NeurIPS 2024

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

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

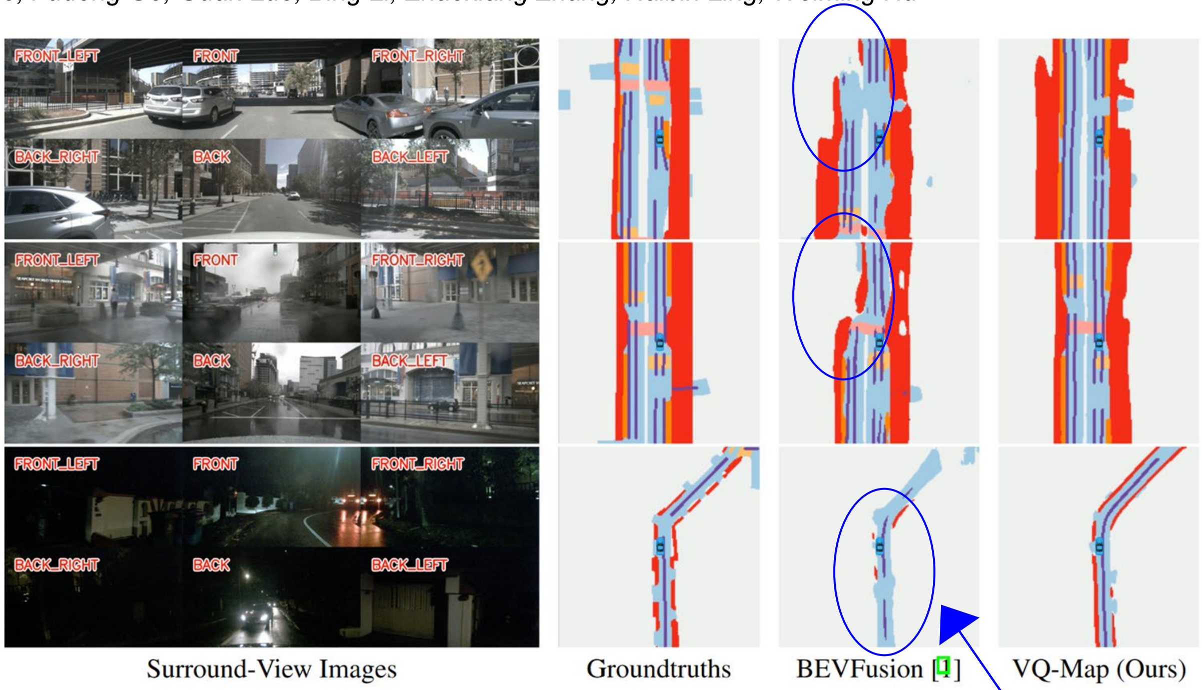
- Problem/Objective
  - BEV map segmentation (Multi / Monocular - Camera)

- Contribution/Key Idea
  - VQ-VAE에서 영감을 받은 Token Embedding / Codebook 네트워크 제안
  - PV - BEV 사이의 새로운 연결고리 제시
  - SOTA at BEV map segmentation Task

김범준

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

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

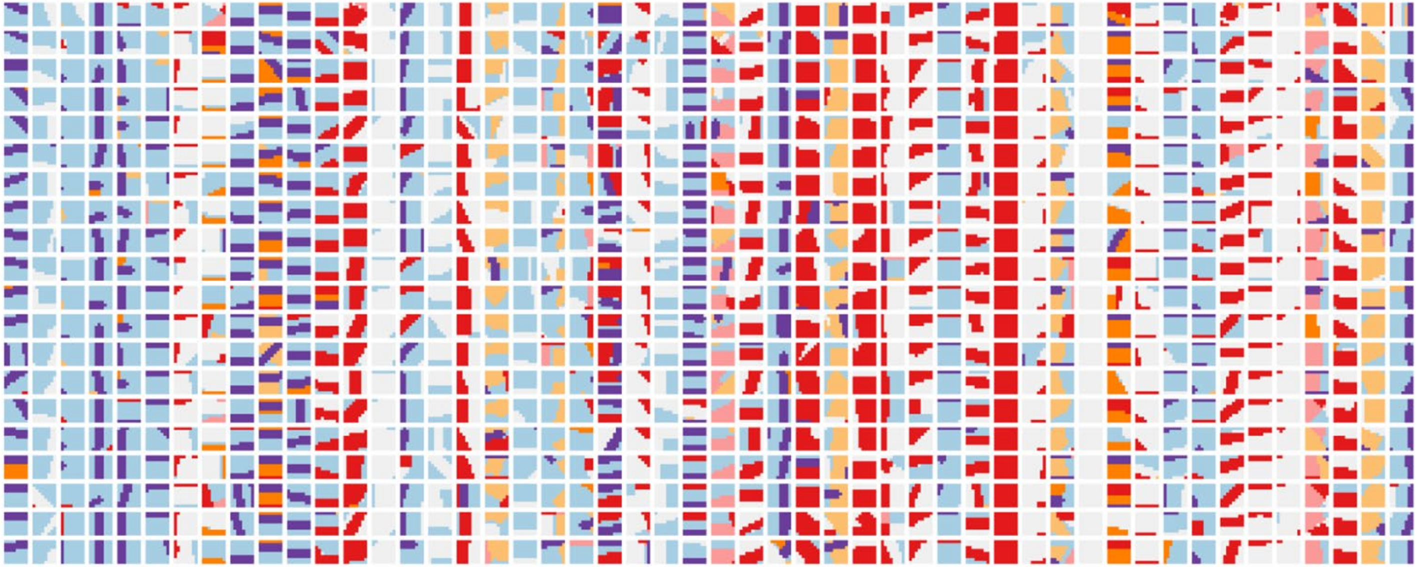


- Limitation
  - Semantic map을 generate 하는 것에만 주목하여 <u>“Map prior knowledge”를 사용하지 않음</u>
  - PV ←→ BEV 사이의 관계점에 덜 주목

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

NeurIPS 2024

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

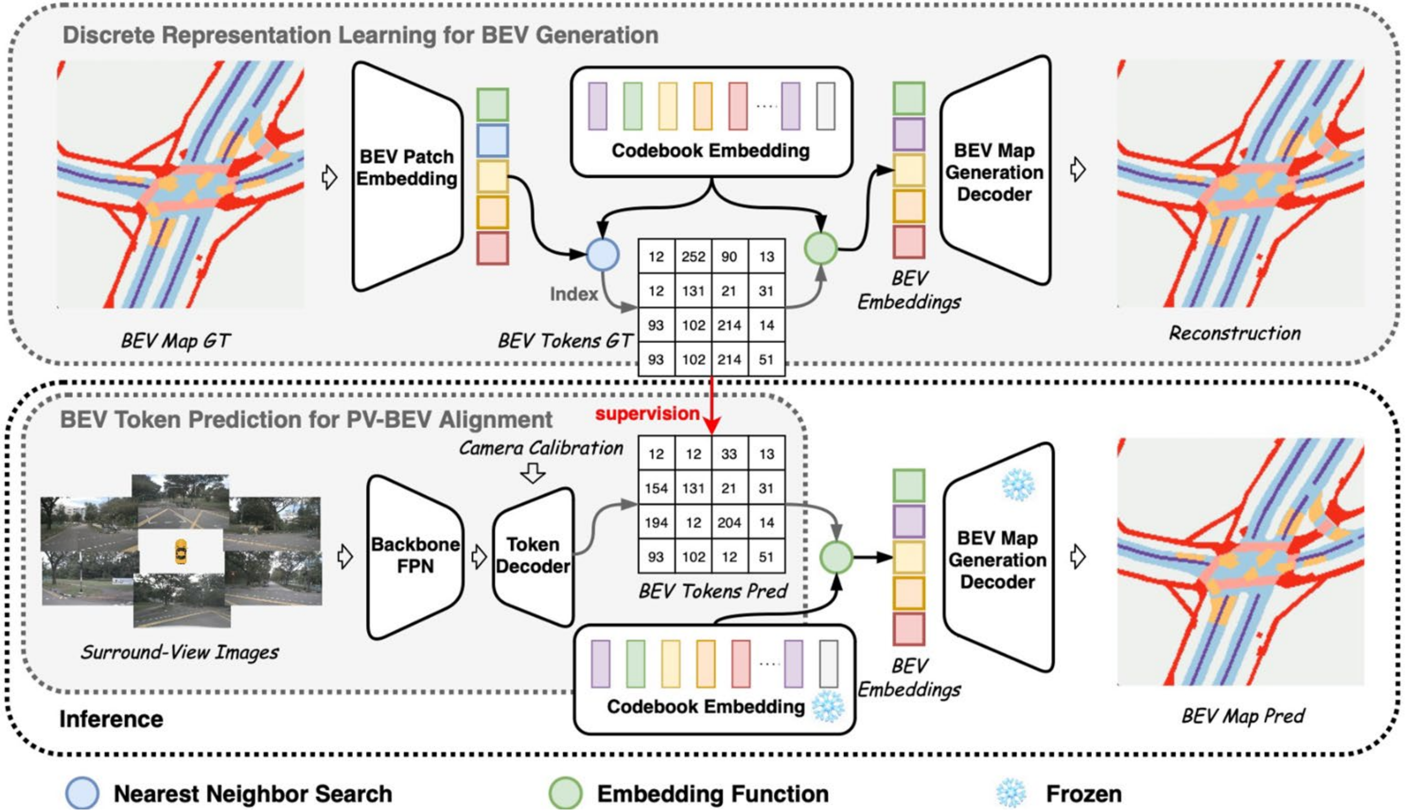


- Overcome / Contribution
  - GT semantic map을 나누어 BEV tokenize 하여 Codebook으로 저장하여 사용
  - PV ←→ BEV의 정보를 같이 사용하는 attention 모듈 제안

김범준

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

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu



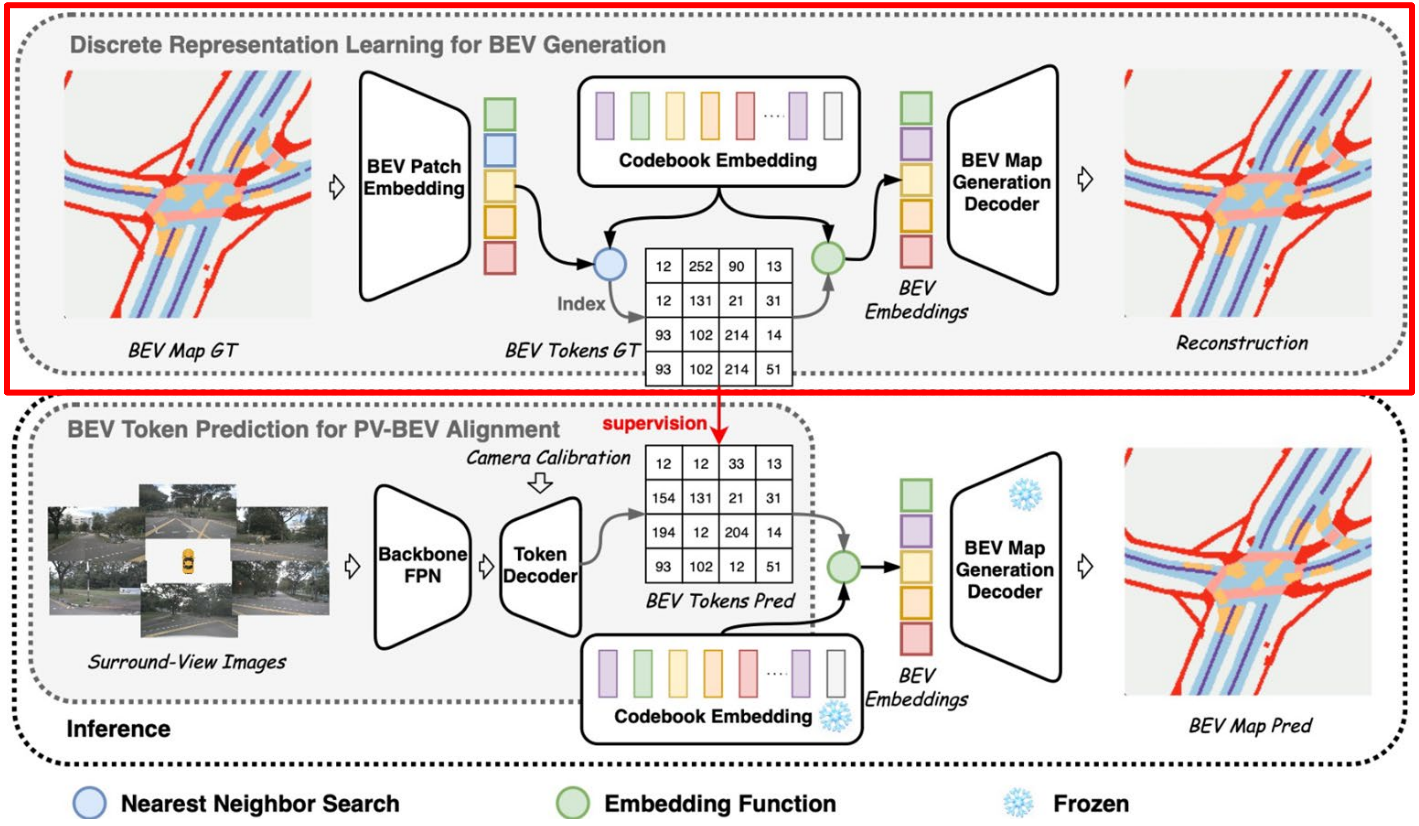
김범준

[VQ-Map: Bird’s-Eye-View Map Layout Estimation in Tokenized Discrete Space via Vector Quantization](#)

NeurIPS 2024

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

Stage 1 (Training)



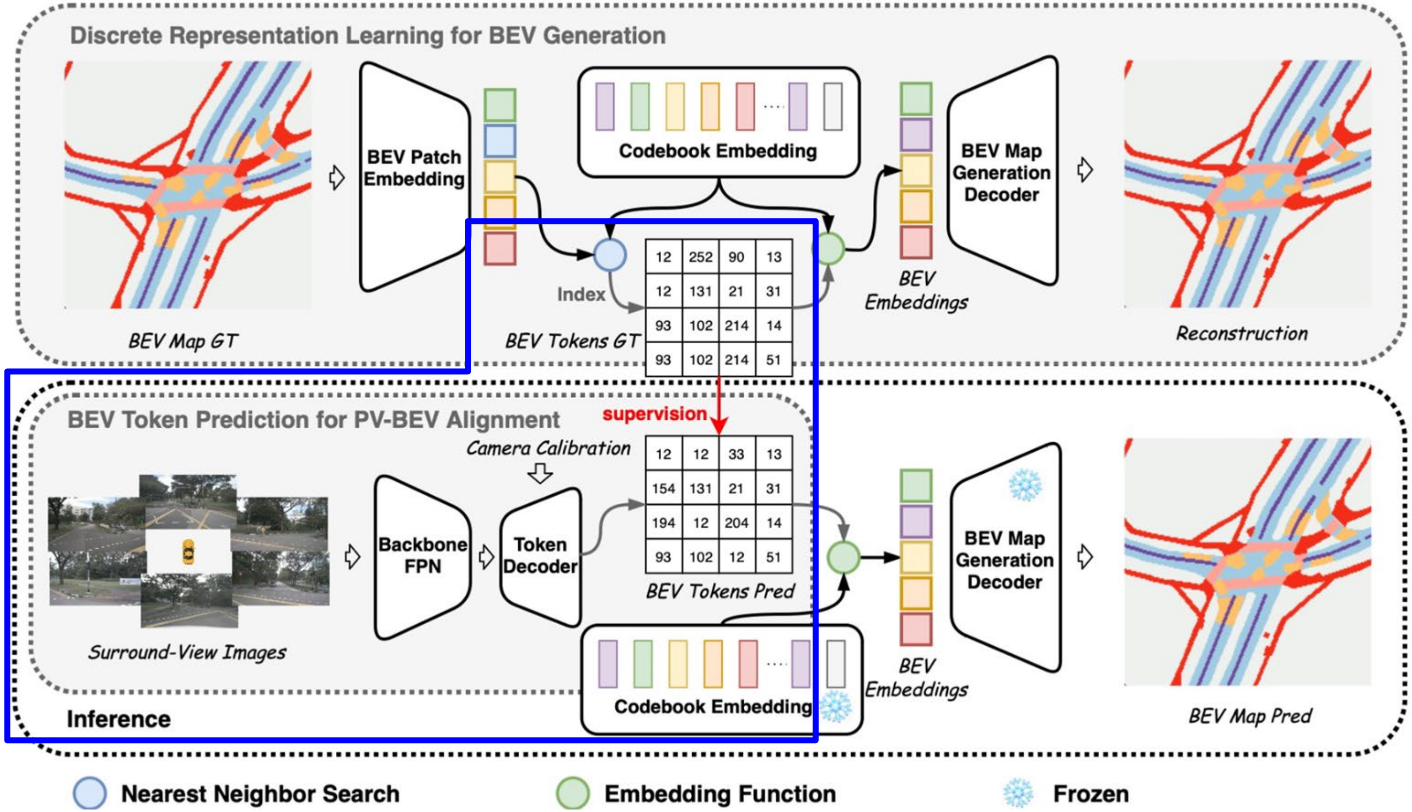
김범준

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

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

NeurIPS 2024

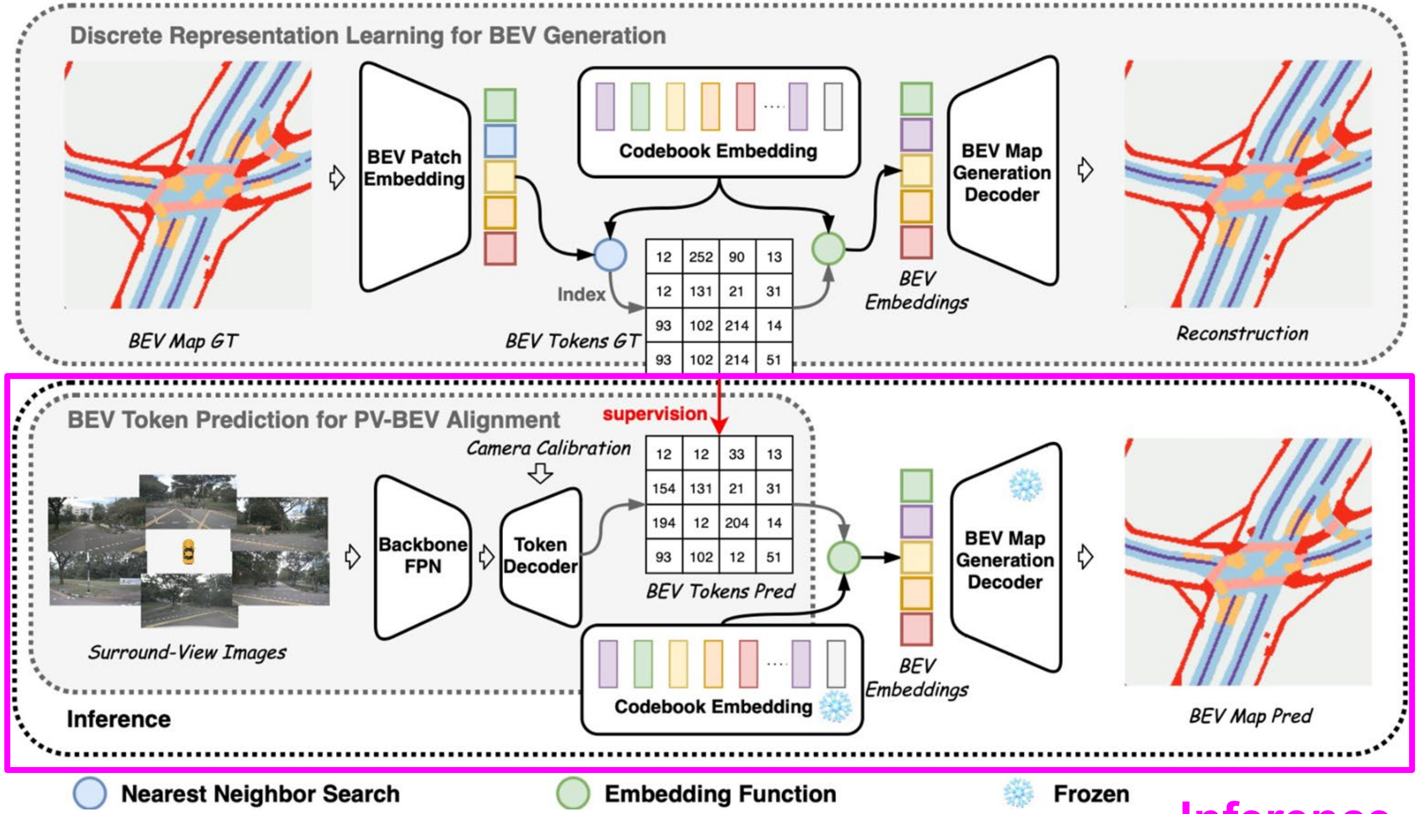
Stage 2 (Training)



김범준

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

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu



Inference

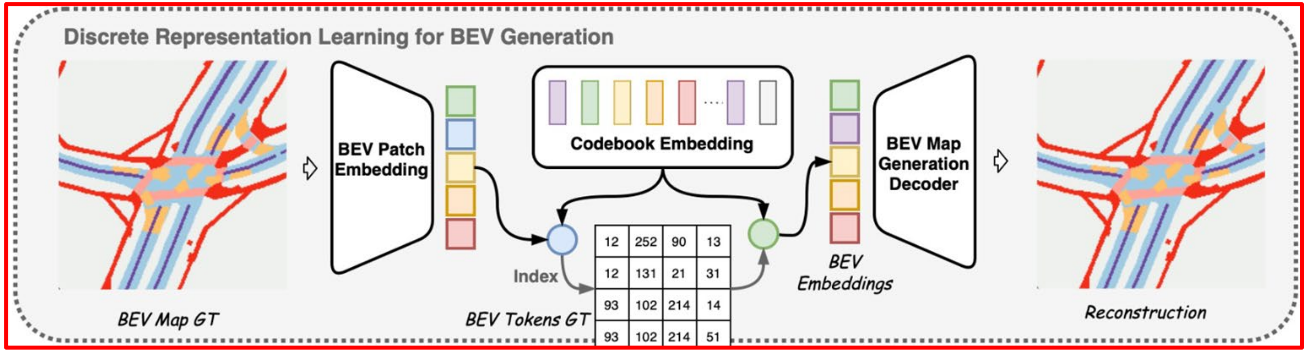
김범준

NeurIPS 2024

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

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

Stage 1 (Training)



$$\mathbf{M} \in \mathbb{B}^{C \times H \times W}$$

$$\{\mathbf{M}^i \in \mathbb{B}^{C \times P \times P}\}_{i=1}^{N} \xrightarrow{\mathcal{E}} \mathbf{z}^i \in \mathbb{R}^D \xleftrightarrow{\mathcal{Q}} \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k, \ldots, \mathbf{v}_K\}$$

$$N = HW/P^2$$

BEV patch embedding

Vector Quantization

Codebook

Embedded D : 512
K = 256

GT의 H=W=200
Patch size P =8
→ N=625

Embedded D : 512

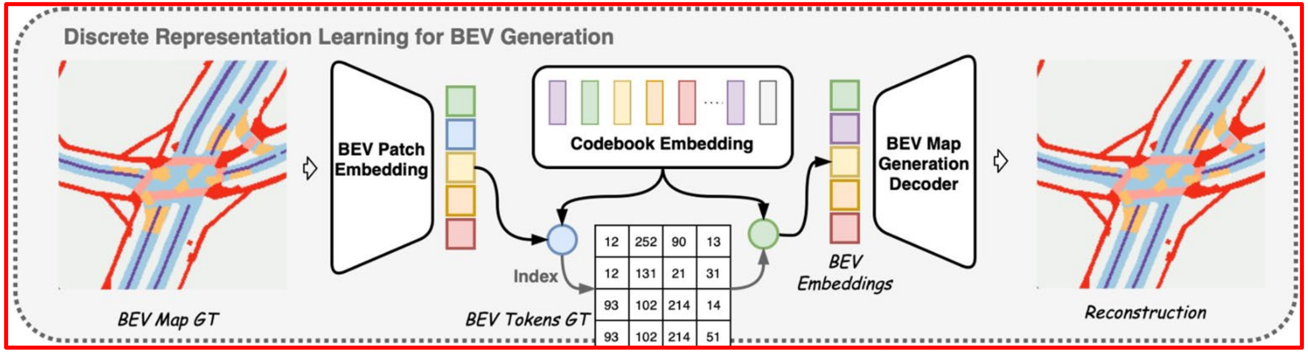
김범준

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

NeurIPS 2024

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

Stage 1 (Training)



continuous latent vector $\mathbf{z}_c$ $\xleftrightarrow{}$ $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k, \ldots, \mathbf{v}_K\}$ $\longrightarrow$ $\mathbf{M}' = \mathcal{D}(\mathcal{Q}(\mathcal{E}(\mathbf{M})))$

$\mathcal{Q}$

Vector Quantization

$\mathcal{D}$

BEV Map Generation Decoder

$$\{\mathbf{z}_q^i = \ell_2(\mathbf{v}_{k_q^i})\}_{i=1}^N$$

$$\mathbf{z}_q = \mathcal{Q}(\mathbf{z}_c) = \arg \min_{\ell_2(\mathbf{v}_k)} \|\ell_2(\mathbf{z}_c) - \ell_2(\mathbf{v}_k)\|_2$$

$$k_q = \arg\min_k \|\ell_2(\mathbf{z}_c) - \ell_2(\mathbf{v}_k)\|_2$$

Zc에 가장 가까운 codebook vector

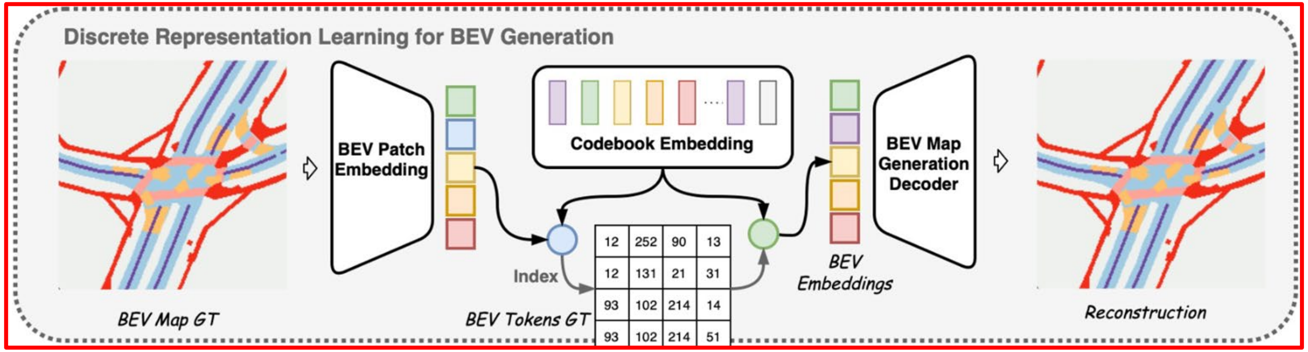
그때의 index 를 저장하여 **GT로 사용**

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

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

NeurIPS 2024

Stage 1 (Training)



$$\mathbf{M}' = \mathcal{D}(\mathcal{Q}(\mathcal{E}(\mathbf{M})))$$

Patch augmentation (ex. translation.. etc)

$$\mathcal{L}_{vq} = \frac{1}{N}\sum_{i=1}^{N}\left(\left\|\mathbf{z}_q^i - \mathrm{sg}(\ell_2(\mathbf{z}_c^i))\right\|_2^2 + \left\|\mathrm{sg}(\mathbf{z}_q^i) - \ell_2(\mathbf{z}_c^i)\right\|_2^2 + \sum_{j=1}^{N_{\mathrm{aug}}}\left\|\mathrm{sg}(\mathbf{z}_q^i) - \ell_2(\tilde{\mathbf{z}}_c^{i,j})\right\|_2^2\right)$$

$$\mathcal{L}_{re} = \frac{1}{C}\sum_{c=1}^{C}\frac{\left\|\mathbf{M}_c - \mathbf{M}'_c\right\|_2^2}{1 + \left\|\mathbf{M}_c\right\|_1}$$

Class 별 가중치 보정

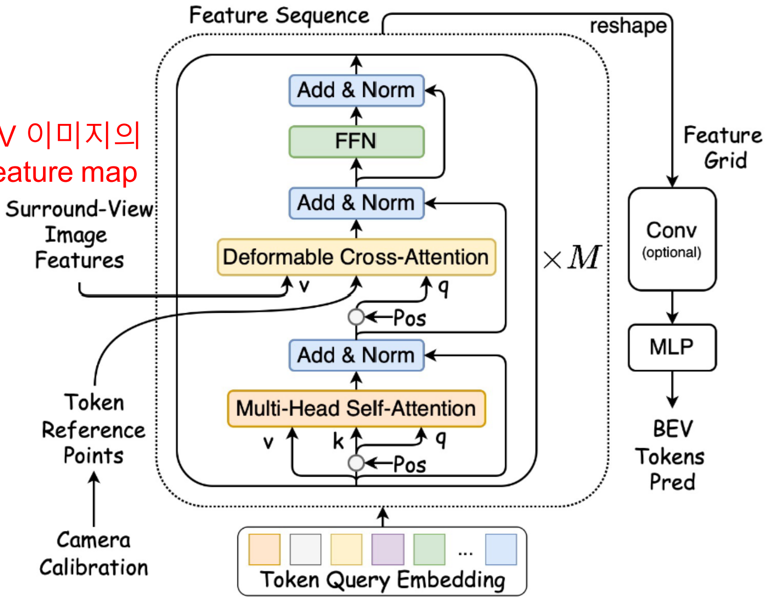
$$\mathcal{L} = \mathcal{L}_{re} + \mathcal{L}_{vq}$$

김범준

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

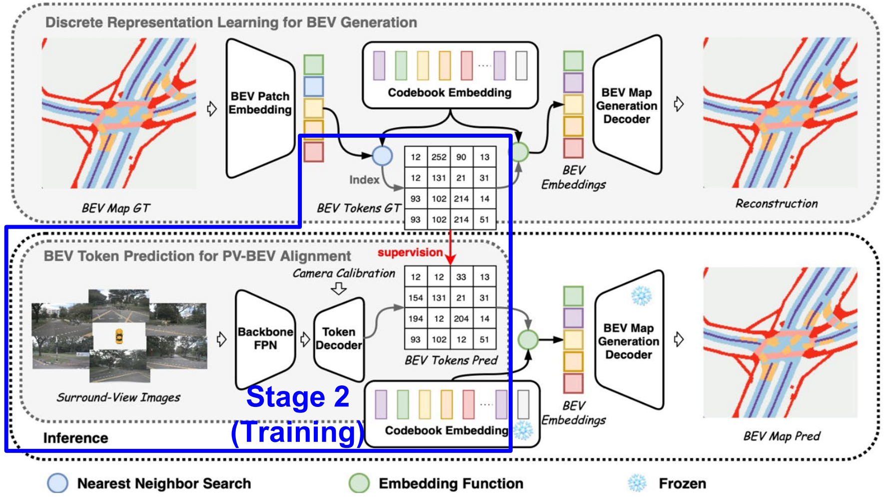
Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

PV 이미지의 feature map



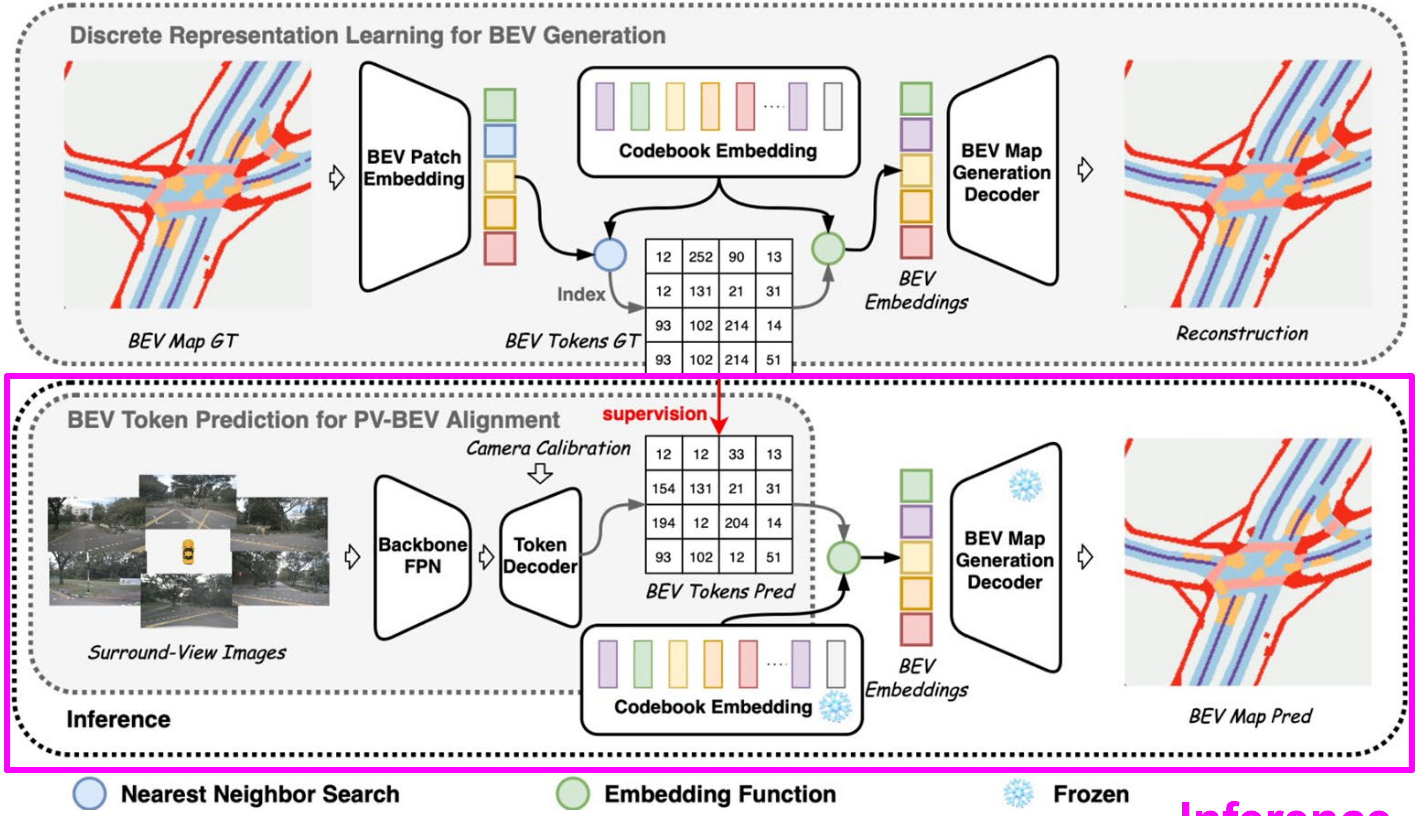
PV 상에서의 위치 파악

Patch 별 초기 embedding (학습)



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

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu



Inference

김범준

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

Table 1: State-of-the-art comparison for the surround-view BEV map layout estimation on the nuScenes **validation** set. MapPrior [17] uses a fixed IoU threshold of 0.5, while other methods apply the threshold that maximizes IoU according to their original settings. In our method, we adopt a constant IoU threshold of 0.5 to ensure a fairer comparison across all existing approaches. We only evaluate different approaches in the camera-only setting.

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^2BEV [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**

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

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

Table 2: State-of-the-art comparison for the monocular BEV map layout estimation on the nuScenes and Argoverse **validation** sets using the IoU (%) metric. Our VQ-Map uses the IoU threshold of 0.5 while other methods choose the best threshold following their original settings. During the evaluation process, grid cells that cannot be reached by LiDAR are ignored [39].

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**

김범준

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

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

NeurIPS 2024

Table 3: Ablation experiments on some key parameters of the token decoder. We perform ablations on the token decoder layer number M using layer dimension of 512, and ablations on different layer dimension by setting M to 8.

(a) Ablation for M of token decoder.

M	2	4	6	8
Drivable	81.1	82.7	83.6	**83.8**
Ped. Cross.	55.9	58.2	60.1	**60.9**
Walkway	59.2	61.7	63.5	**64.2**
Stop Line	50.9	55.1	56.8	**57.7**
Carpark	49.9	52.2	**56.2**	55.7
Divider	47.3	49.0	50.3	**50.8**
Mean	57.4	59.8	61.8	**62.2**

(b) Ablation for the layer dimension of token decoder.

Layer Dimension	256	512	768
Drivable	83.0	**83.8**	82.9
Ped. Cross.	58.4	**60.9**	57.9
Walkway	62.4	**64.2**	61.6
Stop Line	54.5	**57.7**	54.1
Carpark	53.6	**55.7**	52.5
Divider	48.5	**50.8**	48.6
Mean	60.1	**62.2**	59.6

김범준

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

Yiwei Zhang, Jin Gao, Fudong Ge, Guan Luo, Bing Li, Zhaoxiang Zhang, Haibin Ling, Weiming Hu

Table 5: Computational overhead analysis. Training time is measured in GPU hours using NVIDIA A100 (40G). Our method, even in its tiny version, surpasses the previous SOTA DDP (3 steps). Additionally, the computational cost (MACs) of our tiny version is significantly lower than previous methods. For the standard version of the model, it achieves substantial performance improvements while maintaining a relatively low computational cost.

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