ICCV 2023

UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation

Haiyang Wang1 Hao Tang1,4* Shaoshuai Shi2† Aoxue Li3 Zhenguo Li3 Bernt Schiele2 Liwei Wang*

- Problem/Objective
  - Bird's eye view representation (3D OD & segmentation)
  - Reduce model size(fusion model) and fast inference time

- Contribution/Key Idea
  - Modality-agnostic Transformer that interact cross-modal
  - Geometry-aware sparse neighborhood relations (2D & 3D)
  - NDS +1.1% & mIoU +12.0%

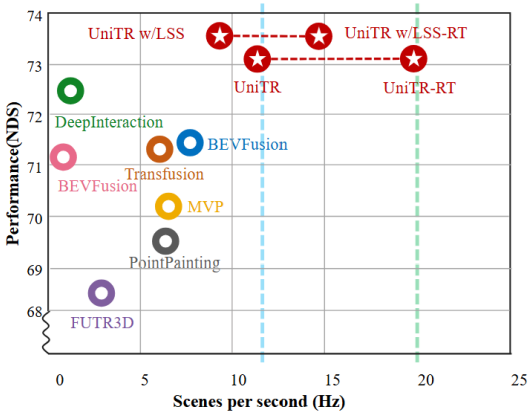


Figure 1. 3D object detection performance (NDS) vs speed (Hz) on nuScenes [3] validation set. Latency is measured on an A100 GPU with AMD EPYC 7513 CPU. Blue and green lines are the operating frequency of the camera and LiDAR in nuScenes.

김범준

UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation

Haiyang Wang1 Hao Tang1,4* Shaoshuai Shi2† Aoxue Li3 Zhenguo Li3 Bernt Schiele2 Liwei Wang*

- **Remaining Method**

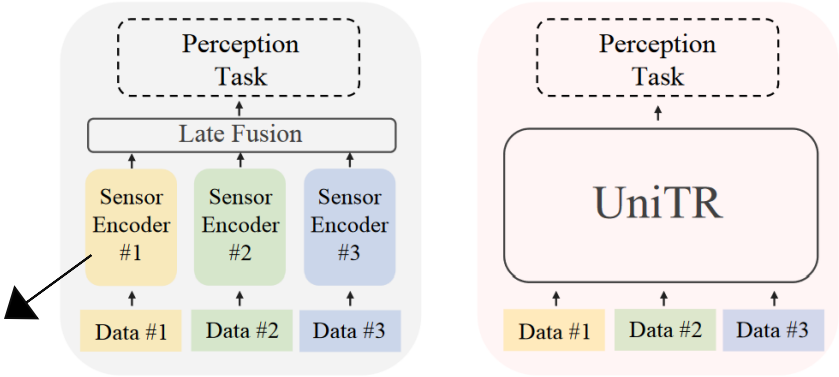


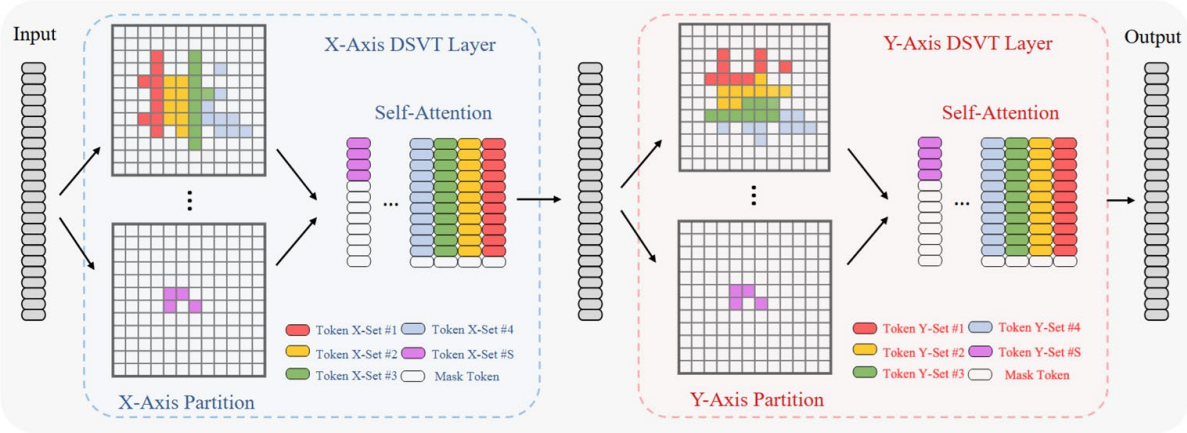
Figure 2. Comparison between sequential modality-specific encoders and our proposed UniTR, which processes various modalities in parallel with a single model and shared parameters.

- 기존의 모델 → 각각의 sensor data를 각자의 encoding layer를 통과하여 Fusion
- UniTR → 데이터를 fusion 하는 early fusion 방식 채택
  - but) 다른점은 modality에 관계없는 (modality-agnostic transformer 제안)

UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation

Haiyang Wang1 Hao Tang1,4* Shaoshuai Shi2† Aoxue Li3 Zhenguo Li3 Bernt Schiele2 Liwei Wang*

- **DSVT (Dynamic Sparse Window attention) - CVPR 2023**

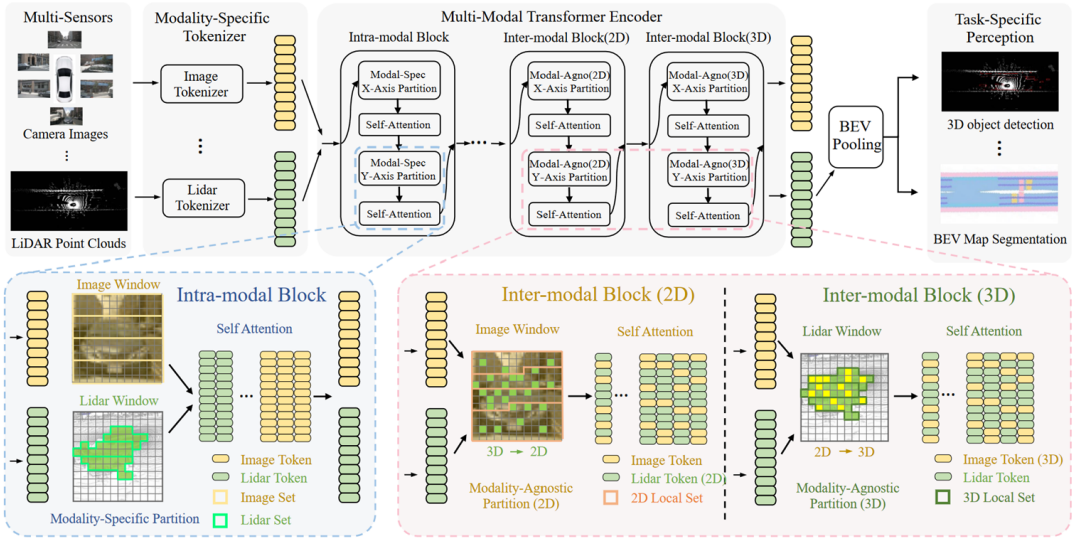


- VFE (voxel feature encoder) 통과한 feature들을 token → token이 비어있는 것이 LiDAR의 특징
  - Window 내부 X-axis 방향으로 non-empty token n개를 concat하여 같은 색으로 지정
  - Self-Attention
  - Y-axis 반복
  - 이것을 여러 Window size에서 반복 (L x W x H size)

UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation

Haiyang Wang1 Hao Tang1,4* Shaoshuai Shi2† Aoxue Li3 Zhenguo Li3 Bernt Schiele2 Liwei Wang*

- **Method**

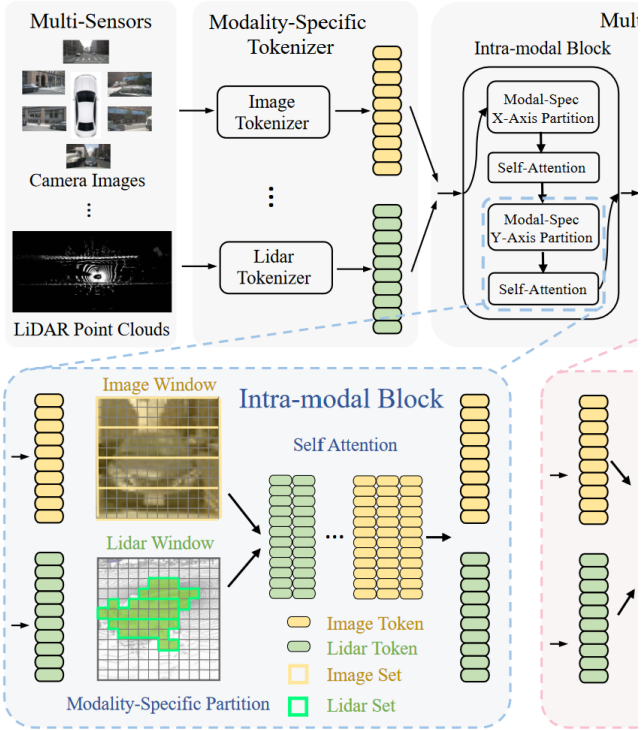


- 기존 모델들 처럼 각 모달리티에 대한 별도의 처리 네트워크 x
- Image Tokenize & LiDAR Tokenizer만 개별로 통과 후 Block들 통과

UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation

Haiyang Wang1 Hao Tang1,4* Shaoshuai Shi2† Aoxue Li3 Zhenguo Li3 Bernt Schiele2 Liwei Wang*

- **Method**



- 첫번째 모듈에서는 sensor의 raw-data를 tokenization하는 과정
  - LiDAR data → voxelization & dynamic 하게 encoding하는 VFE 과정
  - Image data → ViT에서 사용하는 image patch tokenizer

$$\mathcal{T}^P = \{t_i^P | t_i^P = [(x_i^P, y_i^P, z_i^P); f_i^P]\}_{i=1}^N,$$
$$\mathcal{T}^I = \{t_i^I | t_i^I = [(x_i^I, y_i^I, b_i^I); f_i^I]\}_{i=1}^M,$$

- Image: 6x256x704x3 (b,h,w,c) → m x c 개의 patch
- pointcloud : voxel size (0.3, 0.3, 8.0)m → n x c 개의 patch with Dynamic Set partition

$$\{\mathcal{Q}_n^P\}_{n=0}^{\mathcal{N}} = \mathrm{DSP}(\mathcal{T}^P, L^P \times W^P \times H^P, \tau),$$
$$\{\mathcal{Q}_m^I\}_{m=0}^{\mathcal{M}} = \mathrm{DSP}(\mathcal{T}^I, L^I \times W^I \times 1, \tau),$$

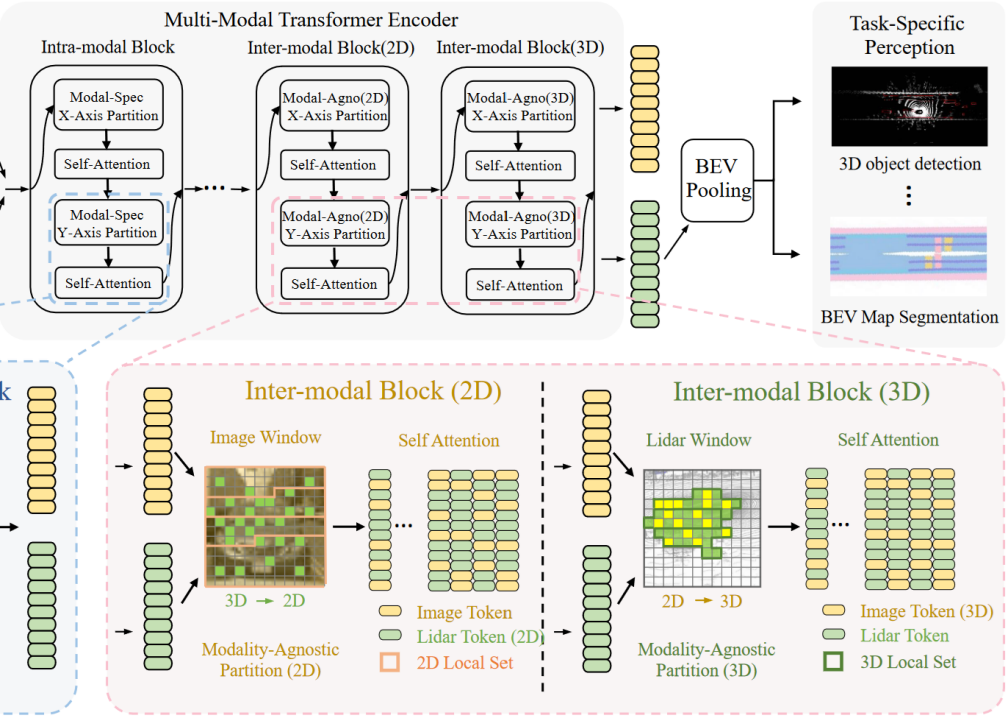
- Intra-modal Block 에서 self-attention 하는 과정 (각 센서별로)
  - Image는 Swin-Transformer 처럼
  - LiDAR는 DSVT 처럼

$$\mathcal{F}_l, \mathcal{C}_l = \mathrm{INDEX}([\mathcal{T}_{l-1}^P, \mathcal{T}_{l-1}^I], [\{\mathcal{Q}_n^P\}_{n=0}^{\mathcal{N}}, \{\mathcal{Q}_m^I\}_{m=0}^{\mathcal{M}}]),$$
$$\widetilde{\mathcal{T}}_l^P, \widetilde{\mathcal{T}}_l^I = \mathrm{MHSA}(\mathcal{F}_l, \mathrm{PE}(\mathcal{C}_l)),$$

UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation

Haiyang Wang1 Hao Tang1,4* Shaoshuai Shi2† Aoxue Li3 Zhenguo Li3 Bernt Schiele2 Liwei Wang*

- **Method**



- Multi-sensor의 경우 view-discrepant 문제를 해결하기 위함
  - 하나의 model에서 여러 sensor의 feature를 뽑을 수 있어서 굉장한 시간 이점이라고 주장

- LiDAR token을 calibration parameter를 이용해서 투영 (L→C)
  - image token과의 unify는 동일 채널을 가지기에 sum or interpolation
  - X-axis / Y-axis에서 self-attention

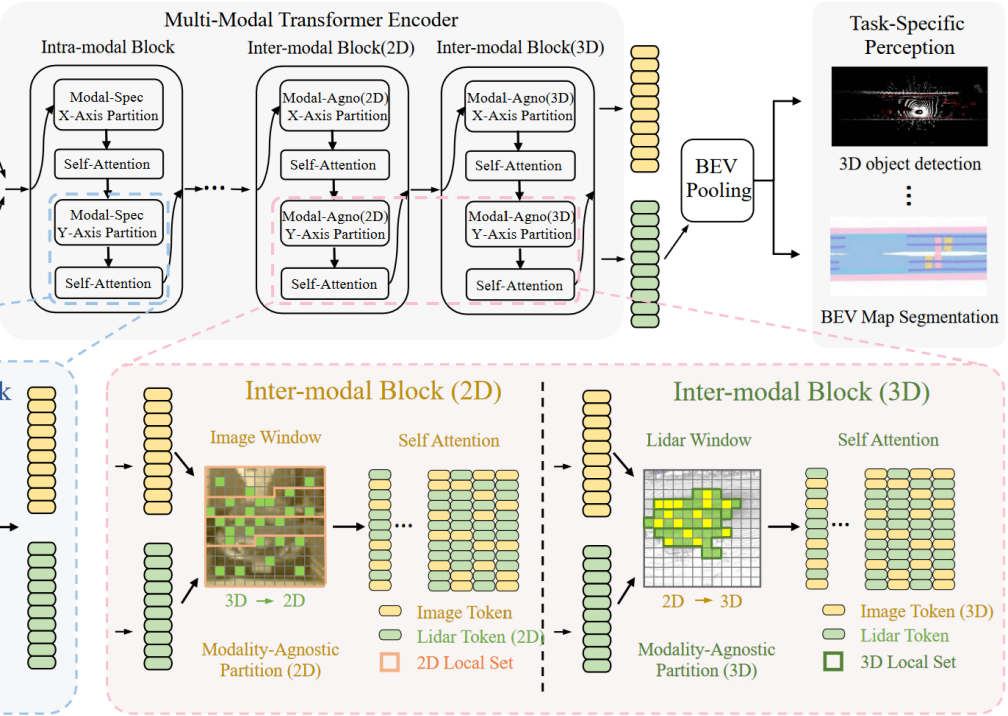
$$\mathcal{T}^P \rightarrow \mathcal{T}^P_{2D} : (x^P, y^P, z^P) \rightarrow (x^P_{2D}, y^P_{2D}, b^P_{2D}),$$

$$\{\mathcal{Q}^{2D}_m\}^{\widetilde{\mathcal{M}}}_{m=0} = \text{DSP}([\mathcal{T}^P_{2D}, \mathcal{T}^I], L^I \times W^I \times 1, \tau),$$

UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation

Haiyang Wang1 Hao Tang1,4* Shaoshuai Shi2† Aoxue Li3 Zhenguo Li3 Bernt Schiele2 Liwei Wang*

- **Method**



- Image token을 LiDAR token에 투영 (C→L)
  - depth estimation (오차 발생, ill-posed problem)
    → MVP [1] 아이디어 차용
    : L→C 매칭쌍 이용하여 depth retrieval,

매칭 안된거는

가장 가까운 depth를 retrieve

  - X-axis / Y-axis에서 self-attention

$$\{v_i^{\text{nearest}} | (x_i, y_i); b_i; d_i\}_{i=0}^{M} = \text{Nearest}(\mathcal{V}^I, \mathcal{T}^I),$$

$$\mathcal{T}^I \rightarrow \mathcal{T}_{3D}^I : (x^I, y^I, b^I) \rightarrow (x_{3D}^I, y_{3D}^I, z_{3D}^I),$$
$$f^I \rightarrow f^I + \text{MLP}(||v^{\text{nearest}}(x, y) - (x^I, y^I)||).$$

$$\{\mathcal{Q}_n^{3D}\}_{n=0}^{\tilde{\mathcal{N}}} = \text{DSP}([\mathcal{T}^P, \mathcal{T}_{3D}^I], L^P \times W^P \times H^P, \tau),$$

[1] Yin, Tianwei, Xingyi Zhou, and Philipp Krähenbühl. "Multimodal virtual point 3d detection." Advances in Neural Information Processing Systems 34 (2021): 16494-16507.

UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation

Haiyang Wang1 Hao Tang1,4* Shaoshuai Shi2† Aoxue Li3 Zhenguo Li3 Bernt Schiele2 Liwei Wang*

- **Experiment**

Modality	Intra-B	Inter-B (2D)	Inter-B (3D)	BEVLSS	NDS	mAP
L	4	0	0		70.5	65.9
C+L	3	0	1		72.0	68.5
C+L	3	1	0		72.5	69.0
C+L	2	1	1		72.9	69.8
C+L	1	2	1		73.1	70.0
C+L	1	2	1	✓	73.3	70.5

Table 10. The number of the intra- and inter-modal blocks on the ablation of 2D & 3D fusion. Camera (C), LiDAR (L).

- 각 블럭 개수에 대한 ablation study

UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation

Haiyang Wang1 Hao Tang1,4* Shaoshuai Shi2† Aoxue Li3 Zhenguo Li3 Bernt Schiele2 Liwei Wang*

- **Experiment**

Methods	Present at	Modality	NDS (*val*)	mAP (*val*)	NDS (*test*)	mAP (*test*)	Latency (*ms*)
BEVFormer [34]	ECCV'22	C	-	-	56.9	48.1	–
BEVDepth [32]	AAAI'23	C	-	-	60.0	50.3	–
BEVFormer v2 [69]	CVPR'23	C	-	-	63.4	55.6	–
SECOND [68]	Sensors'18	L	63.0	52.6	63.3	52.8	53.2
PointPillars [27]	CVPR'19	L	61.3	52.3	–	–	28.1
CenterPoint [74]	CVPR'21	L	66.8	59.6	67.3	60.3	62.7
PointPainting [56]	CVPR'20	C+L	69.6	65.8	-	-	151.8
PointAugmenting [57]	CVPR'21	C+L	-	-	71.0†	66.8†	188.4
MVP [75]	NeurIPS'21	C+L	70.0	66.1	70.5	66.4	148.1
FusionPainting [67]	ITSC'21	C+L	70.7	66.5	71.6	68.1	-
FUTR3D [5]	ArXiv'22	C+L	68.3	64.5	-	-	302.6
AutoAlign [8]	IJCAI'22	C+L	71.1	66.6	-	-	-
TransFusion [1]	CVPR'22	C+L	71.3	67.5	71.6	68.9	164.6
AutoAlignV2 [7]	ECCV'22	C+L	71.2	67.1	72.4	68.4	207.0
UVTR [30]	NeurIPS'22	C+L	70.2	65.4	71.1	67.1	-
BEVFusion (PKU) [35]	NeurIPS'22	C+L	71.0	67.9	71.8	69.2	1231.0
DeepInteraction [71]	NeurIPS'22	C+L	72.6	69.9	73.4	**70.8**	541.1
BEVFusion (MIT) [39]	ICRA'23	C+L	71.4	68.5	72.9	70.2	130.5
UniTR (Ours)	ICCV'23	C+L	**73.1**	**70.0**	**74.1**	70.5	**88.7 (50.2‡)**
UniTR w/ LSS (Ours)	ICCV'23	C+L	**73.3**	**70.5**	**74.5**	**70.9**	**107.5 (69.1‡)**

Table 1. Performance of 3D detection on nuScenes (val and test) dataset [3]. Notion of modality: Camera (C), LiDAR (L). †: with test-time augmentation. ‡: deployed by TensorRT). We highlight the top-2 entries with **bold** font in each column.

- 매우 빠른 Inference & +1.9% 정도의 성능 향상 in 3D object detection

UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation

Haiyang Wang1 Hao Tang1,4* Shaoshuai Shi2† Aoxue Li3 Zhenguo Li3 Bernt Schiele2 Liwei Wang*

- **Experiment**

Approach	Clean		Missing F		Preserve F		Stuck	
	mAP	NDS	mAP	NDS	mAP	NDS	mAP	NDS
DETR3D* [64]	34.9	43.4	25.8	39.2	3.3	20.5	17.3	32.3
PointAugmenting [57]	46.9	55.6	42.4	53.0	31.6	46.5	42.1	52.8
MVX-Net [54]	61.0	66.1	47.8	59.4	17.5	41.7	48.3	58.8
TransFusion [1]	66.9	70.9	65.3	70.1	64.4	69.3	65.9	70.2
BEVFusion [35]	67.9	71.0	65.9	70.7	65.1	69.9	66.2	70.3
UniTR (Ours)	70.5	73.3	68.5	72.4	66.5	71.2	68.1	71.8

Table 5. Results on robustness setting of camera failure cases. F denotes the front camera, and * means camera-only inputs. All the experiments are on nuScenes validation set.

Aug	Metrics	LiDAR	BEVFusion[35]	Ours
	mAP	31.3	40.2 (+8.9)	38.3(+7.0)
	NDS	50.7	54.3 (+3.6))	55.6(+4.9)
✓	mAP	-	54.0(+22.7)	60.2(+28.9)
✓	NDS	-	61.6(+10.9)	66.0(+15.3)

Table 6. Results on robustness setting of object failure cases. We refer readers to [35] for more details.

- 일부로 만든 failure case에서도 높은 성능

[UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Representation]

Haiyang Wang1 Hao Tang1,4* Shaoshuai Shi2† Aoxue Li3 Zhenguo Li3 Bernt Schiele2 Liwei Wang*

- **Experiment**

Modality	Camera	Lidar	Serial	Parallel
C+L(1-beam)	36.2	42.0	57.6	**59.5**
C+L(4-beam)	36.2	62.1	67.3	**68.5**
C+L(16-beam)	36.2	69.1	71.6	**72.2**
C+L(32-beam)	36.2	70.5	73.2	**73.3**

Table 7. Ablation of low beam setting with NDS evaluation metric.

Block config	NDS	mAP
3D → 2D	73.0	70.0
2D → 3D	**73.3**	**70.5**
Inter → Intra	72.9	69.8
Intra → Inter	**73.3**	**70.5**

Table 8. Results of different block configurations.

- Table 7: low beam LiDAR에서도 잘 작동한다.
- Table 8: Block들의 순서를 바꿨을때 현재의 순서가 가장 잘 작동한다.