BEV-MAE: Bird's Eye View Masked Autoencoders for Point Cloud Pre-training in Autonomous Driving Scenarios

AAAI 2024

Zhiwei Lin1, Yongtao Wang1, Shengxiang Qi2, Nan Dong2, Ming-Hsuan Yang3*

- Problem/Objective
  - 3D detection module의 pre-trained encoder 제시

- Contribution/Key Idea
  - self-supervised pre-training Autoencoder module
  - Learnable point token
  - Saving training cost(time, label cost .. etc) + SOTA

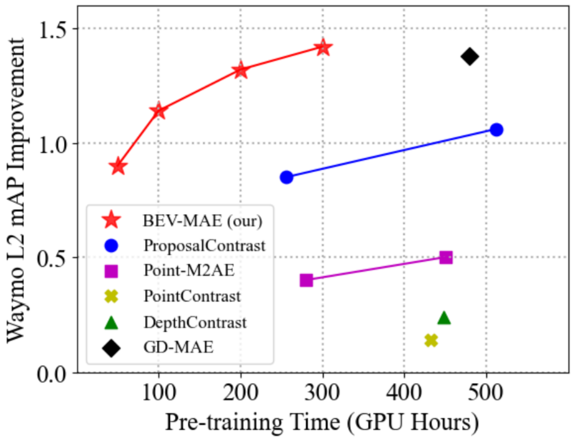


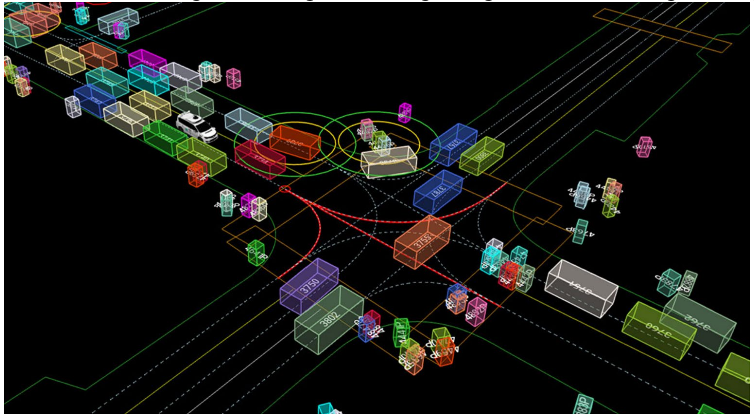
Figure 1: **Performance improvement *vs.* Pre-training time trade-off.** All entries are benchmarked by a P40 GPU. The 3D object detector is CenterPoint (Yin, Zhou, and Krahenbuhl 2021). All models are pre-trained on full Waymo and then fine-tuned with 20% training samples on Waymo.

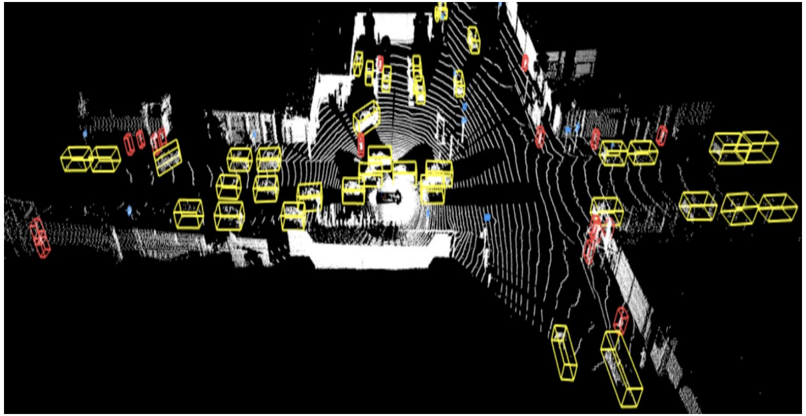
김범준

BEV-MAE: Bird's Eye View Masked Autoencoders for Point Cloud Pre-training in Autonomous Driving Scenarios

AAAI 2024

Zhiwei Lin1, Yongtao Wang1, Shengxiang Qi2, Nan Dong2, Ming-Hsuan Yang3*





- 3D object detection
  - 최근에는 모두 scratch로 부터 학습 (pre-train을 모델을 활용 x)
    - 학습 시간이 오래 걸림
    - Labeled data에 매우 의존
      - bounding box / classification label → expensive
      - one object average 114s

  - self-supervised을 활용하여 pre-training 해보자

    → self-supervised learning in **masked modeling**

김범준

BEV-MAE: Bird's Eye View Masked Autoencoders for Point Cloud Pre-training in Autonomous Driving Scenarios

AAAI 2024

Zhiwei Lin1, Yongtao Wang1, Shengxiang Qi2, Nan Dong2, Ming-Hsuan Yang3*

● Related Work

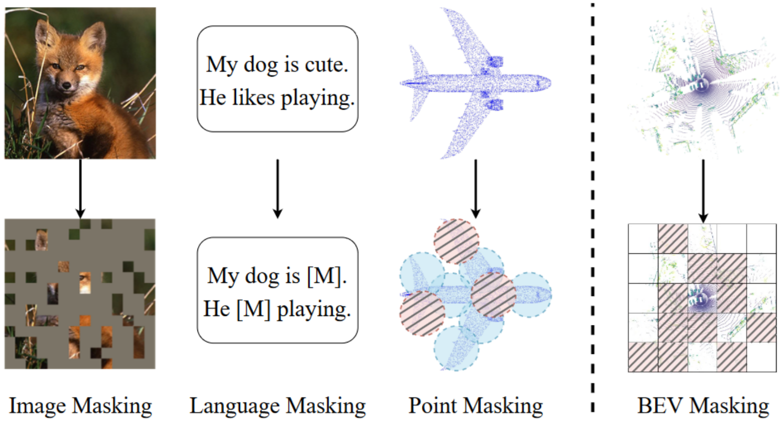


Figure 2: **Illustration of several masking strategies in the masked modeling.** MAE (He et al. 2022) masks non-overlapping image patches. BERT (Devlin et al. 2019) masks words or sentences. Point-MAE (Pang et al. 2022) uses furthest point sampling to create overlapping point patches. Our method (right) projects point clouds into a BEV plane, and masks points in non-overlapping BEV grids.

- Image, Language Masking 에 이어 **points** 분야에서도 masked modeling
  - Target = Input data
  - Generalization ability
  - 기존: voxel-based masking
    - BEV 변환시에 representation gap
    - Voxel 처리 decoder complicated
    - density 문제

→ Points을 **voxel이 아닌 BEV mask**에서..!

김범준

[BEV-MAE: Bird's Eye View Masked Autoencoders for Point Cloud Pre-training in Autonomous Driving Scenarios]

Zhiwei Lin1, Yongtao Wang1, Shengxiang Qi2, Nan Dong2, Ming-Hsuan Yang3*

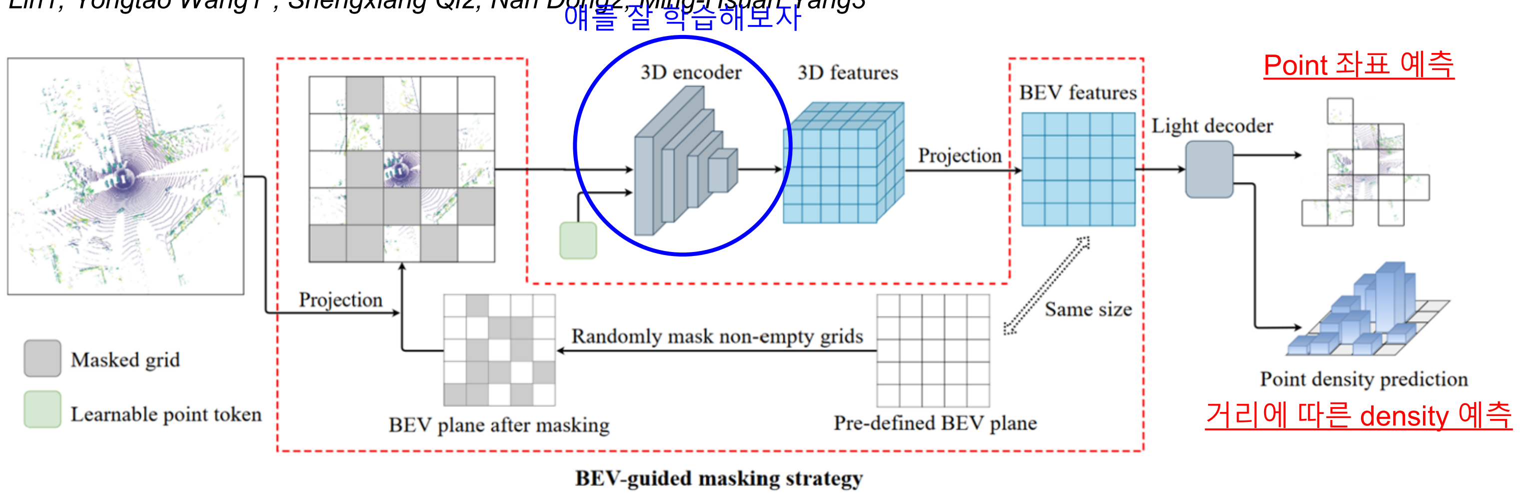


Figure 3: **Overall pipeline of BEV-MAE.** We first mask point clouds with the BEV-guided masking strategy. Then, the masked points are replaced with a shared learnable point token. After extracting BEV features by a 3D encoder from visible points, we send the features to a light decoder to reconstruct masked point clouds and predict the point density of masked grids.

- BEV projection을 통해서 masking할 points 선정
  → Points 는 3D input으로 network에 input
  → Points 좌표 예측 / 공간적 분포인 density 예측

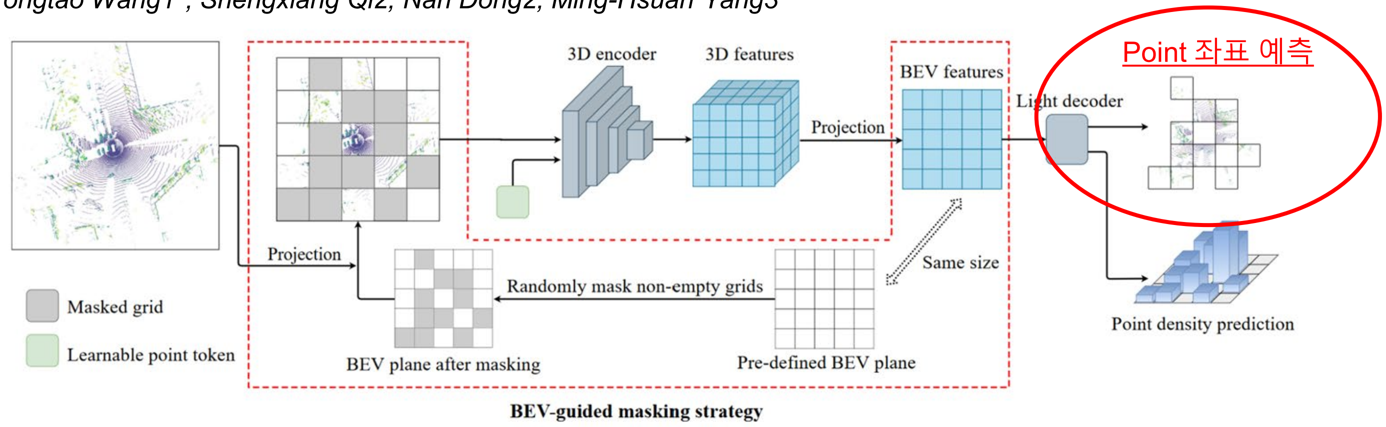
$$g_{i,j} = \{p_k \mid \lfloor x_{p_k}/d \rfloor = i,\ \lfloor y_{p_k}/d \rfloor = j\}$$

point cloud   BEV grid resolution

AAAI 2024

BEV-MAE: Bird's Eye View Masked Autoencoders for Point Cloud Pre-training in Autonomous Driving Scenarios

Zhiwei Lin1, Yongtao Wang1, Shengxiang Qi2, Nan Dong2, Ming-Hsuan Yang3*



- Sperse convolution(SECOND, Centerpoint etc)은 모두 masking으로 인해 receptive field가 제한됨
  - masked 된 point 대신 learnable point token 사용
  - 각 BEV mask 마다 L(=20)개의 point token 고정
  - Chamfer distance metric 이용

$$P_i = \{p_l \mid l = 1, 2, ..., L\}$$

$$\hat{P}_i = \{\hat{p_k} \mid k = 1, 2, ..., N\}$$

Original points

$$\mathcal{L}_c^i = \frac{1}{L}\sum_{p_l \in P_i} \min_{\hat{p_k} \in \hat{P}_i} ||p_l - \hat{p_k}||_2^2 + \frac{1}{N}\sum_{\hat{p_k} \in \hat{P}_i} \min_{p_l \in P_i} ||\hat{p_k} - p_l||_2^2.$$

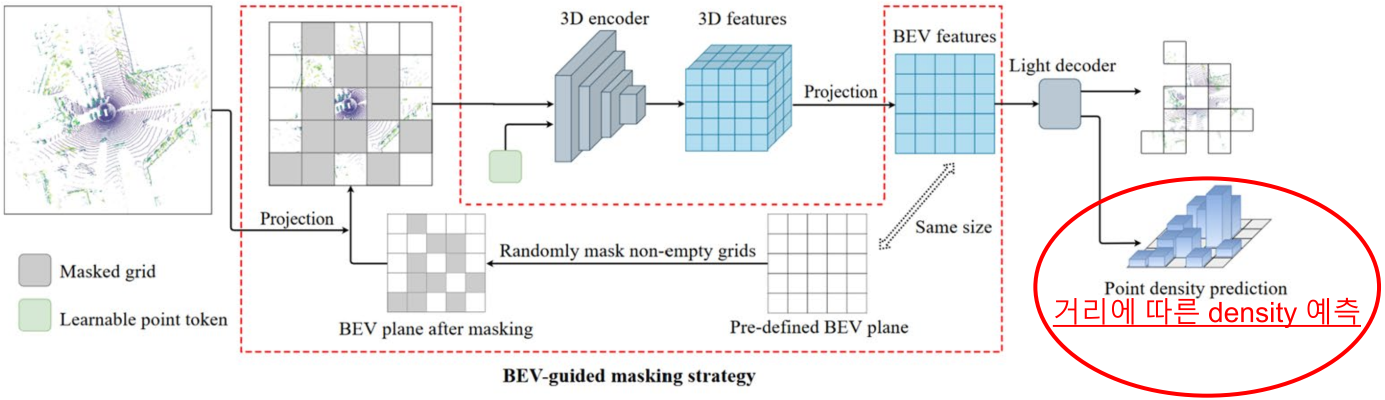
$$\mathcal{L}_c = \frac{1}{n_m}\sum_{i=1}^{n_m} \mathcal{L}_c^i,$$

김범준

AAAI 2024

BEV-MAE: Bird's Eye View Masked Autoencoders for Point Cloud Pre-training in Autonomous Driving Scenarios

Zhiwei Lin1, Yongtao Wang1, Shengxiang Qi2, Nan Dong2, Ming-Hsuan Yang3*



- LiDAR point cloud data는 거리에 따라 sparse 해지는 특성
  - 이를 고려한 feature를 만드는 encoder를 학습
  - 더 나은 localization ability

pred real

$$\mathcal{L}_d^i = Smooth\text{-}\ell_1(\rho_i - \hat{\rho}_i).$$

$$\mathcal{L}_d = \frac{1}{n_m}\sum_{i=1}^{n_m}\mathcal{L}_d^i.$$

김범준

AAAI 2024

BEV-MAE: Bird's Eye View Masked Autoencoders for Point Cloud Pre-training in Autonomous Driving Scenarios

Zhiwei Lin1, Yongtao Wang1, Shengxiang Qi2, Nan Dong2, Ming-Hsuan Yang3*

Pre-traing Method	Epochs	Time	Dataset fraction	L2 (mAP/APH)			
				Overall	Vehicle	Pedestrian	Cyclist
From-scratch	-	-	-	65.60 / 63.21	64.18 / 63.69	65.22 / 59.68	67.41 / 66.25
GCC-3D (Liang et al. 2021)*	40	-	100%	65.29 / 62.79	63.97 / 63.47	64.23 / 58.47	67.68 / 66.44
PointContrast (Xie et al. 2020)	50	54h	100%	$65.88^{+0.28}$ / $63.28^{+0.07}$	63.81 / 63.33	66.67 / 60.51	67.17 / 66.00
DepthContrast (Zhang et al. 2021)	50	56h	100%	$65.84^{+0.24}$ / $63.33^{+0.12}$	64.45 / 63.95	65.61 / 59.86	67.43 / 66.22
Point-M2AE (Zhang et al. 2022)	30	56h	100%	$66.10^{+0.50}$ / $63.59^{+0.38}$	64.26 / 63.77	65.64 / 60.00	68.20 / 67.01
ProposalContrast (Yin et al. 2022)	50	64h	100%	$66.42^{+0.82}$ / $63.85^{+0.64}$	65.03 / 64.53	65.93 / 59.95	68.26 / 67.04
MSP (Jiang et al. 2023)	30	-	100%	- / $64.26^{+1.05}$	- / -	- / -	- / -
GD-MAE† (Yang et al. 2023)	30	60h	100%	$66.98^{+1.38}$ / $64.53^{+1.32}$	65.64 / 64.95	66.39 / 61.12	68.92 / 67.52
BEV-MAE (Ours)	20	5h	20%	$66.70^{+1.10}$ / $64.25^{+1.04}$	64.71 / 64.22	66.21 / 60.59	69.11 / 67.93
BEV-MAE (Ours)	30	38h	100%	**$67.02^{+1.42}$ / $64.55^{+1.34}$**	65.01 / 64.53	66.58 / 60.87	69.46 / 68.25

Table 1: Comparisons between BEV-MAE and state-of-the-art self-supervised learning methods on Waymo validation set. All detectors are fine-tuning with 20% training samples on Waymo following the OpenPCDet configuration. Here, the entry with * denotes the results are from the paper (Liang et al. 2021); the entry with † indicates the results are implemented by the released official code[1]. 'Epochs' indicates the pre-training epochs; 'Dataset fraction' means the data fraction of the Waymo training set used for pre-training; and 'Time' refers to the pre-training time estimated by 8 P40 GPU.

3D object detector, *i.e.*, TransFusion-L.

김범준

AAAI 2024

BEV-MAE: Bird's Eye View Masked Autoencoders for Point Cloud Pre-training in Autonomous Driving Scenarios

Zhiwei Lin1, Yongtao Wang1, Shengxiang Qi2, Nan Dong2, Ming-Hsuan Yang3*

Data amount	Initialization	Overall		Car		Pedestrian		Cyclist	
		L2 mAP	L2 mAPH	L2 mAP	L2 mAPH	L2 mAP	L2 mAPH	L2 mAP	L2 mAPH
5%	Random	44.41	40.34	51.01	50.49	52.74	42.26	29.49	28.27
	PointContrast(Xie et al. 2020)	45.32	41.30	52.12	51.61	53.68	43.22	30.16	29.09
	ProposalContrast(Yin et al. 2022)	46.62	42.58	52.67	52.19	54.31	43.82	32.87	31.72
	MV-JAR (Xu et al. 2023)	50.52	46.68	56.47	56.01	57.65	47.69	37.44	36.33
	BEV-MAE (Ours)	**51.63**	**47.77**	56.35	55.81	58.11	48.37	40.44	39.13
10%	Random	54.31	50.46	54.84	54.37	60.55	50.71	47.55	46.29
	PointContrast(Xie et al. 2020)	53.69	49.94	54.76	54.30	59.75	50.12	46.57	45.39
	ProposalContrast(Yin et al. 2022)	53.89	50.13	55.18	54.71	60.01	50.39	46.48	45.28
	MV-JAR (Xu et al. 2023)	57.44	54.06	58.43	58.00	63.28	54.66	50.63	49.52
	BEV-MAE (Ours)	**58.16**	**54.75**	58.51	57.94	63.83	55.23	52.13	51.07
20%	Random	60.16	56.78	58.79	58.35	65.63	57.04	56.07	54.94
	PointContrast(Xie et al. 2020)	59.35	55.78	58.64	58.18	64.39	55.43	55.02	53.73
	ProposalContrast(Yin et al. 2022)	59.52	55.91	58.69	58.22	64.53	55.45	55.36	54.07
	MV-JAR (Xu et al. 2023)	62.28	59.15	61.88	61.45	66.98	59.02	57.98	57.00
	BEV-MAE (Ours)	**62.88**	**59.97**	61.79	61.37	67.35	59.39	59.51	59.14
50%	Random	66.43	63.36	63.81	63.38	70.78	63.05	64.71	63.66
	PointContrast(Xie et al. 2020)	65.51	62.21	62.66	62.23	69.82	61.53	64.04	62.86
	ProposalContrast(Yin et al. 2022)	65.76	62.49	62.93	62.50	70.09	61.86	64.26	63.11
	MV-JAR (Xu et al. 2023)	66.70	63.69	64.30	63.89	71.14	63.57	64.65	63.63
	BEV-MAE (Ours)	**67.16**	**64.07**	64.33	63.84	71.38	63.61	65.76	64.77
100%	Random	68.50	65.54	64.96	64.56	72.38	64.89	68.17	67.17
	PointContrast(Xie et al. 2020)	68.06	64.84	64.24	63.82	71.92	63.81	68.03	66.89
	ProposalContrast(Yin et al. 2022)	68.17	65.01	64.42	64.00	71.94	63.94	68.16	67.10
	MV-JAR (Xu et al. 2023)	69.16	66.20	65.52	65.12	72.77	65.28	69.19	68.20
	BEV-MAE (Ours)	**69.35**	**66.46**	65.54	65.02	72.84	65.31	69.67	69.05

Table 3: Results about data efficiency on Waymo. The detectors are fine-tuned on various fractions of Waymo training split following MV-JAR (Xu et al. 2023). 'Random' denotes the training-from-scratch baseline.

김범준

BEV-MAE: Bird's Eye View Masked Autoencoders for Point Cloud Pre-training in Autonomous Driving Scenarios

AAAI 2024

Zhiwei Lin1, Yongtao Wang1, Shengxiang Qi2, Nan Dong2, Ming-Hsuan Yang3*

Pre-train \ Fine-tune	nuScenes		Waymo	
	mAP	NDS	L2 mAP	L2 APH
Random init.	48.6	58.4	63.97	61.53
nuScenes	$49.7^{+1.1}$	$58.9^{+0.5}$	$64.79^{+0.82}$	$62.28^{+0.75}$
Waymo	$49.4^{+0.8}$	$58.8^{+0.4}$	$65.13^{+1.16}$	$62.63^{+1.10}$
nuScenes + Waymo	$\mathbf{50.1}^{+1.5}$	$\mathbf{59.1}^{+0.7}$	$\mathbf{65.36}^{+1.39}$	$\mathbf{62.89}^{+1.36}$

Table 4: Results of transfer learning. The contents in the column and row show the datasets for pre-training and fine-tuning, respectively.

ting, *i.e.*, pre-trained on a combined dataset and fine-tuned on the target dataset. The performance of the 3D object de-

김범준

BEV-MAE: Bird's Eye View Masked Autoencoders for Point Cloud Pre-training in Autonomous Driving Scenarios

Zhiwei Lin1, Yongtao Wang1, Shengxiang Qi2, Nan Dong2, Ming-Hsuan Yang3*

Pre-train	Reconstruction target	LT	L2 mAP	L2 APH
None	-	-	65.60	63.21
BEV-MAE	Coord. (w/o norm)	✓	65.66	63.09
	Coord. (w norm)	✓	66.20	63.71
	Density	✓	65.80	63.27
	Number of points	✓	65.32	62.88
	Coord. (w norm) + Density	×	66.49	63.99
	Coord. (w norm) + Density	✓	**66.70**	**64.25**

Table 5: Ablation on main components. 'LT' denotes the shared learnable point token. Each component brings performance improvement for BEV-MAE.

train the 3D encoder on 20% training data of Waymo with BEV-MAE and then evaluate its performance by fine-tuning with CenterPoint on the same 20% training data.

김범준

BEV-MAE: Bird's Eye View Masked Autoencoders for Point Cloud Pre-training in Autonomous Driving Scenarios

Zhiwei Lin1, Yongtao Wang1, Shengxiang Qi2, Nan Dong2, Ming-Hsuan Yang3*

Decoder	L2 mAP	L2 APH	Training cost
One-layer 3×3 Conv	**66.70**	**64.25**	**1×**
Residual Conv block	66.61	64.09	1.2×
Transformer block	65.80	63.26	1.4×

Table 6: Ablation on different deocders. One-layer 3×3 Conv achieves the best results with the least training cost.

Masking strategy	L2 mAP	L2 APH	Memory	Training cost
BEV-guided masking	**66.70**	**64.25**	**4.1G**	**1×**
Voxel masking	66.63	64.16	12.6G	1.4×

Table 7: Ablation on the masking strategy. Pre-training with the BEV-guided masking strategy performs better with less GPU memory consumption and pre-training cost.

Masking ratio	L2 mAP	L2 APH
50%	66.45	64.00
60%	66.62	64.13
70%	**66.70**	**64.25**
80%	66.52	64.06

Table 8: Ablation on masking ratio. The fine-tuning results are less sensitive to the masking ratio.