VQ-Map: Bird's-Eye-View Map Layout Estimation in Tokenized Discrete Space via Vector Quantization

Yiwei Zhang· Jin Gao · Fudong Ge · Bing Li

Zhaoxiang Zhang · Haibin Ling · Weiming Hu

State Key Laboratory of Multimodal Artificial Intelligence Systems(MAIS), CASIA

School of Artificial Intelligence, University of Chinese Academy of Sciences

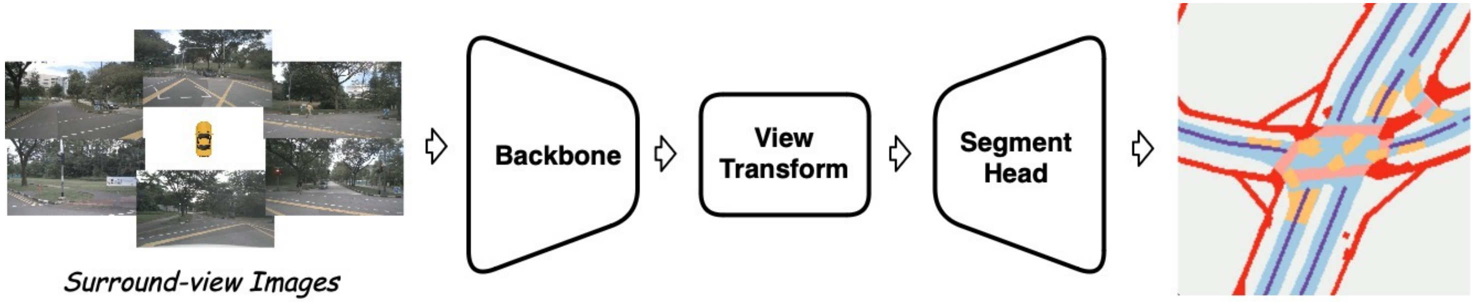
School of Information Science and Technology, ShanghaiTech University

Center for Artificial Intelligence and Robotics, HKISI, CAS,

People AI Stony Brook University

Background and Motivations

BEV Map Segmentation Pipeline for Map Layout Estimation

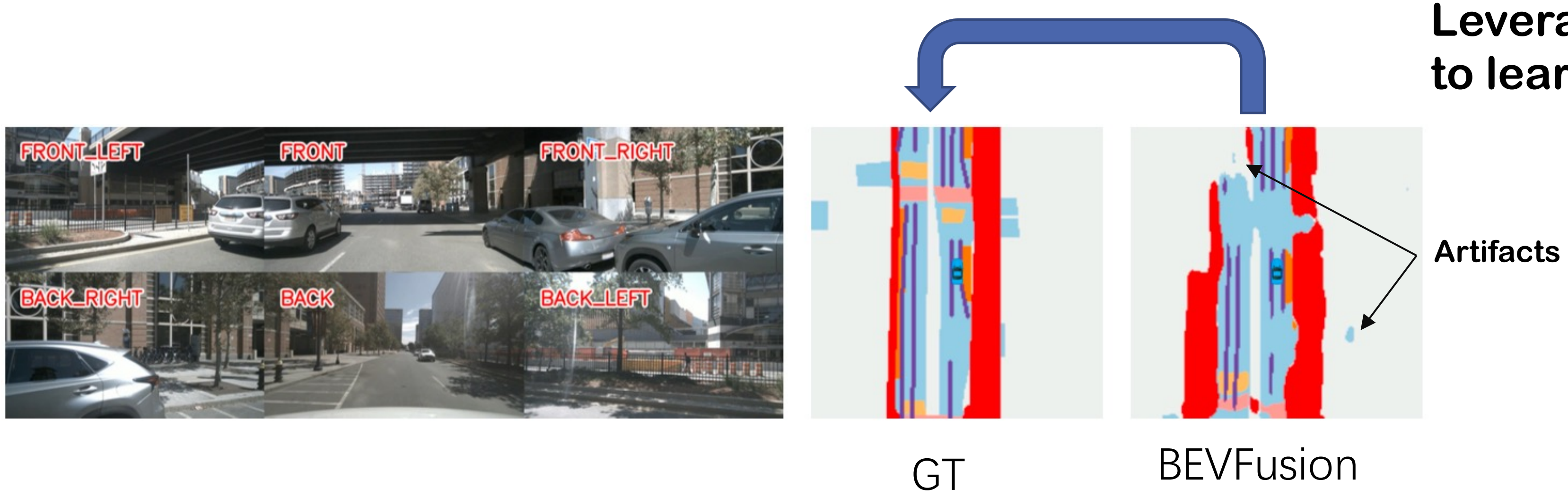


Many challenges
- **Occlusion**
- **Low resolution**
- **Unfavourable imaging conditions**

Inaccurate dense BEV features

Incoherent and Unrealistic results

Lack of Prior knowledge

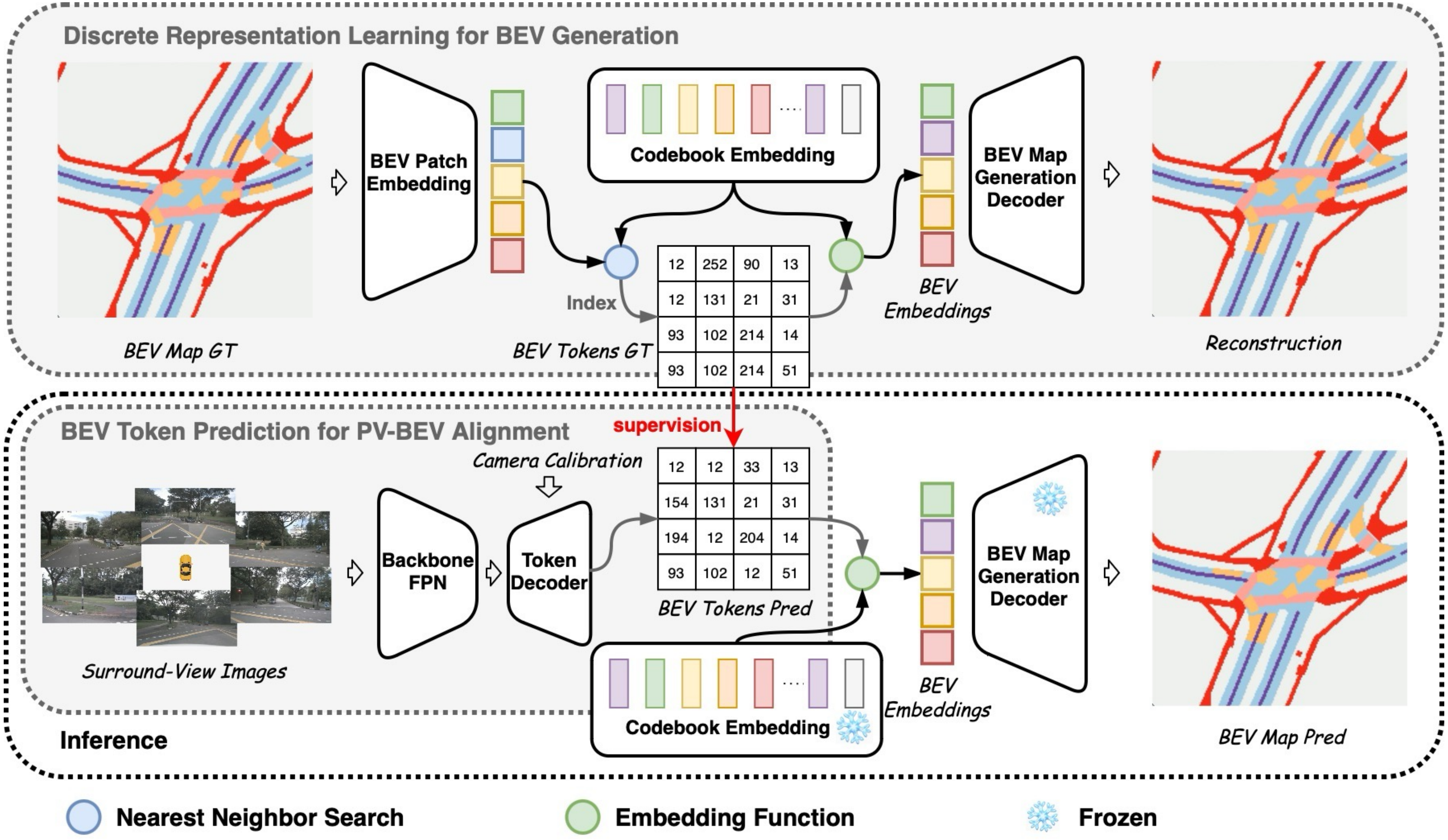


Leverage generative models to learn this prior knowledge

How to align the PV features with the generative models to facilitate BEV map estimation

NEURAL INFORMATION PROCESSING SYSTEMS

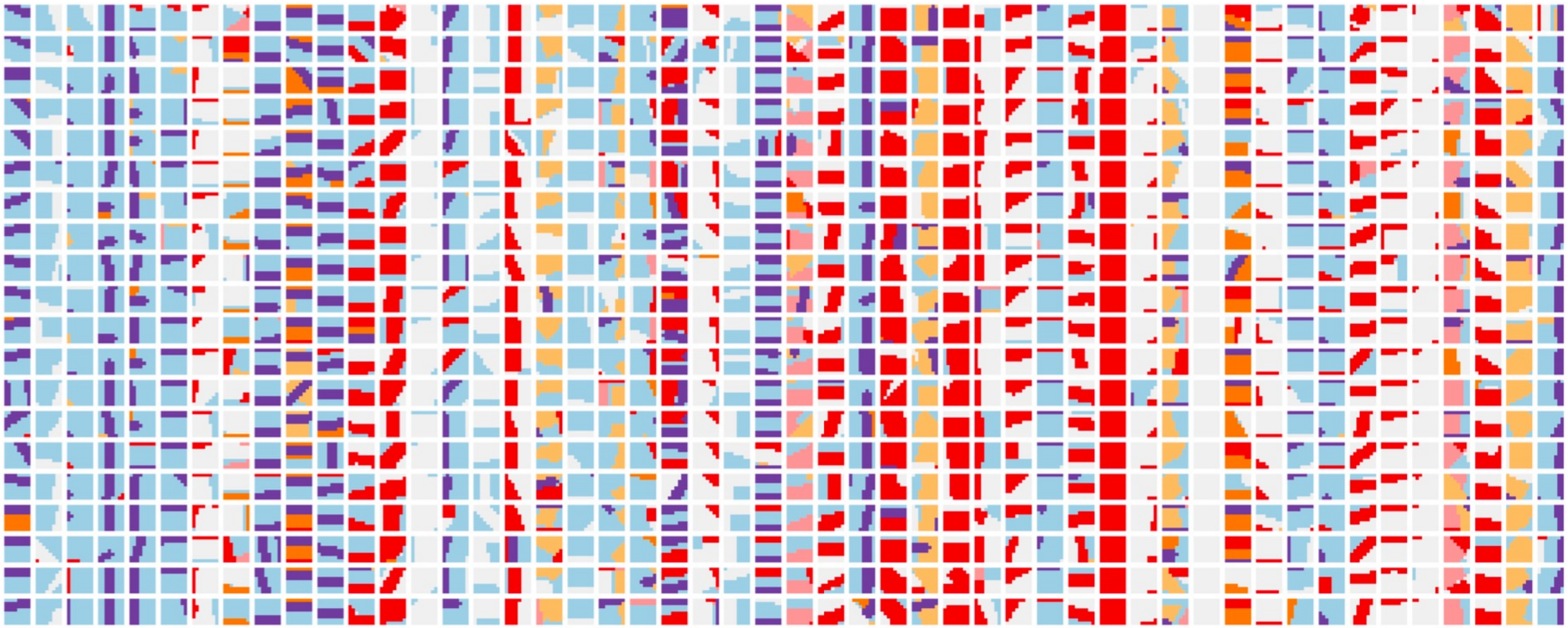
VQ-Map



Discrete Representation Learning for BEV Generation
BEV Map GT
BEV Patch Embedding
Codebook Embedding
Index
12 252 90 13
12 131 21 31
93 102 214 14
93 102 214 51
BEV Tokens GT
BEV Embeddings
BEV Map Generation Decoder
Reconstruction
BEV Token Prediction for PV-BEV Alignment
supervision
Camera Calibration
Surround-View Images
Backbone FPN
Token Decoder
12 12 33 13
154 131 21 31
194 12 204 14
93 102 12 51
BEV Tokens Pred
Codebook Embedding
BEV Embeddings
BEV Map Generation Decoder
BEV Map Pred
Inference
Nearest Neighbor Search
Embedding Function
Frozen

Results

What Has the Codebook Embedding Learned?



All BEV patch images in the same column correspond to the same token.

Results

Surround-View Map Layout Estimation Result

Methods	IoU ↑ (%)						
	Drivable	Ped. Cross.	Walkway	Stopline	Carpark	Divider	Mean
OFT [36]	74.0	35.3	45.9	27.5	35.9	33.9	42.1
LSS [3]	75.4	38.8	46.3	30.3	39.1	36.5	44.4
CVT [37]	74.3	36.8	39.9	25.8	35.0	29.4	40.2
M^{2}BEV [38]	77.2	-	-	-	-	40.5	-
BEVFusion [1]	81.7	54.8	58.4	47.4	50.7	46.4	56.6
MapPrior [17]	81.7	54.6	58.3	46.7	53.3	45.1	56.7
X-Align [34]	82.4	55.6	59.3	49.6	53.8	47.4	58.0
MetaBEV [35]	83.3	56.7	61.4	50.8	55.5	48.0	59.3
DDP [19]	83.6	58.3	61.6	52.4	51.4	49.2	59.4
VQ-Map	**83.8**	**60.9**	**64.2**	**57.7**	**55.7**	**50.8**	**62.2**

Monocular Map Layout Estimation Result

Methods	nuScenes [9]					Argoverse [10]
	Drivable	Crossing	Walkway	Carpark	Mean	Drivable
IPM [39]	40.1	-	14.0	-	-	43.7
Depth Unpr. [39]	27.1	-	14.1	-	-	33.0
VED [40]	54.7	12.0	20.7	13.5	25.2	62.9
VPN [41]	58.0	27.3	29.4	12.9	31.9	64.9
PON [39]	60.4	28.0	31.0	18.4	34.5	65.4
DiffBEV [20]	65.4	41.3	41.1	28.4	44.1	-
GitNet [42]	65.1	41.6	42.1	31.9	45.2	67.1
TaDe [16]	65.9	40.9	42.3	30.7	45.0	68.3
VQ-Map	**70.0**	**43.9**	**43.8**	**32.7**	**47.6**	**73.4**

VQ-Map establishes new state-of-the-art performance on camera-based BEV semantic map layout estimation.

Ablation Study

		(a)	(b)	(c)	(d)	(e)	(f)	(g)
Arch.	Sparse Feature			✓	✓	✓	✓	✓
	Codebook Embedding		✓		✓	✓	✓	✓
Supervision		**M**	$\{k_q^i\}_{i=1}^N$	**M**	$\{\mathbf{z}_c^i\}_{i=1}^N$	$\{\mathbf{z}_q^i\}_{i=1}^N$	$\{k_q^i\}_{i=1}^N$	$\{k_q^i\}_{i=1}^N$
Drivable		81.5	80.4	**83.9**	82.5	82.5	83.6	83.5
Ped. Cross.		54.2	52.9	60.0	59.7	59.1	**60.1**	59.9
Walkway		58.1	58.2	63.5	62.2	62.1	**63.5**	63.4
Stop Line		46.1	47.2	53.2	55.1	54.9	**56.8**	56.8
Carpark		53.2	52.7	51.0	53.4	53.1	**56.2**	55.1
Divider		45.3	46.6	46.9	48.9	49.0	50.3	**50.7**
Mean		56.4	56.3	59.8	60.3	60.1	**61.8**	61.6
Improvements			-0.1	3.4	3.9	3.7	**5.4**	5.2

As a PV-BEV alignment method, token classification is more effective than value regression.

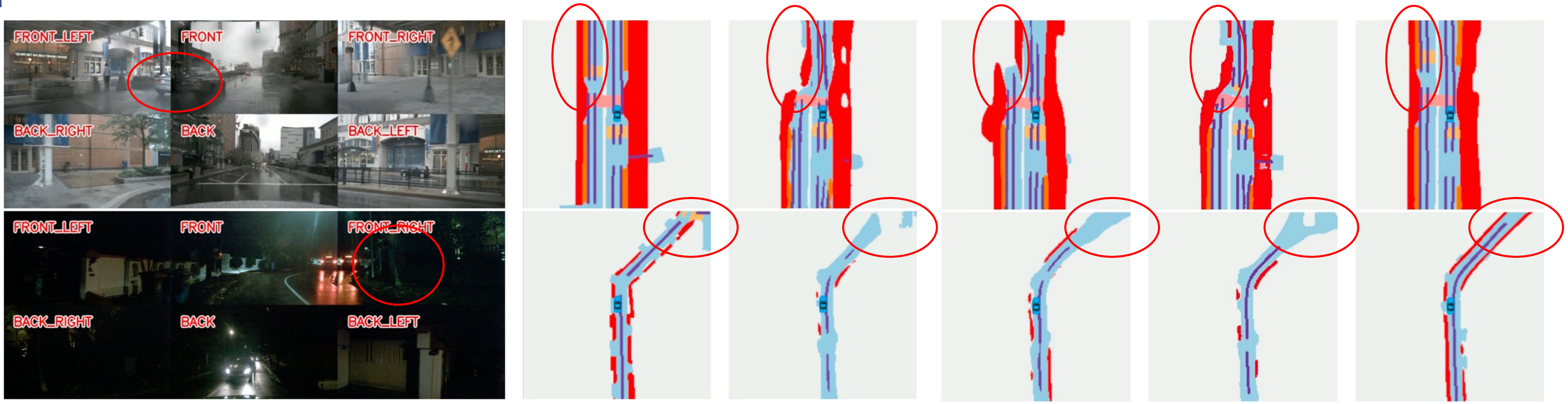
Computational overhead analysis

Method	mIoU↑(%)	Params↓(M)	MACs↓(G)	Training Time↓(h)
BEVFusion	56.6	50.1	155.5	**100**
MapPrior	56.7	719.1	396.0	>200
DDP(3 steps)	59.4	53.6	614.1	160
VQ-Map(tiny)	59.6	**44.2**	**86.8**	30+74=104
VQ-Map(light)	60.1	81.9	137.3	35+80=115
VQ-Map	**62.2**	108.3	231.6	35+96=131

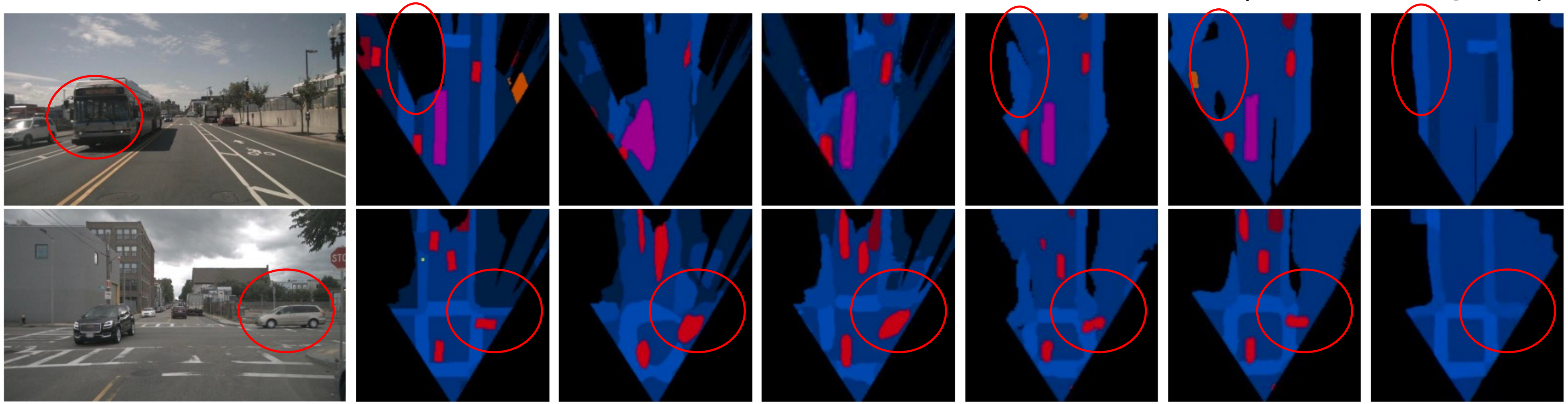
Our approach not only demonstrates strong performance, but also saves much computational cost in comparison to the recent SOTA methods MapPrior and DDP.

NEURAL INFORMATION PROCESSING SYSTEMS

Visualization



GT BEVFusion DDP MapPrior VQ-Map



GT VPN PON GitNet TaDe VQ-Map

Conclusion

A novel pipeline VQ-Map by aligning with the generative models using discrete BEV tokens for BEV Layout Estimation

- VQ-Map gets new state-of-the-art performance on camera-based BEV semantic segmentation.
- The core components of our method are the codebook embedding constructed via vector quantization, serving as a bridge between PV and BEV.
- As a PV-BEV alignment method, token classification is more effective than value regression.

Paper

Code