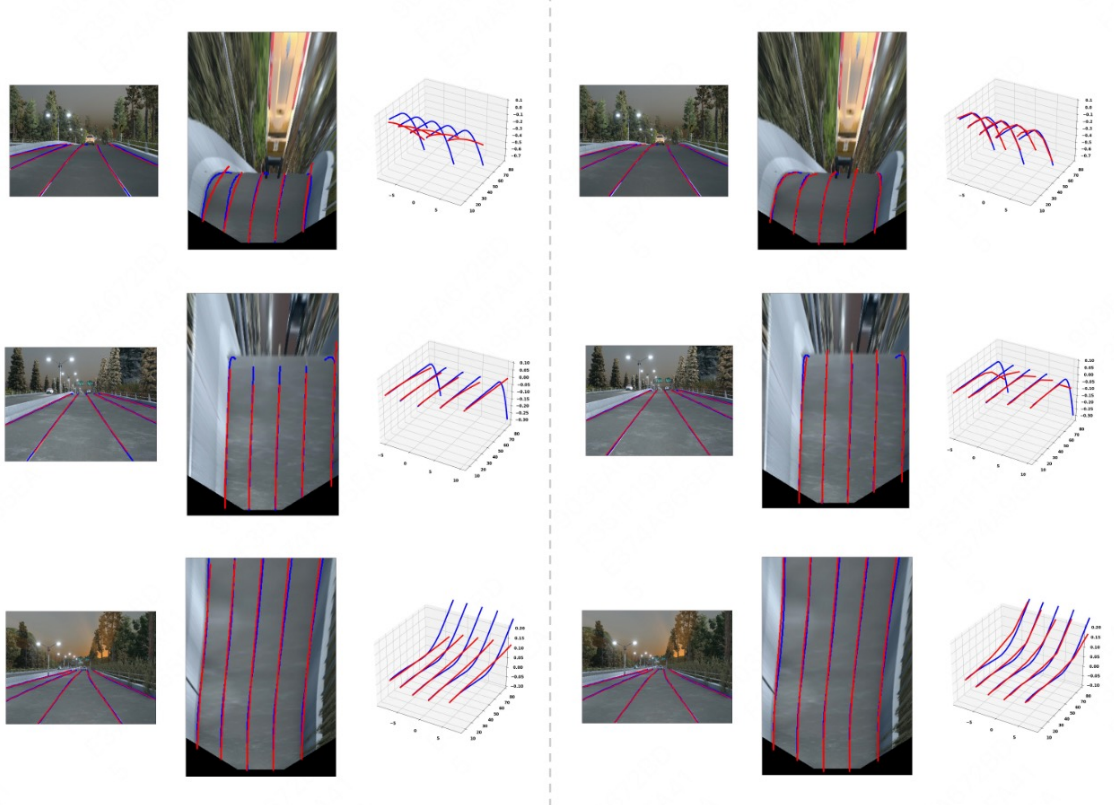
Table 2. Comparison with state-of-the-art methods on OpenLane validation set. † denotes iterative regression. Anchor3DLane-T denotes incorporating multi-frame information. "Cate Acc" means category accuracy.

Method	All	Up & Down	Curve	Extreme Weather	Night	Intersection	Merge & Split
3D-LaneNet [7]	44.1	40.8	46.5	47.5	41.5	32.1	41.7
GenLaneNet [8]	32.3	25.4	33.5	28.1	18.7	21.4	31.0
PersFormer [5]	50.5	42.4	55.6	48.6	46.6	40.0	50.7
Anchor3DLane (Ours)	53.1	45.5	56.2	51.9	47.2	44.2	50.5
Anchor3DLane† (Ours)	53.7	46.7	57.2	52.5	47.8	45.4	51.2
Anchor3DLane-T† (Ours)	**54.3**	**47.2**	**58.0**	**52.7**	**48.7**	**45.8**	**51.7**

Table 3. Comparison with state-of-the-art methods on OpenLane validation set. F1 score is presented for each scenario. † denotes iterative regression. Anchor3DLane-T denotes incorporating multi-frame information.

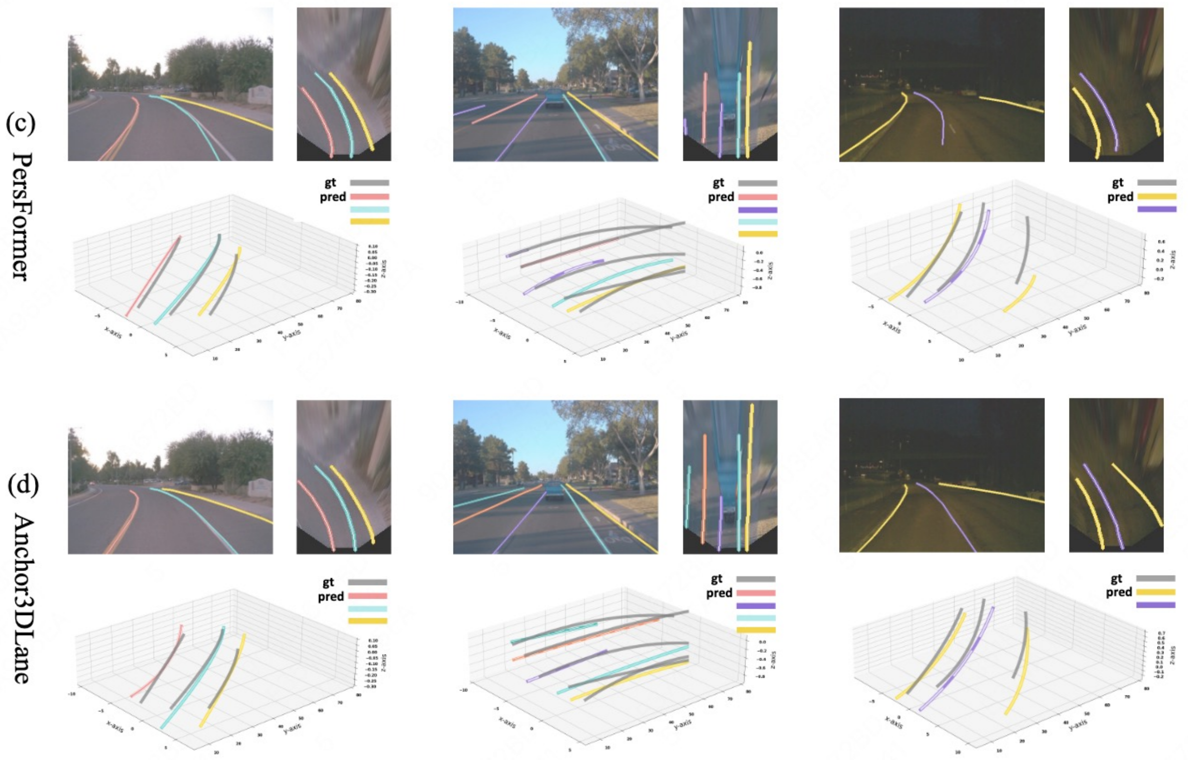
Qualitative Results

ApolloSim



(a) CLGo (b) Anchor3DLane

OpenLane



Ground-truth

Prediction

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thanks for Listening!

Code is available at: https://github.com/tusen-ai/Anchor3DLane

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences

北京航空航天大学

tu 图森未来