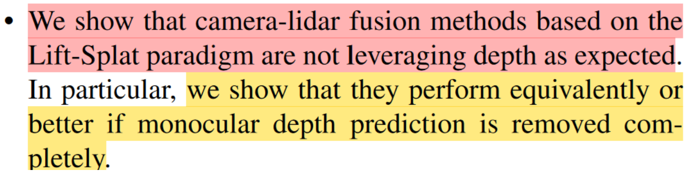
Lift-Attend-Splat: Bird's-eye-view camera-lidar fusion using transformers

James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller

CVPR 2024

- Problem/Objective
  - 3D object detection
  - Camera depth estimation의 불필요성(?) 분석



- We show that camera-lidar fusion methods based on the Lift-Splat paradigm are not leveraging depth as expected. In particular, we show that they perform equivalently or better if monocular depth prediction is removed completely.

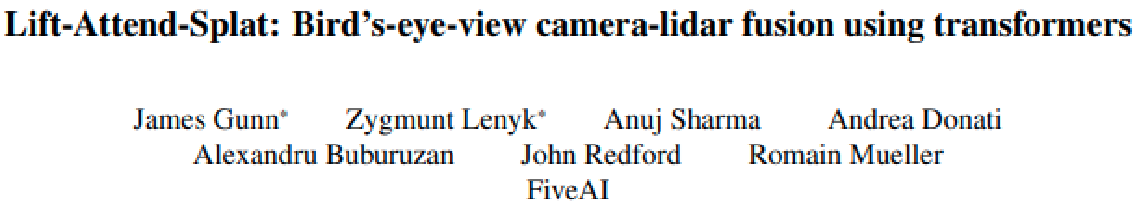
- Contribution/Key Idea
  - LiDAR가 존재하는 fusion에서 LSS의 활용도가 낮다고 주장
  - Attention을 이용한 새로운 L+C fusion 모듈 제시
  - SOTA 는 x, equal or better

김범준

Lift-Attend-Splat: Bird’s-eye-view camera-lidar fusion using transformers

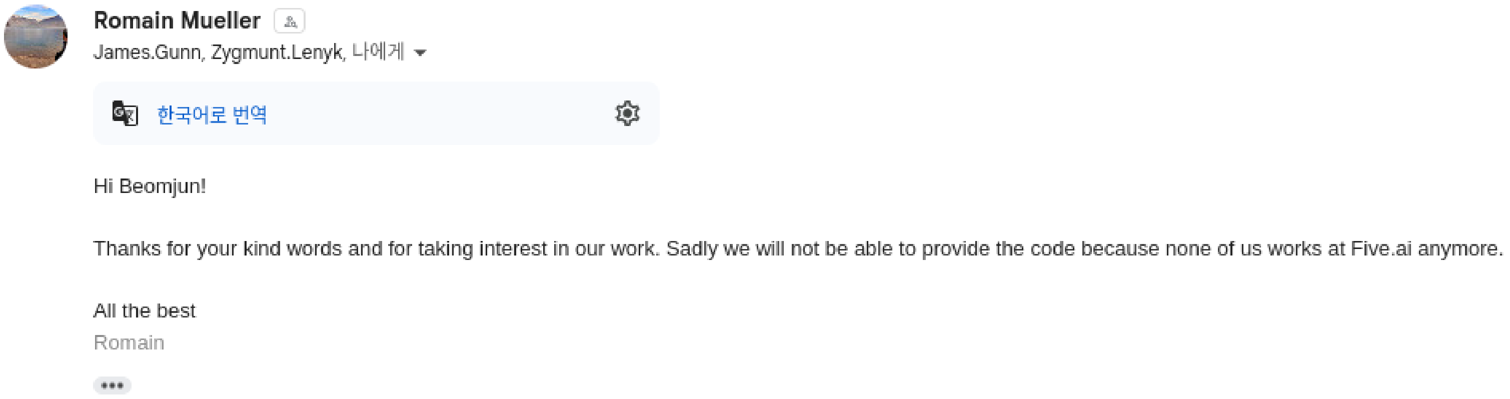
CVPR 2024

James Gunn, *Zygmunt Lenyk*, *Anuj Sharma*, *Andrea Donati*, *Alexandru Buburuzan*, *John Redford*, *Romain Mueller*



Lift-Attend-Splat: Bird’s-eye-view camera-lidar fusion using transformers

James Gunn* Zygmunt Lenyk* Anuj Sharma Andrea Donati
Alexandru Buburuzan John Redford Romain Mueller
FiveAI



Romain Mueller

James.Gunn, Zygmunt.Lenyk, 나에게 ▾

Hi Beomjun!

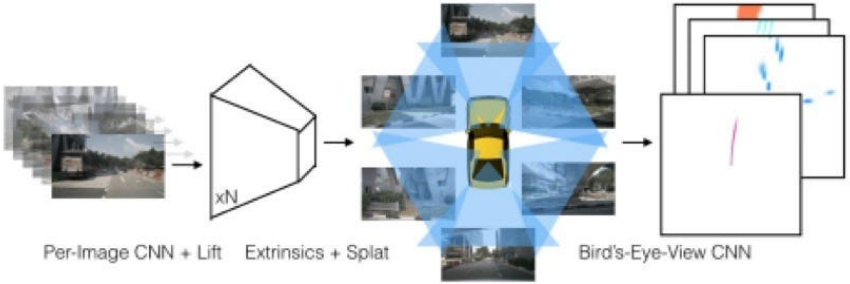
Thanks for your kind words and for taking interest in our work. Sadly we will not be able to provide the code because none of us works at Five.ai anymore.

All the best
Romain

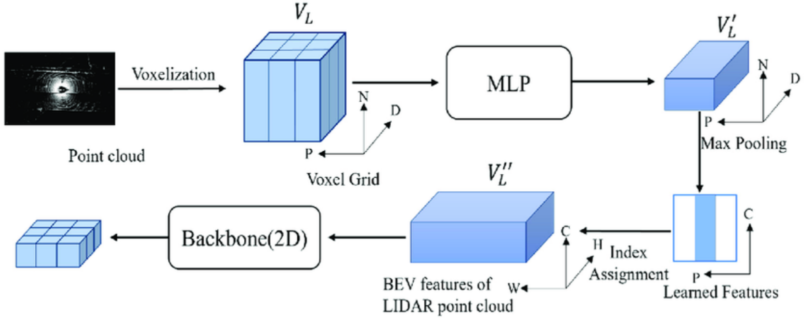
CVPR 2024

Lift-Attend-Splat: Bird’s-eye-view camera-lidar fusion using transformers

James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller



Camera to BEV



LiDAR to BEV

- BEV로의 변환 과정
  - Camera : depth estimation → 3d frustum → (flatten) → BEV 변환
  - LiDAR : Voxelization → Pillar → (flatten) → BEV 변환

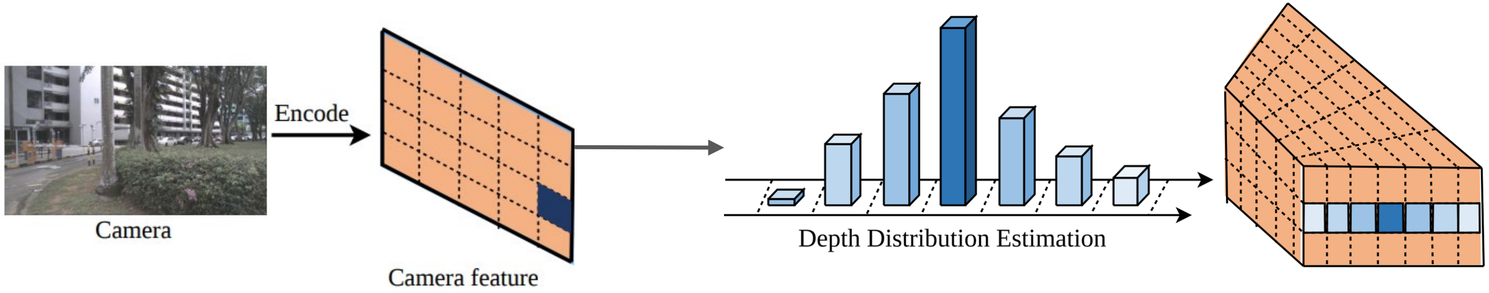
- L+C 모듈은 C only에 비해 depth estimation에 대한 의존도 ⬇
  - 이미 LiDAR에 geometric 정보가 있기 때문..!
  - 더군다나 이 논문의 task인 3d object detection에서는 LiDAR 네트워크가 더 중요

김범준

Lift-Attend-Splat: Bird's-eye-view camera-lidar fusion using transformers

James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller

CVPR 2024



Lift - Splat 네트워크

- 기존 LSS의 Lift - Splat 네트워크

$$F^{\text{cam}} \in \mathbb{R}^{C_c \times H \times W} \rightarrow \begin{array}{l} F'^{\text{cam}} \in \mathbb{R}^{C'_c \times H \times W} \\ D \in \mathbb{R}^{N_D \times H \times W} \end{array}$$

$$\downarrow$$

$$\text{Proj}_{\text{Lift-Splat}} = \text{Splat}\left(F'^{\text{cam}} \otimes D\right)$$

김범준

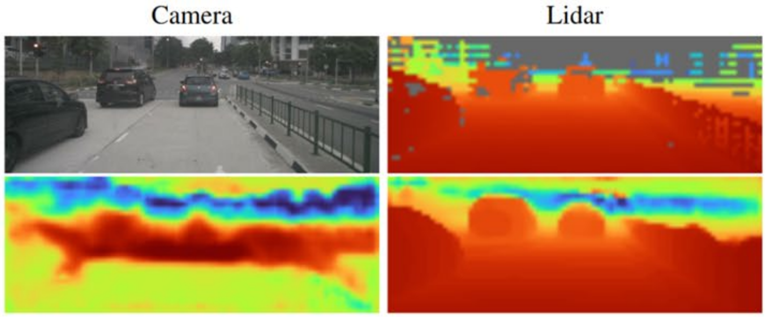
Lift-Attend-Splat: Bird's-eye-view camera-lidar fusion using transformers

James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller

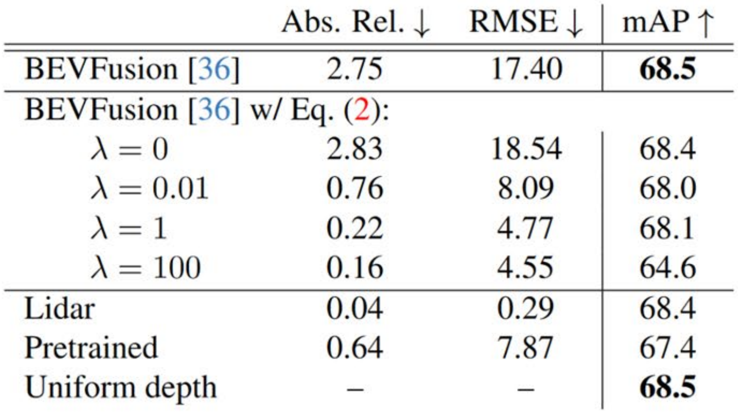
CVPR 2024

$$D \in \mathbb{R}^{N_D \times H \times W} \longleftrightarrow$$ LiDAR depth supervise

$$L_{\text{total}} = L_{\text{sup}} + \lambda L_{\text{depth}}$$



BEVFusion [36] BEVFusion [36] w/ $\lambda = 1$



	Abs. Rel. ↓	RMSE ↓	mAP ↑
BEVFusion [36]	2.75	17.40	**68.5**
BEVFusion [36] w/ Eq. (2):			
$\lambda = 0$	2.83	18.54	68.4
$\lambda = 0.01$	0.76	8.09	68.0
$\lambda = 1$	0.22	4.77	68.1
$\lambda = 100$	0.16	4.55	64.6
Lidar	0.04	0.29	68.4
Pretrained	0.64	7.87	67.4
Uniform depth	–	–	**68.5**

$$\text{Proj}_{\text{no-depth}} = \text{Splat}\left(F'^{\text{cam}} \otimes 1\right)$$

CVPR 2024

Lift-Attend-Splat: Bird's-eye-view camera-lidar fusion using transformers

James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller

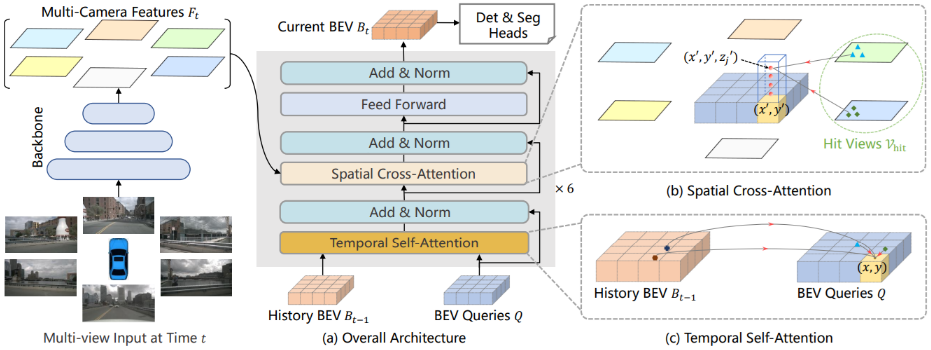


Figure 2: **Overall architecture of BEVFormer.** (a) The encoder layer of BEVFormer contains

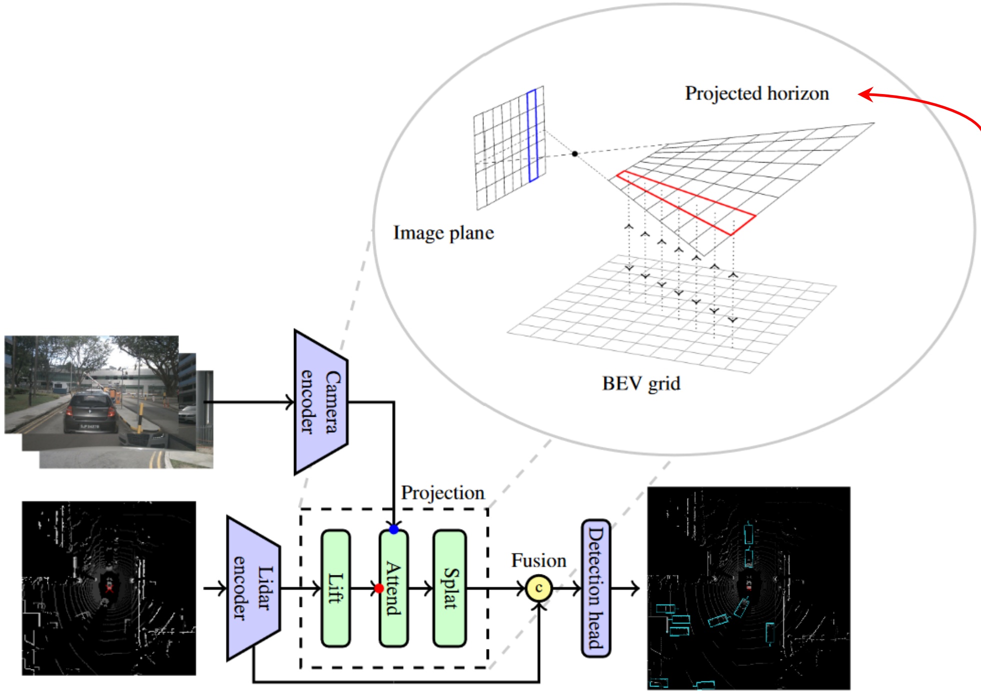
BEVFormer (ECCV 2022)

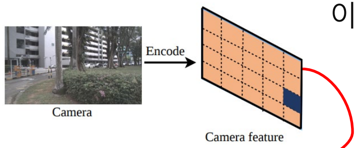
- Attention 네트워크를 활용하고자 기존의 BEV & attention 메커니즘 분석
  - BEV feature 전체 사용 → too much computation /
  - **Camera의 Column과 LiDAR의 Polar ray '만'을 활용해보자 !!**

CVPR 2024

Lift-Attend-Splat: Bird's-eye-view camera-lidar fusion using transformers

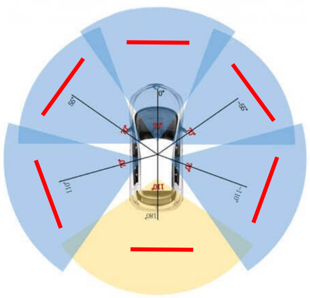
James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller





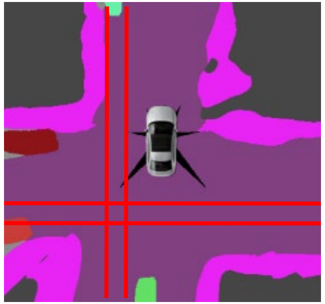
이미지의 센터를 통과하는 horizontal plane 생성

$$\mathbf{Cx} \sim (u, h/2, 1),$$



PV feature

vs



BEV feature

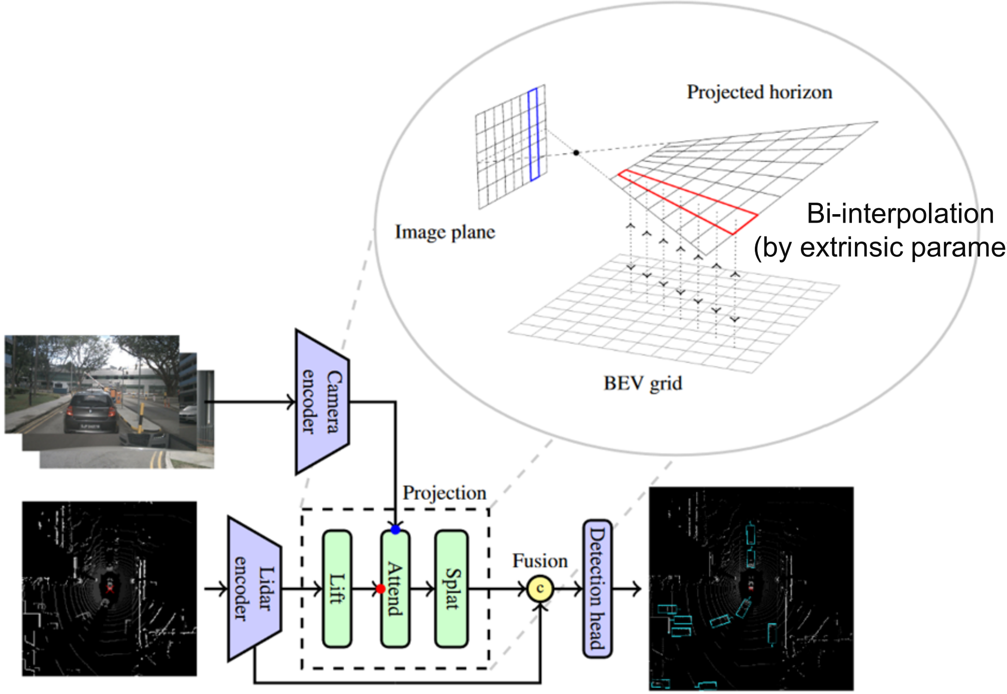
김범준

CVPR 2024

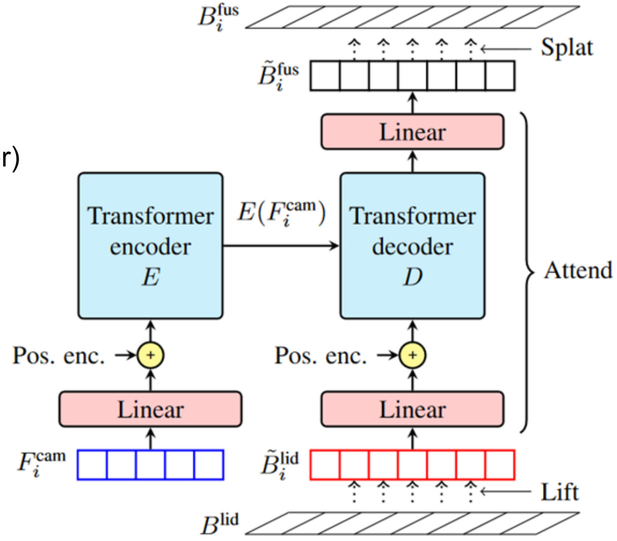
Lift-Attend-Splat: Bird's-eye-view camera-lidar fusion using transformers

James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller

$$B_i^{\text{fus}} = \text{Splat}_i\left(D\left(\text{Lift}_i\left(B^{\text{lid}}\right), E\left(F_i^{\text{cam}}\right)\right)\right),$$



Bi-interpolation
(by extrinsic parameter)



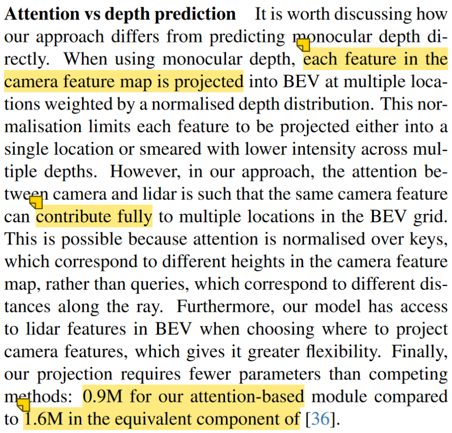
- 카메라 feature가 LiDAR feature에 입히는 과정

Lift-Attend-Splat: Bird’s-eye-view camera-lidar fusion using transformers

James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller

CVPR 2024

- Depth prediction(LSS) VS Attention(본인들)
  - depth prediction은 부정확한 depth로 feature들이 여기저기, But 본인들의 방법은 이 과정을 생략하고 attention으로 **fully contribute**
  - 파라미터 수 줄이면서 동등한 성능 **(1.6M vs 0.9M)**



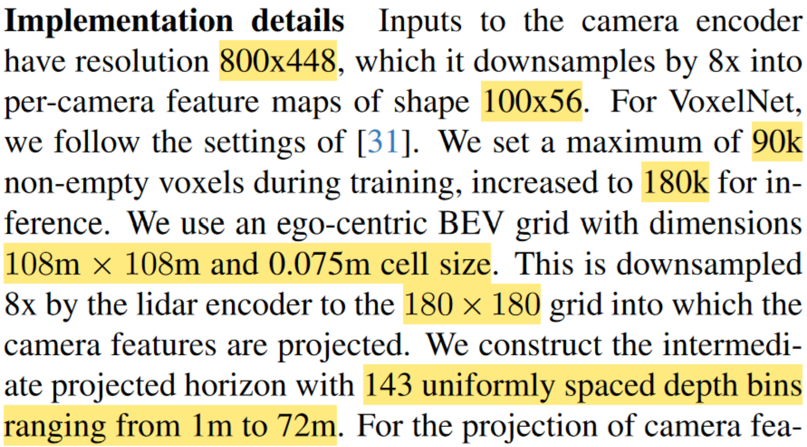
Attention vs depth prediction It is worth discussing how our approach differs from predicting monocular depth directly. When using monocular depth, each feature in the camera feature map is projected into BEV at multiple locations weighted by a normalised depth distribution. This normalisation limits each feature to be projected either into a single location or smeared with lower intensity across multiple depths. However, in our approach, the attention between camera and lidar is such that the same camera feature can contribute fully to multiple locations in the BEV grid. This is possible because attention is normalised over keys, which correspond to different heights in the camera feature map, rather than queries, which correspond to different distances along the ray. Furthermore, our model has access to lidar features in BEV when choosing where to project camera features, which gives it greater flexibility. Finally, our projection requires fewer parameters than competing methods: 0.9M for our attention-based module compared to 1.6M in the equivalent component of [36].

김범준

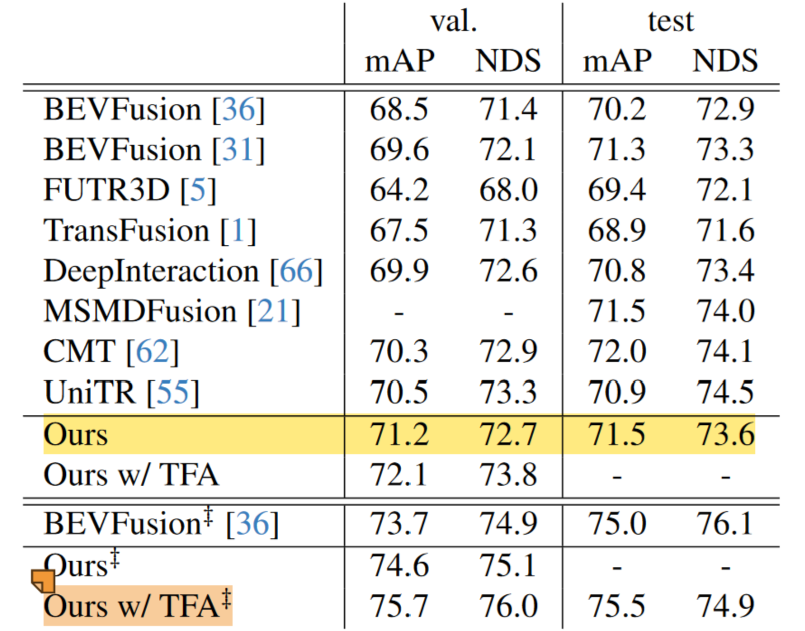
CVPR 2024

Lift-Attend-Splat: Bird's-eye-view camera-lidar fusion using transformers

James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller



Implementation details Inputs to the camera encoder have resolution 800x448, which it downsamples by 8x into per-camera feature maps of shape 100x56. For VoxelNet, we follow the settings of [31]. We set a maximum of 90k non-empty voxels during training, increased to 180k for inference. We use an ego-centric BEV grid with dimensions 108m × 108m and 0.075m cell size. This is downsampled 8x by the lidar encoder to the 180 × 180 grid into which the camera features are projected. We construct the intermediate projected horizon with 143 uniformly spaced depth bins ranging from 1m to 72m. For the projection of camera fea-



	val.		test	
	mAP	NDS	mAP	NDS
BEVFusion [36]	68.5	71.4	70.2	72.9
BEVFusion [31]	69.6	72.1	71.3	73.3
FUTR3D [5]	64.2	68.0	69.4	72.1
TransFusion [1]	67.5	71.3	68.9	71.6
DeepInteraction [66]	69.9	72.6	70.8	73.4
MSMDFusion [21]	-	-	71.5	74.0
CMT [62]	70.3	72.9	72.0	74.1
UniTR [55]	70.5	73.3	70.9	74.5
Ours	71.2	72.7	71.5	73.6
Ours w/ TFA	72.1	73.8	-	-
BEVFusion‡ [36]	73.7	74.9	75.0	76.1
Ours‡	74.6	75.1	-	-
Ours w/ TFA‡	75.7	76.0	75.5	74.9

Table 1. Object detection performance on the validation and test splits of the nuScenes dataset. TFA: Temporal Feature Aggregation. ‡ denotes test-time augmentations and model ensembling.

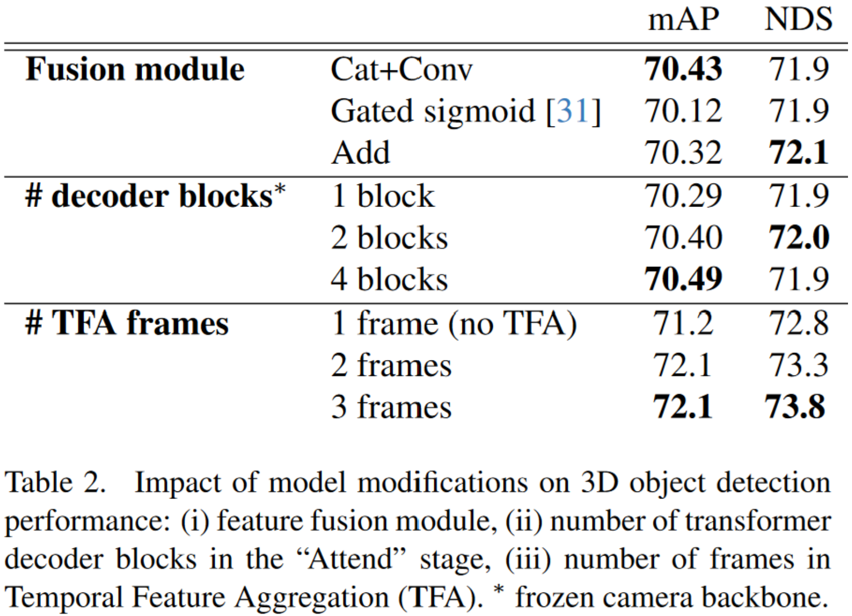
TTA: mirror + rotation 증강을 조합하여 적용.각각의 셀 해상도(0.05m, 0.075m, 0.10m)에서 TTA 적용.
Model Ensemble: TTA를 개별 해상도에서 적용한 뒤, Weighted Boxes Fusion를 사용, 결과 바운딩 박스들을 병합.

김범준

CVPR 2024

Lift-Attend-Splat: Bird's-eye-view camera-lidar fusion using transformers

James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller



		mAP	NDS
Fusion module	Cat+Conv	**70.43**	71.9
	Gated sigmoid [31]	70.12	71.9
	Add	70.32	**72.1**
# decoder blocks*	1 block	70.29	71.9
	2 blocks	70.40	**72.0**
	4 blocks	**70.49**	71.9
# TFA frames	1 frame (no TFA)	71.2	72.8
	2 frames	72.1	73.3
	3 frames	**72.1**	**73.8**

Table 2. Impact of model modifications on 3D object detection performance: (i) feature fusion module, (ii) number of transformer decoder blocks in the "Attend" stage, (iii) number of frames in Temporal Feature Aggregation (TFA). * frozen camera backbone.

- 과거 정보 사용시 성능 향상의 가능성을 열어둠

김범준

CVPR 2024

Lift-Attend-Splat: Bird’s-eye-view camera-lidar fusion using transformers

James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller

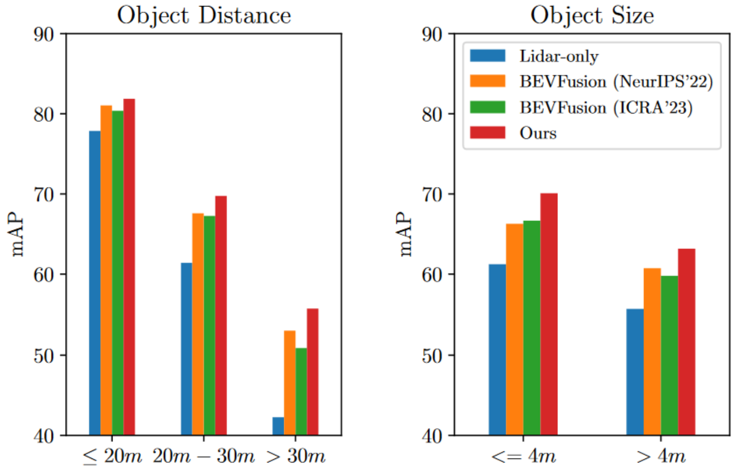


Figure 3. Object detection performance measured using mAP for objects at different distances from the ego and of different sizes. Our model consistently outperforms baselines based on Lift-Splat, especially at large distances and for small objects.

- 개선의 대부분은 **먼 거리** 개체 → 기존 depth estimation으로는 먼거리 객체 정확도 낮았음을 의미
- **작은 객체**에서 성능 개선 → LiDAR point 적어도, attention 모듈을 통해 카메라 feature를 잘 활용하여 입혔음을 의미

Lift-Attend-Splat: Bird's-eye-view camera-lidar fusion using transformers

James Gunn, Zygmunt Lenyk, Anuj Sharma, Andrea Donati, Alexandru Buburuzan, John Redford, Romain Mueller

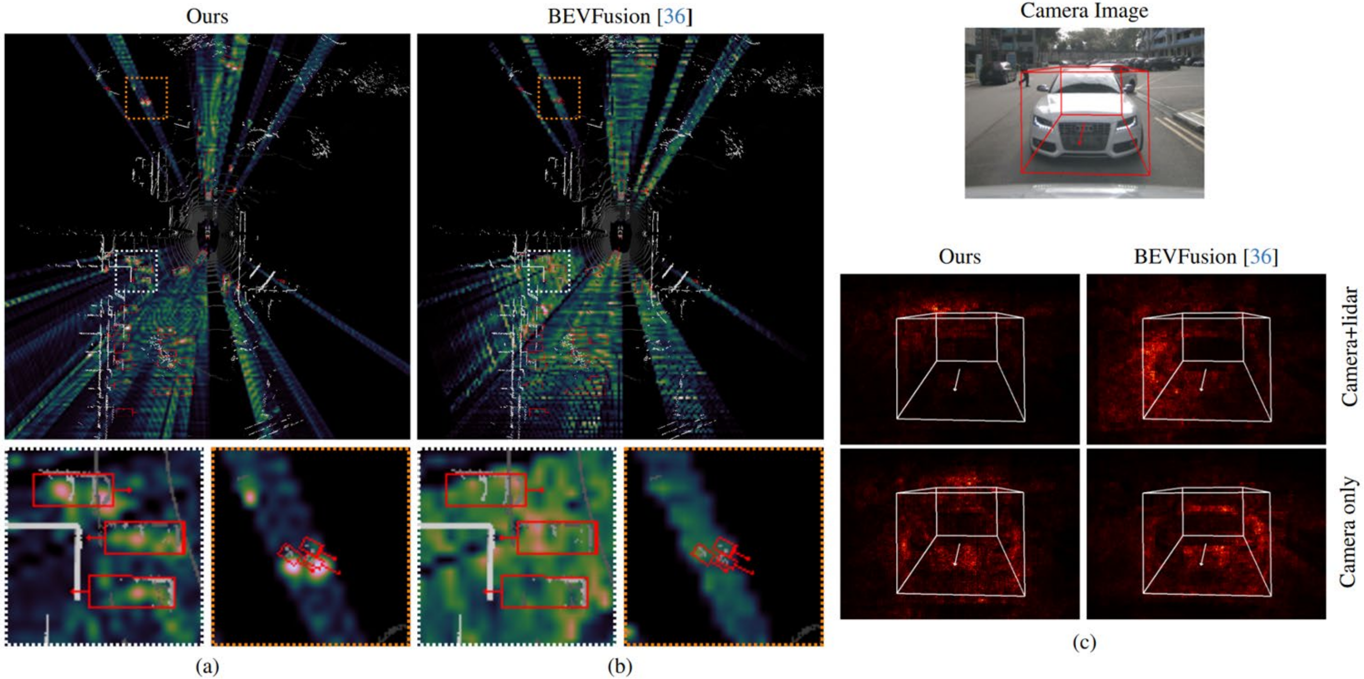


Figure 4. **(a, b)** Visualisation of where camera features of ground-truth objects are projected onto the BEV grid for our method compared to BEVFusion [36]. We observe that our method is able to place camera features around objects more narrowly than BEVFusion, which is based on monocular depth estimation. **(c)** Comparison of saliency maps, cropped to aid visualisation, given the camera image (top) for models trained with camera-lidar (middle) and camera only (bottom). When trained with both camera and lidar, our model selects camera features in an area that is different than when trained with camera only, while [36] behaves similarly in both settings.