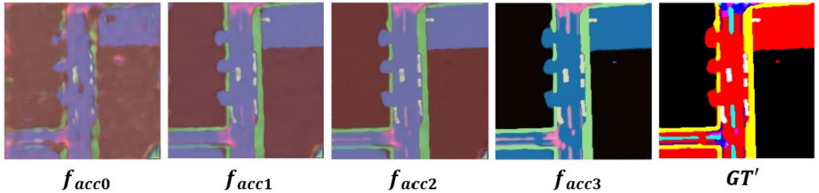
RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation

Arxiv 2025
(T-ITS summit)

Zhiwen Zeng, Yunfei Yin*, *Member, IEEE*, Zheng Yuan, Argho Dey, and Xianjian Bao

- Problem/Objective
  - Bird’s Eye View Segmentation



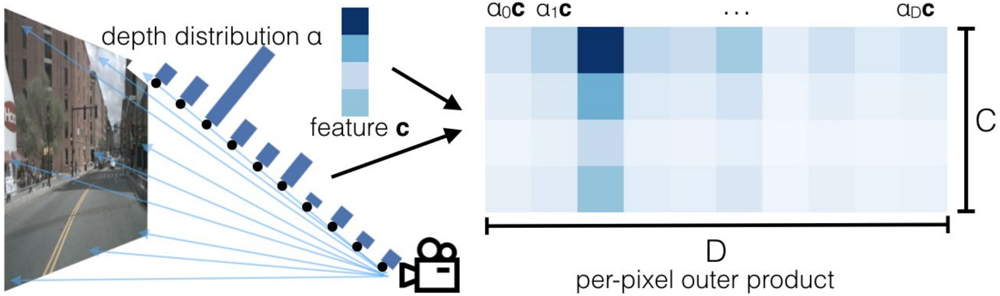
- Contribution/Key Idea
  - Coarse - to - Fine progressive refinement BEV segmentation
  - Residual learning + Autoregressive
  - 향상된 결과 및 real-time capability

김범준

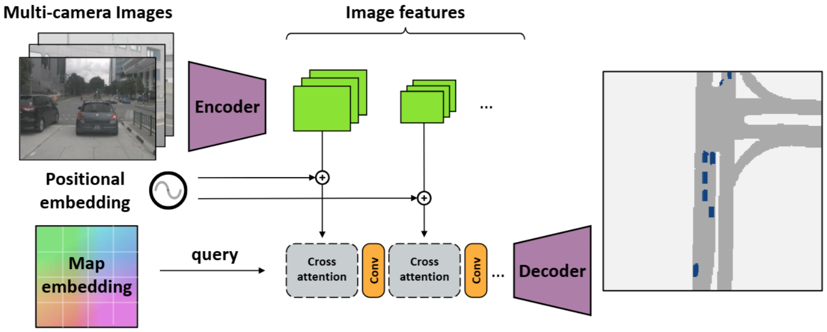
RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation

Arxiv 2025
(T-ITS summit)

- ## Introduction



Lift, Splat, Shoot 부류



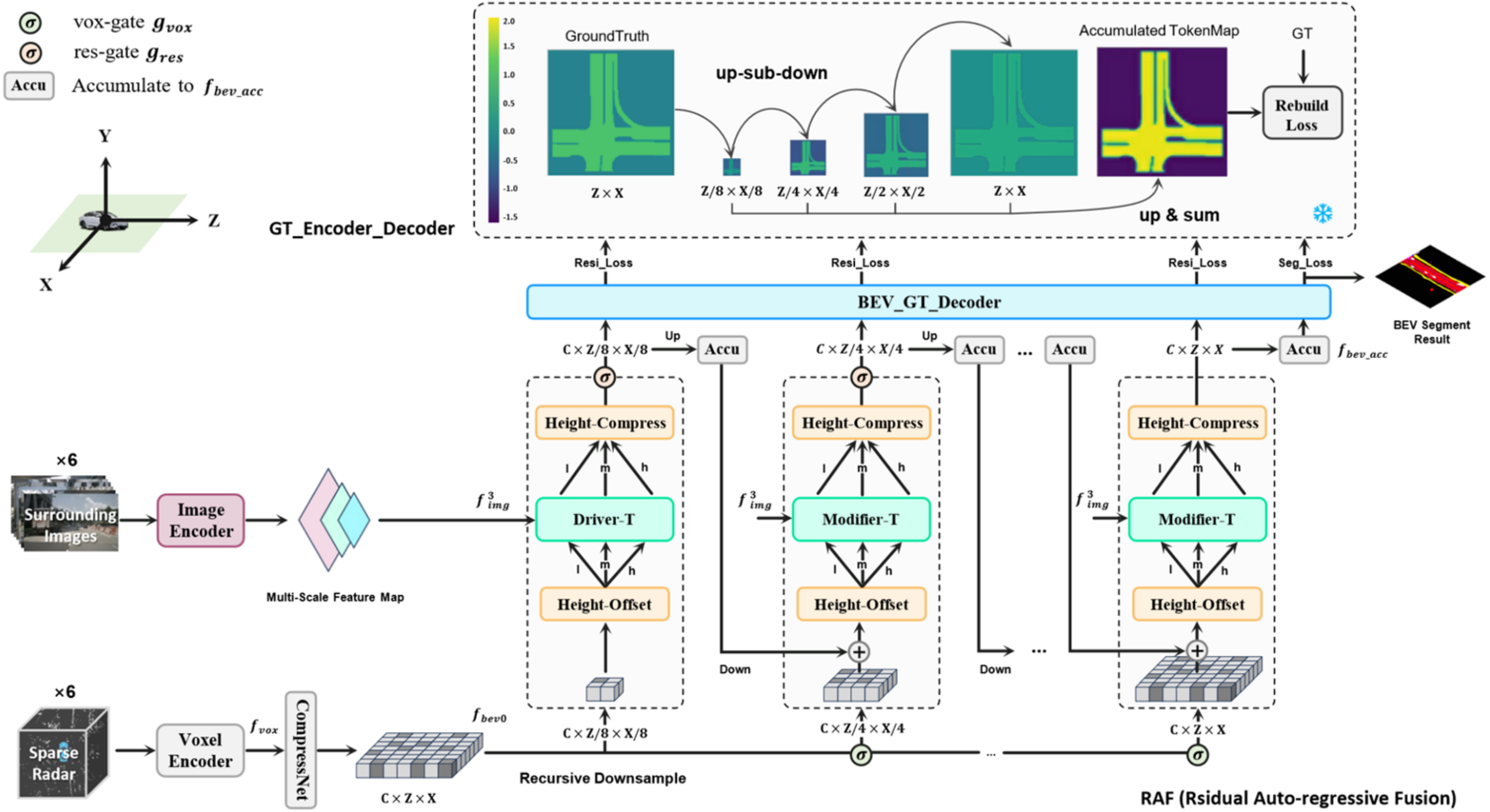
Transformer 부류

- *Geometry-based* or *Learning-based* 두 갈래가 존재 → 어느 쪽이든 sensor misalignment에 취약
- + remaining method는 모두 “single step end-to-end” 라는 점에 주목 (한번에 모든 픽셀 예측)
    - 이는 occlusion / long-range 등 challenging 상황에서 **오류가 발생하면 복구 불가, 에러가 고착됨**
        - By Depth estimation limitation / global attention error
        - + 2D → 3D 변환의 ambiguity도 한 번에 예측해서 error ↑
        - **도로 구조에서 세부 차선까지 “hierarchical(계층적) reasoning”이 부족**

RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation

Arxiv 2025
(T-ITS summit)

- **Method** - ***Overall***

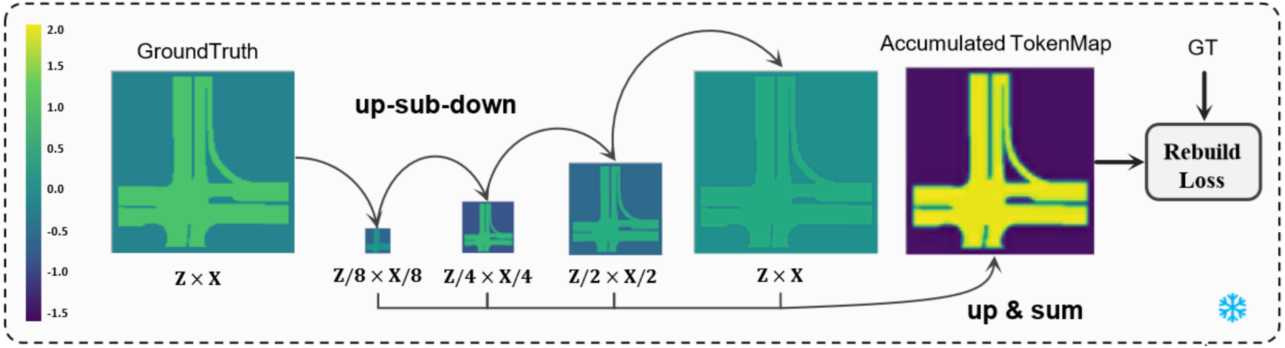


김범준

RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation

Arxiv 2025
(T-ITS summit)

- **Method** - ***Multi-Scale Ground-Truth Token Maps Decomposition***



Algorithm 1 Multi-scale Ground Truth Decomposition

Input: Original mask $GT \in \mathbb{R}^{C \times Z \times X}$, levels N

Output: Token maps $\{TP_i\}_{i=1}^{N}$, reconstructed $\hat{GT}$

```
DECOMPOSE($\mathbf{GT}, N$)
    $R_1 \leftarrow \mathbf{GT}$
    $\hat{\mathbf{GT}} \leftarrow \mathbf{0}$
    for $i = 1$ to $N - 1$ do
        $\mathbf{TP}_i \leftarrow \tanh(\textsc{AvgPoolConv}(R_i, (\frac{Z}{2^{N-i}}, \frac{X}{2^{N-i}})))$
        $\hat{\mathbf{TP}}_i \leftarrow \textsc{Bicubic}(\mathbf{TP}_i, (Z, X))$
        $R_{i+1} \leftarrow R_i - \sigma(\theta_i^{(C)}) \odot \hat{\mathbf{TP}}_i$
        $\hat{\mathbf{GT}} \leftarrow \hat{\mathbf{GT}} + \hat{\mathbf{TP}}_i$
    end for
    $\mathbf{TP}_N \leftarrow \tanh(R_N)$
    $\hat{\mathbf{GT}} \leftarrow \hat{\mathbf{GT}} + \mathbf{TP}_N$
    return $\{\mathbf{TP}_i\}_{i=1}^{N}, \hat{\mathbf{GT}}$
```

- Offline 과정으로 GT를 여러 resolution으로 나누어 학습 및 저장

-1 ~ 1 으로 값 강제 조정하여 token map 안정성 도모

가우시안의 w 값(표준편차)를 조절

GT : $C \times Z \times X,\quad \text{GT as residual } R_1$

Step by step: $Z/2^{N-i} \times X/2^{N-i}$

$$R_{i+1} = R_i - \sigma(\theta) \odot \tanh(\text{Down}(R_i))$$

가우시안의 w 값(표준편차)를 조절

learnable parameter

김범준

RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation

Arxiv 2025
(T-ITS summit)

- ## Method - *Encoder part + voxel encoding*

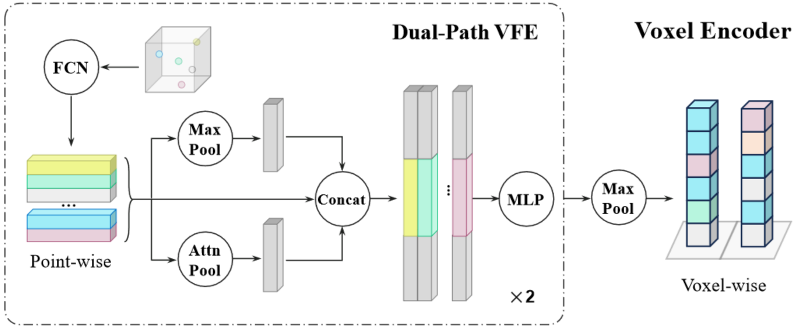
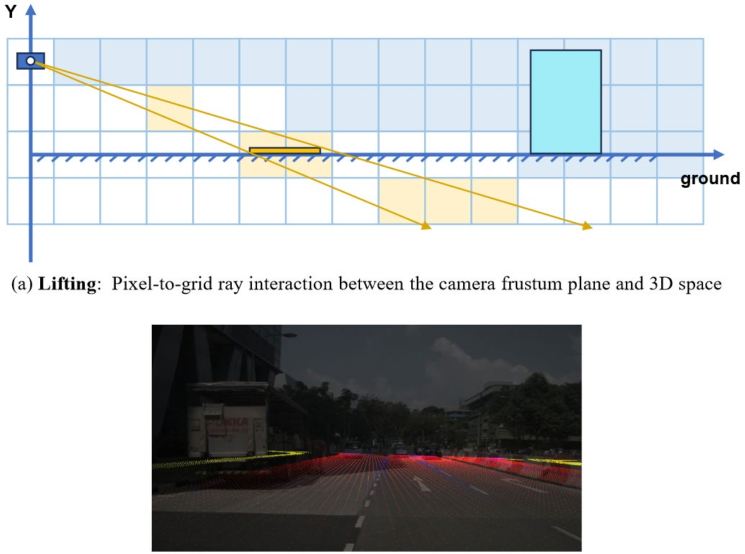


Fig. 3. Voxel feature extraction: we normalize each voxel to 10 points, then extract $C \times 10$ features via point-wise encoding. Apply parallel max/attention-pooling, concatenate with original features $(3C \times 10)$, and compress to C channels via MLP. Repeat twice, then max-pool for final voxel features.



(a) **Lifting**: Pixel-to-grid ray interaction between the camera frustum plane and 3D space

(b) **Unlifting**: BEV grid back-projection onto the image plane

Fig. 4. Lifting and Unlifting Visualization Based on Camera Sensor Intrinsics.

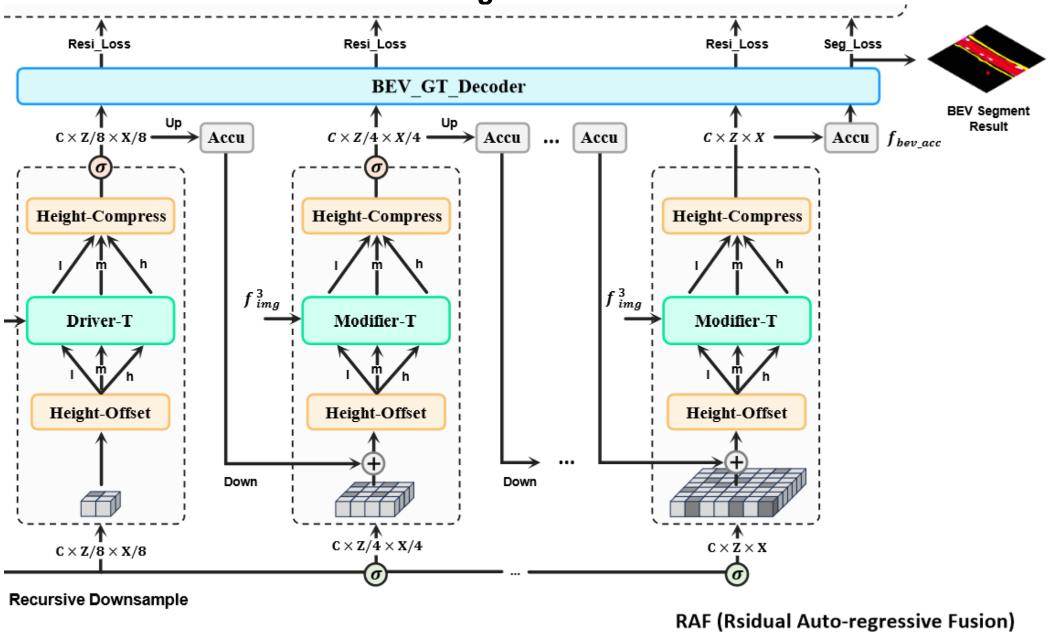
$$Y_{new}^{(i)} = Y_{gr}^{(i)} + ofst_{min} + Y_{drift}^{(i)} \cdot (ofst_{max} - ofst_{min})$$

- Ground level이 일정하지 않은 것을 고려하고 싶다.
  - ground-proximity 영역을 학습적으로 최적화

RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation

Arxiv 2025
(T-ITS summit)

- **Method** - ***Residual Auto-regressive Fusion***



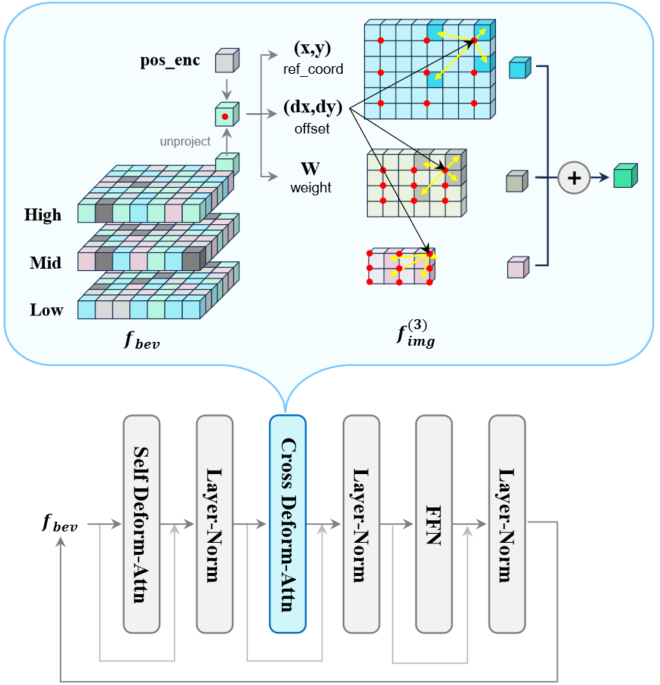


Fig. 5. Architecture of Driver/Modifier Transformer decoders. Cascaded decoders process learnable 3-layer f_{bev}, where Cross Deformable Attention enables BEV-to-multi-view semantic interaction. Modifier stages maintain independent cross-attention modules while sharing other components.

- Driver-T / Modifier-T 구조 동일
  - cross attn/샘플링 point 수 차이 (고해상도일수록 더 넓게 샘플링)

$$GT = \sigma(\theta_0) \cdot TP_0 + \sigma(\theta_1) \cdot TP_1 + \sigma(\theta_2) \cdot TP_2 + TP_3 \quad (3)$$

김범준

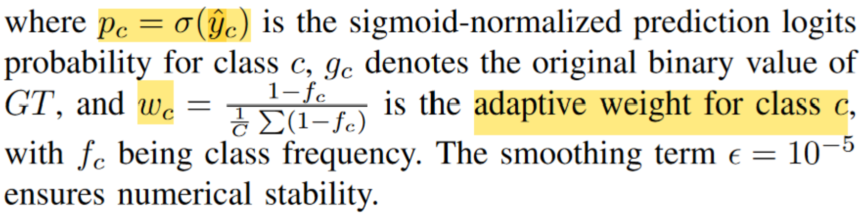
RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation

Arxiv 2025
(T-ITS summit)

- **Method** - ***Overall training loss***

$$\mathcal{L}_{TP_s} = \begin{cases} \frac{1}{ZX} \sum_{z,x} \|TP_i - \hat{TP}_i\|_p & \text{(Spatial-wise)} \\ \frac{1}{BC} \sum_{b,c} \|TP_i - \hat{TP}_i\|_p & \text{(Channel-wise)} \end{cases}$$

$$\mathcal{L}_{seg} = \frac{1}{C} \sum_{c=1}^{C} w_c \left(1 - \frac{2 \sum p_c \cdot g_c + \epsilon}{\sum p_c + \sum g_c + \epsilon} \right)$$



where $p_c = \sigma(\hat{y}_c)$ is the sigmoid-normalized prediction logits probability for class c, g_c denotes the original binary value of GT, and $w_c = \frac{1-f_c}{\frac{1}{C}\sum(1-f_c)}$ is the adaptive weight for class c, with f_c being class frequency. The smoothing term $\epsilon = 10^{-5}$ ensures numerical stability.

RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation

Arxiv 2025
(T-ITS summit)

- **Experiment -**

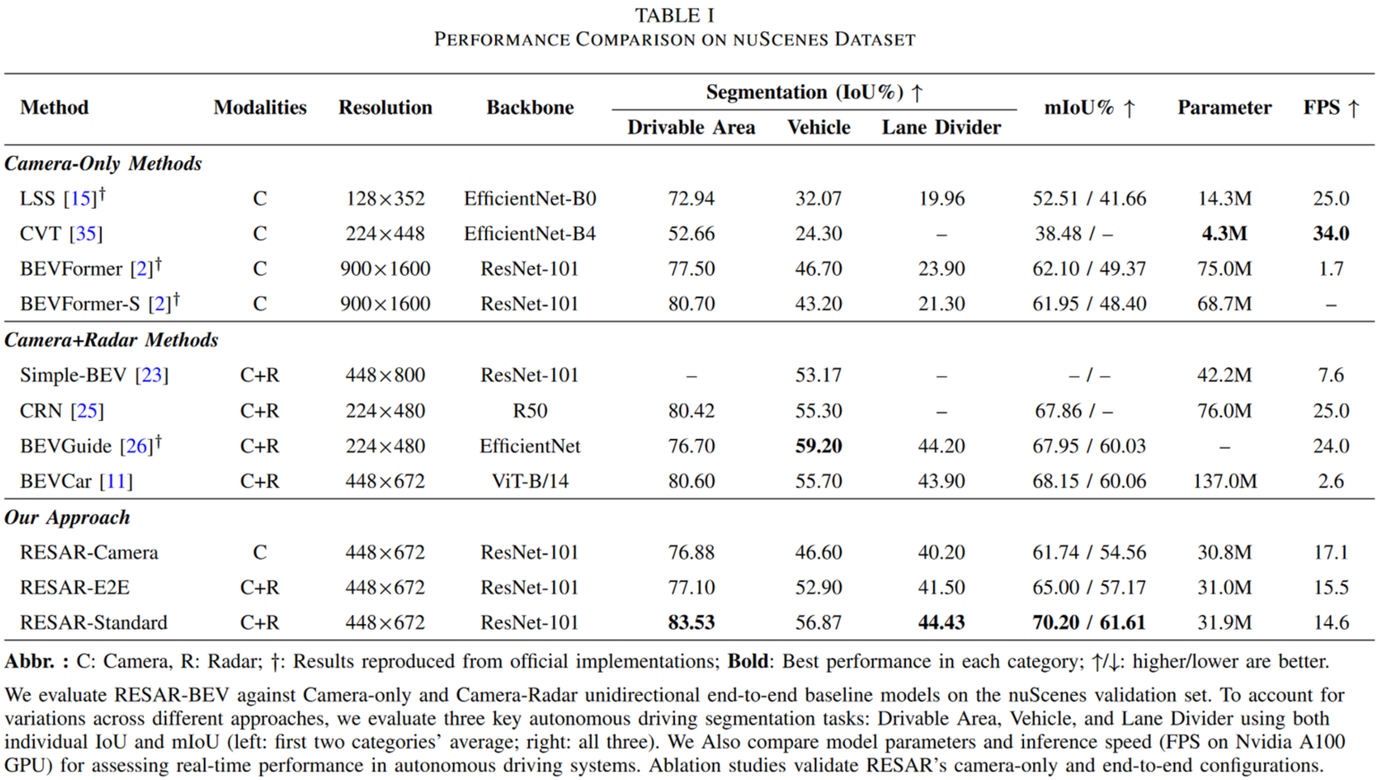


TABLE I
PERFORMANCE COMPARISON ON NUSCENES DATASET

Method	Modalities	Resolution	Backbone	Segmentation (IoU%) ↑			mIoU% ↑	Parameter	FPS ↑
				Drivable Area	Vehicle	Lane Divider			
Camera-Only Methods									
LSS [15]†	C	128×352	EfficientNet-B0	72.94	32.07	19.96	52.51 / 41.66	14.3M	25.0
CVT [35]	C	224×448	EfficientNet-B4	52.66	24.30	–	38.48 / –	**4.3M**	**34.0**
BEVFormer [2]†	C	900×1600	ResNet-101	77.50	46.70	23.90	62.10 / 49.37	75.0M	1.7
BEVFormer-S [2]†	C	900×1600	ResNet-101	80.70	43.20	21.30	61.95 / 48.40	68.7M	–
Camera+Radar Methods									
Simple-BEV [23]	C+R	448×800	ResNet-101	–	53.17	–	– / –	42.2M	7.6
CRN [25]	C+R	224×480	R50	80.42	55.30	–	67.86 / –	76.0M	25.0
BEVGuide [26]†	C+R	224×480	EfficientNet	76.70	**59.20**	44.20	67.95 / 60.03	–	24.0
BEVCar [11]	C+R	448×672	ViT-B/14	80.60	55.70	43.90	68.15 / 60.06	137.0M	2.6
Our Approach									
RESAR-Camera	C	448×672	ResNet-101	76.88	46.60	40.20	61.74 / 54.56	30.8M	17.1
RESAR-E2E	C+R	448×672	ResNet-101	77.10	52.90	41.50	65.00 / 57.17	31.0M	15.5
RESAR-Standard	C+R	448×672	ResNet-101	**83.53**	56.87	**44.43**	**70.20 / 61.61**	31.9M	14.6

Abbr. : C: Camera, R: Radar; †: Results reproduced from official implementations; **Bold**: Best performance in each category; ↑/↓: higher/lower are better.

We evaluate RESAR-BEV against Camera-only and Camera-Radar unidirectional end-to-end baseline models on the nuScenes validation set. To account for variations across different approaches, we evaluate three key autonomous driving segmentation tasks: Drivable Area, Vehicle, and Lane Divider using both individual IoU and mIoU (left: first two categories' average; right: all three). We Also compare model parameters and inference speed (FPS on Nvidia A100 GPU) for assessing real-time performance in autonomous driving systems. Ablation studies validate RESAR's camera-only and end-to-end configurations.

RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation

Arxiv 2025
(T-ITS summit)

- **Experiment -**

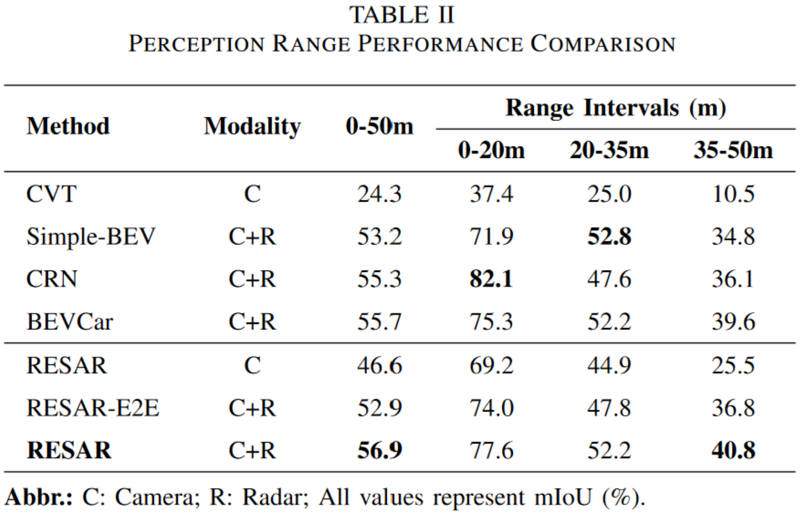


TABLE II
PERCEPTION RANGE PERFORMANCE COMPARISON

Method	**Modality**	**0-50m**	**Range Intervals (m)**		
			0-20m	**20-35m**	**35-50m**
CVT	C	24.3	37.4	25.0	10.5
Simple-BEV	C+R	53.2	71.9	**52.8**	34.8
CRN	C+R	55.3	**82.1**	47.6	36.1
BEVCar	C+R	55.7	75.3	52.2	39.6
RESAR	C	46.6	69.2	44.9	25.5
RESAR-E2E	C+R	52.9	74.0	47.8	36.8
RESAR	C+R	**56.9**	77.6	52.2	**40.8**

Abbr.: C: Camera; R: Radar; All values represent mIoU (%).

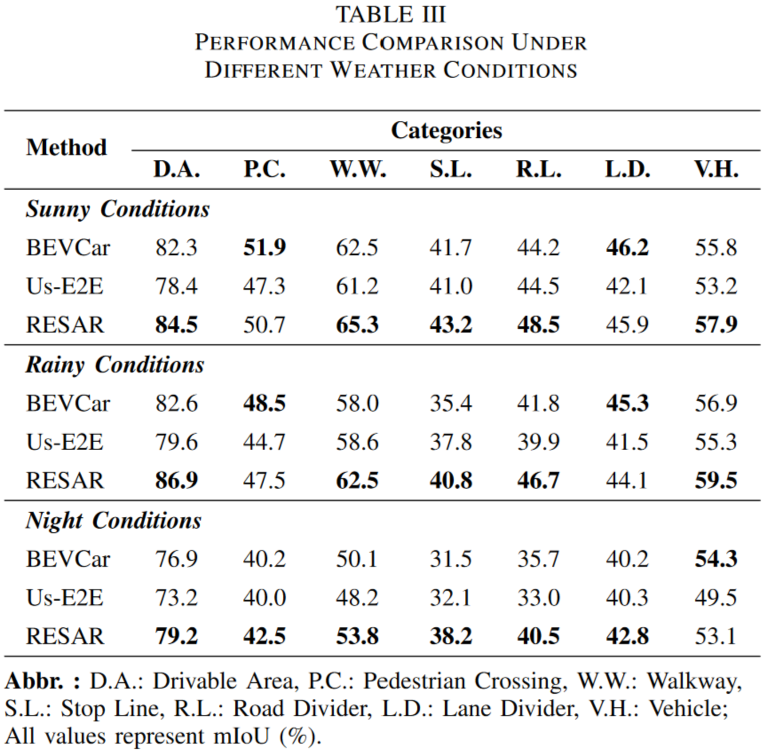


TABLE III
PERFORMANCE COMPARISON UNDER
DIFFERENT WEATHER CONDITIONS

Method	**Categories**						
	D.A.	**P.C.**	**W.W.**	**S.L.**	**R.L.**	**L.D.**	**V.H.**
Sunny Conditions							
BEVCar	82.3	**51.9**	62.5	41.7	44.2	**46.2**	55.8
Us-E2E	78.4	47.3	61.2	41.0	44.5	42.1	53.2
RESAR	**84.5**	50.7	**65.3**	**43.2**	**48.5**	45.9	**57.9**
Rainy Conditions							
BEVCar	82.6	**48.5**	58.0	35.4	41.8	**45.3**	56.9
Us-E2E	79.6	44.7	58.6	37.8	39.9	41.5	55.3
RESAR	**86.9**	47.5	**62.5**	**40.8**	**46.7**	44.1	**59.5**
Night Conditions							
BEVCar	76.9	40.2	50.1	31.5	35.7	40.2	**54.3**
Us-E2E	73.2	40.0	48.2	32.1	33.0	40.3	49.5
RESAR	**79.2**	**42.5**	**53.8**	**38.2**	**40.5**	**42.8**	53.1

Abbr. : D.A.: Drivable Area, P.C.: Pedestrian Crossing, W.W.: Walkway, S.L.: Stop Line, R.L.: Road Divider, L.D.: Lane Divider, V.H.: Vehicle; All values represent mIoU (%).

RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation

- **Experiment -**

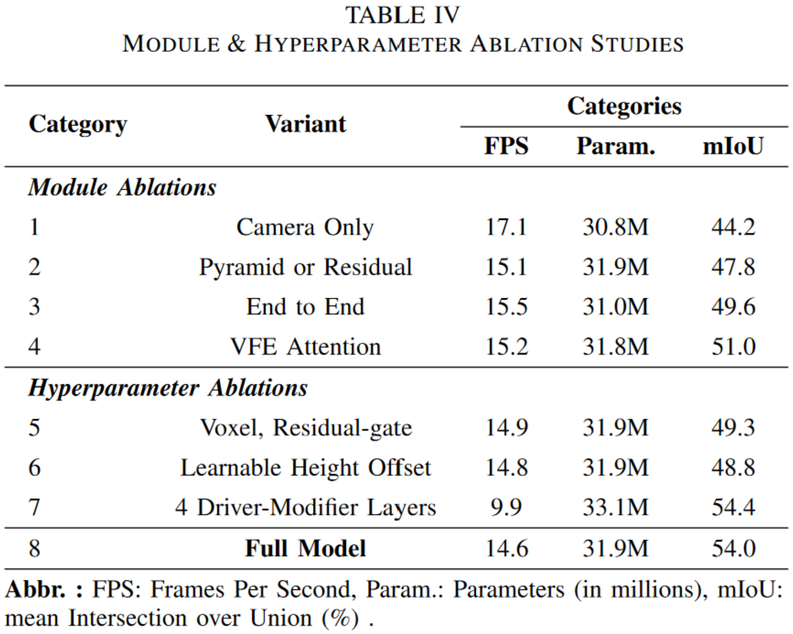


TABLE IV
MODULE & HYPERPARAMETER ABLATION STUDIES

Category	Variant	Categories		
		FPS	Param.	mIoU
Module Ablations				
1	Camera Only	17.1	30.8M	44.2
2	Pyramid or Residual	15.1	31.9M	47.8
3	End to End	15.5	31.0M	49.6
4	VFE Attention	15.2	31.8M	51.0
Hyperparameter Ablations				
5	Voxel, Residual-gate	14.9	31.9M	49.3
6	Learnable Height Offset	14.8	31.9M	48.8
7	4 Driver-Modifier Layers	9.9	33.1M	54.4
8	**Full Model**	14.6	31.9M	54.0

Abbr. : FPS: Frames Per Second, Param.: Parameters (in millions), mIoU: mean Intersection over Union (%) .

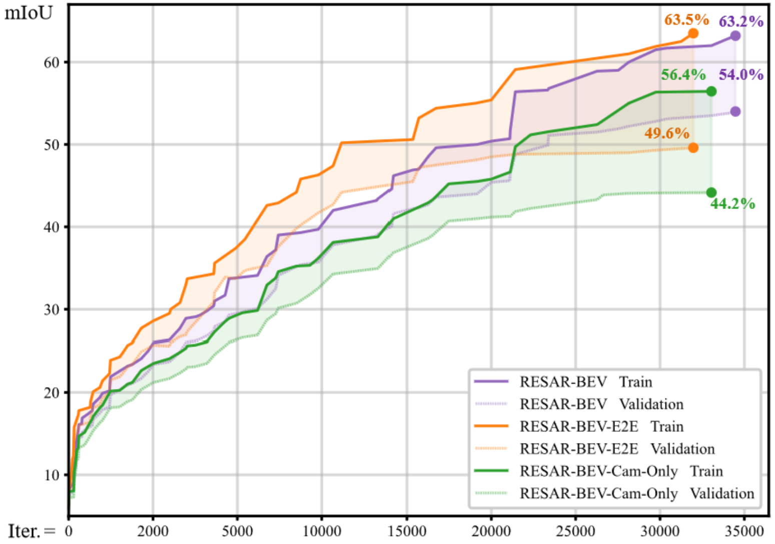


Fig. 6. Training and validation mIoU trajectories for complete RESAR-BEV model versus two ablated models across iterations (32 batches/iteration). The semi-transparent bands indicate performance gaps between training and validation sets.

- **train/validation gap이 크지 않음** (오버피팅 적고, 모델 일반화 성능 좋음)
- 학습 안정성/수렴 속도도 Full Model이 가장 우수

RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation

Arxiv 2025
(T-ITS summit)

- **Experiment -**

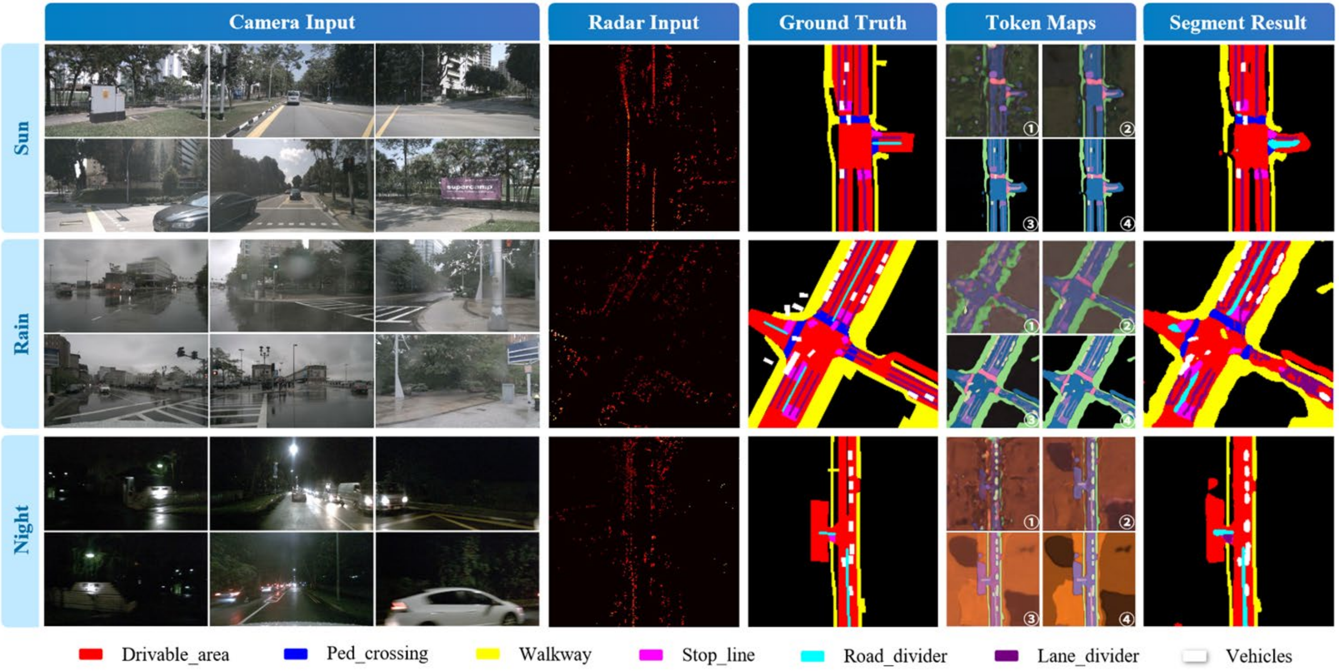
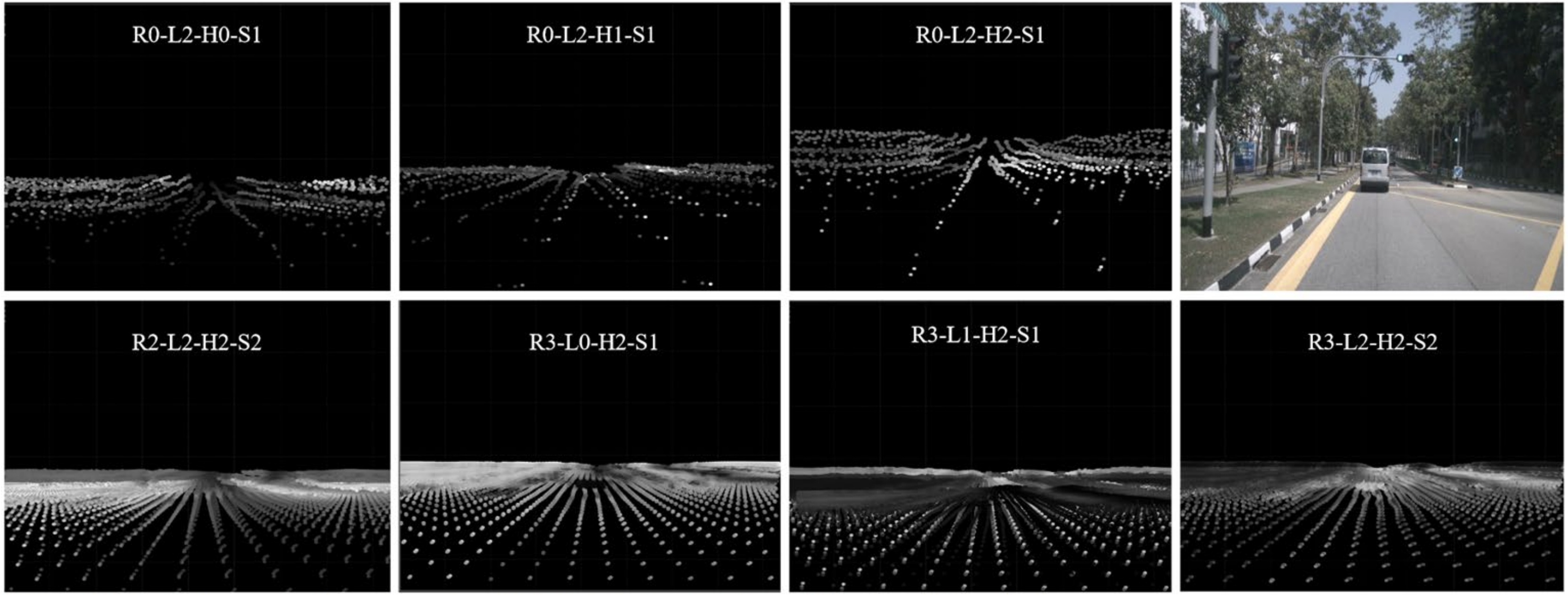


Fig. 7. Progressive Multi-modal BEV Semantic Segmentation via Residual Auto-regression Across Diverse Environmental Conditions. The model integrates synchronized inputs from six surround-view cameras and six consecutive frames of radar point clouds through four-step residual auto-regression (shown as accumulated residuals), outputting seven-class BEV segmentation.

[RESAR-BEV: An Explainable Progressive Residual Autoregressive Approach for Camera-Radar Fusion in BEV Segmentation]

Arxiv 2025
(T-ITS summit)

- **Experiment -**



Abbr. : R: Residual stage; **L**: Image feature level; **H**: Attention head; **S**: Driver-T/Modifier-T Decoder layer.

Fig. 8. The visualization of cross-modal attention weights in Driver-T and Modifier-T under the image view, demonstrating a distinct vertical and road-category hierarchical attention pattern across different residual modules, image feature layers, attention heads, and decoder module levels.