Epipolar Attention Field Transformers for Bird's Eye View Semantic Segmentation

Christian Witte[1,2] Jens Behley[2] Cyrill Stachniss[2,3] Marvin Raaijmakers[1]

[1]CARIAD SE, Germany [2]Center for Robotics, University of Bonn, Germany
[3]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany

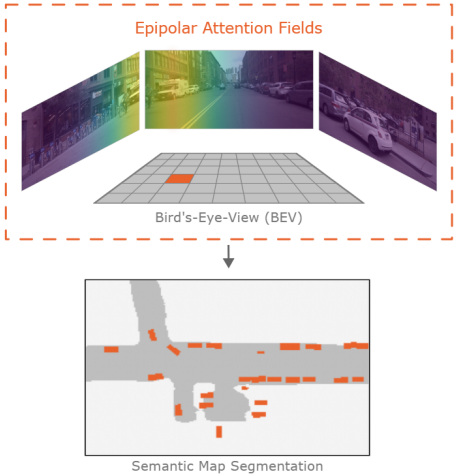
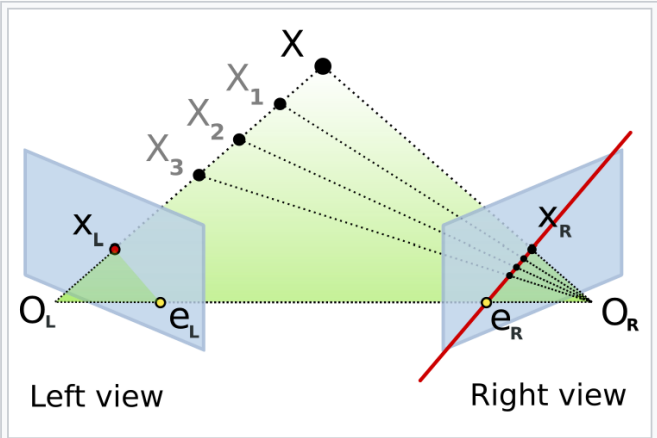


Figure 1. Our approach attends features of multi-view camera images to BEV features by computing Epipolar Attention Fields (top) instead of leveraging learnable positional encoding for BEV semantic segmentation (bottom), where we predict the drivable area (dark grey) and vehicles (orange).

- Problem/Objective
  - Bird's Eye View Semantic Segmentation
- Contribution/Key Idea
  - BEV feature로 변환하는 새로운 방법을 제시
  - Epipolar Attention Field Transformers(EAFormer)를 제안
  - 향상된 결과

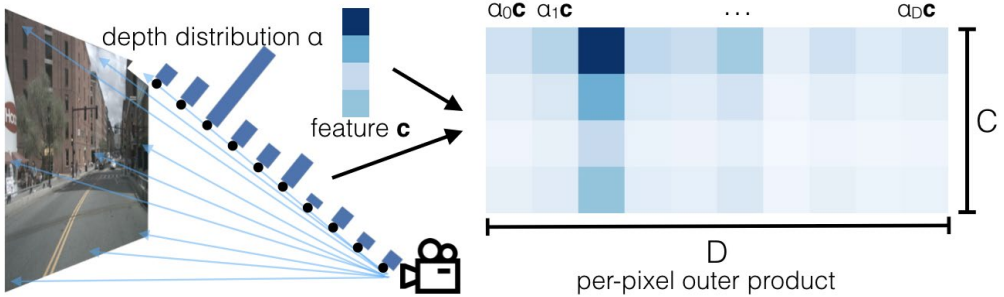
- **Epipolar geometry**



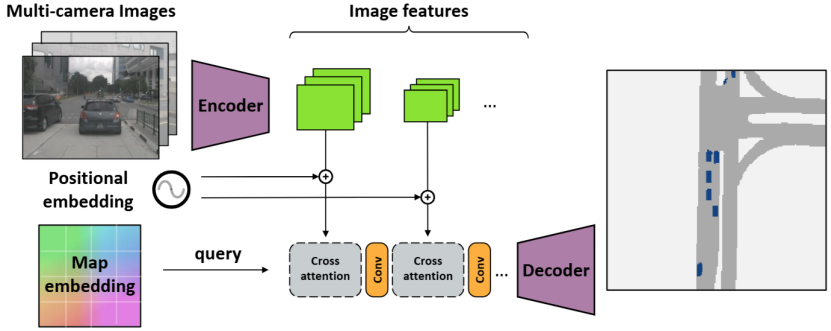
Epipolar geometry

- multi-view에서 3D - 2D 사이의 **geometric 관계**를 설명하는 개념
- line(O_L–X) 는 왼쪽 카메라 world line 상 = 점 but, 오른쪽에서는 선
- 노란색 점: epipole, 빨간 선: epipolar line

- **BEV semantic segmentation 의 두 갈래**



Lift, Splat, Shoot 부류



Transformer 부류

- LSS 부류, CVT 부류 두 갈래가 존재
- 이 논문은 transformer 기반 CVT 부류에 주목
  - Positional Encoding 에 한계점이 있다고 주장
    - PE는 카메라의 intrinsic/extrinsic info → PE vector로
    - 새로운 카메라 세팅, PE가 무용지물이 되어 다시 학습해야함 등등

- **Method - epipolar line with BEV**

- 논문에서는 **BEV grid**를 하나의 "이미지 평면"으로 간주
- BEV에서 본 한 점 x_0(즉, BEV grid의 한 셀)에 대해
  - "O_0에서 x_0로 향하는 직선(ray)" 위의 모든 점과 동치(equivalence class)
  - "O_0-x_0" 직선 위의 모든 점 **= epipolar line**으로 투영

$$\mathbf{l}_i = E_i \mathbf{x}_0$$

essential matrix E_i

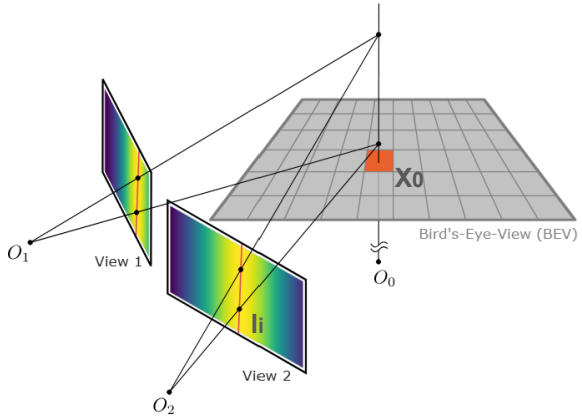


Figure 2. The BEV (grey) is considered the reference view for epipolar geometry. We project each coordinate of the grid cell center onto all other views. A given grid cell (red) thus corresponds to the epipolar lines projected onto the source views 1 and 2. We impose a normal distribution on the distance to the epipolar line and obtain Epipolar Attention Fields, which are visualized as heatmaps in the source views.

• Method - Attention weighting

- 기존 Transformer → Q, K, V = Linear(PE + feature)
  - 이 PE에 extrinsic, intrinsic + BEV grid 내 좌표를 vector로 변환하여 포함

$$\tilde{X} = X + \mathrm{PE}$$

etric 정보가 Q, K, V에 implicit

$$Q = \tilde{X}W_Q, \quad K = \tilde{X}W_K, \quad V = \tilde{X}W_V$$

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- 본 논문에서 Attention weighting
  - attention score 계산에 epipolar geometry 가중치 W 곱하는 방식을 제안

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

ion scor

$$\mathrm{Attention}(W, Q, K, V) = \mathrm{softmax}\left(\frac{W \odot (QK^T)}{\sqrt{d_k}}\right)V$$

● Method - Epipolar Attention Fields

- Attention weight 계산 방법
  - Query / key
    - BEV grid의 q번째 셀: **query q**
    - image feature map의 k번째 위치: **key k**
    - key k가 소속된 카메라 뷰: **i**
  - Key와 Epipolar Line 사이의 거리 계산
    - key k의 이미지 좌표: x_i
    - x_i 와 epipolar line li 사이에 **점-직선 거리** 공식으로 계산
  - $$W_{q,k} = \exp\left(-(\lambda\lambda_{q,i})^2 \left(\mathbf{x}_i \hat{\mathbf{l}}_i^\intercal\right)^2\right)$$

$l_i = E_i x_0$

gaussian width 조절 hyperparameter

key k ⟷ epipolar line 거리

q와 카메라 간의 상대 거리 등에 따라 local하게 width를 조절
(BEV cell이 카메라에서 멀면 line이 더 두꺼워지는 효과 반영)

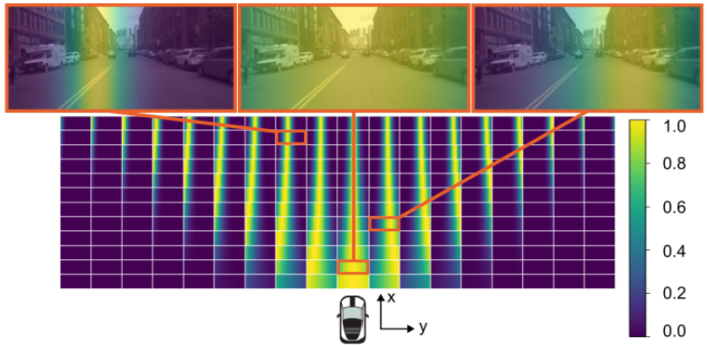
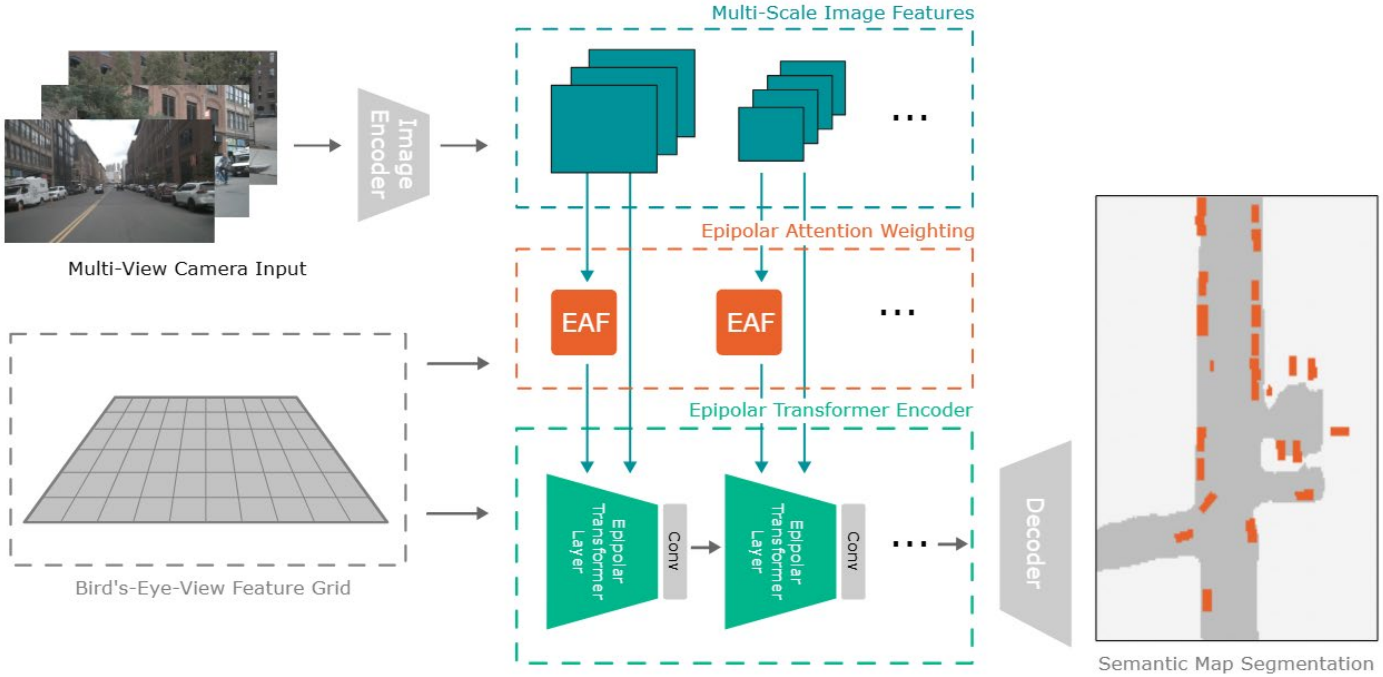


Figure 3. Epipolar Attention Fields for the front camera of nuScenes. The grid (middle) visualizes the BEV cells that are visible to the front view camera. The imaginary location of the vehicle is at the bottom center of the grid, facing upward. For illustrative purposes, each BEV cell is visualized by the Epipolar Attention Field heatmap of a corresponding query. Yellow indicates high attention weights, while blue indicates low attention weights. For three distinct BEV grid cells (red outline), we overlay the heatmap with the actual input image showing the Epipolar Attention Field.

- **Method - Architecture**



- **Experiment**

Table 1. Comparisons with state-of-the-art methods on the nuScenes validation set. We report the mIoU for semantic segmentation of the drivable area (drivable) and the orthogonal vehicle projection (vehicle). LSS uses visibility $> 0\%$, all others use visibility $> 40\%$.
†: Approach uses additional sensors, temporal information and/or a significantly larger backbone.

	Drivable	Vehicle
LSS [23]	72.9	32.1
FIERY static [10]	-	35.8
CVT [40]	74.3	36.0
M^2BEV [33]	77.2	-
GKT [4]	-	38.0
BAEFormer [21]	76.0	38.9
EAFormer (ours)	**78.0**	**39.0**

Table 2. Cross-dataset evaluation for AV2 and nuScenes. The experiments show the influence of changes in camera parameters for different camera setups in a zero-shot transfer setting. Methods are trained on the source dataset and then they are evaluated on the target dataset without retraining. We denote the zero-shot transfer from source dataset to target dataset as: source dataset $\rightarrow$ target dataset. AV2* is the AV2 dataset where the front-left stereo camera was used instead of the front ring camera.

	Training Epochs	CVT [40]	EAFormer (ours)
nuScenes	12	34.31	38.18
$\rightarrow$ nuScenes	30	36.69	38.98
AV2	12	38.00	38.66
$\rightarrow$ AV2	30	38.47	39.60
AV2	12	30.92	32.99
$\rightarrow$ AV2*	30	31.21	33.91
nuScenes	12	7.78	19.72
$\rightarrow$ AV2	30	7.86	14.00
AV2	12	3.07	12.17
$\rightarrow$ nuScenes	30	2.70	11.44

- **Experiment**

Table 5. Distance-based evaluation (mIoU) for vehicle segmentation on nuScenes.

	Epochs	0 - 10 m	10 - 20 m	20 - 30 m	30 - 40 m	40 - 50 m	mIoU
CVT	12	67.28	53.01	37.42	22.60	11.02	34.02
EAFormer (ours)	12	68.59	55.33	40.93	27.54	16.42	36.70
CVT	30	68.62	55.38	40.37	26.85	15.44	36.69
EAFormer (ours)	30	71.53	57.64	41.76	30.8	18.71	38.98

Table 3. Performance comparison for different splits. We report the mIoU for semantic segmentation of the drivable area for the original nuScenes data split and the disjoint split without data leakage [38]. Both models were trained for 30 epochs.

	CVT [40]	EAFormer (ours)
Original Split	76.02	78.04
New Split	54.23	58.06
Difference	-21.79	**-19.98**

Table 7. Ablations about the distance strength parameter λ.

Distance Strength λ	Vehicle	Drivable
1.0	38.76	78.04
1.4	38.00	77.65
learnable	38.77	78.10

- **Experiment**

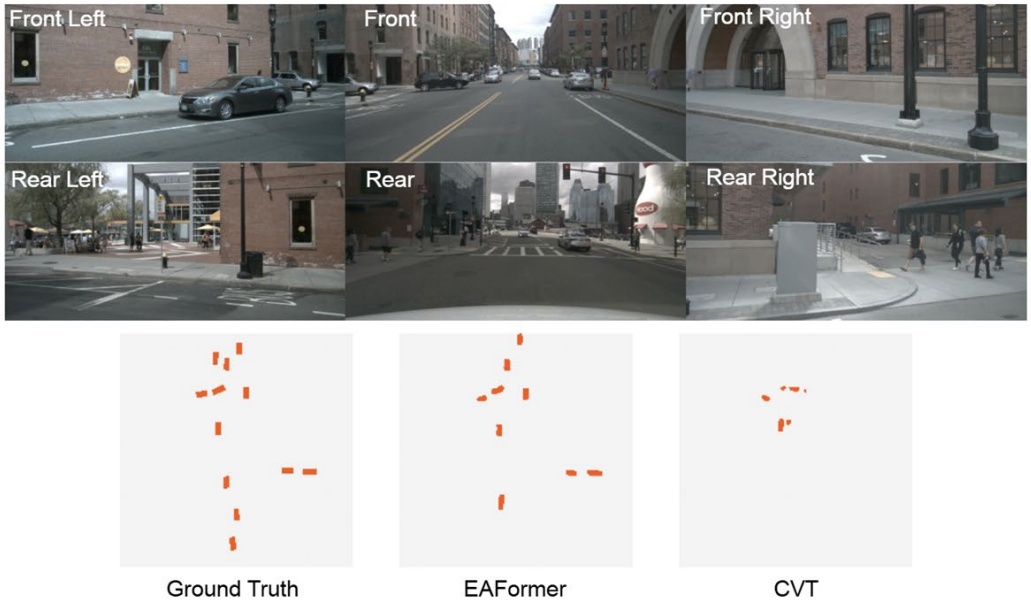


Figure 5. Visualization of zero-shot transfer performance of EAFormer and CVT for vehicle segmentation. The models were trained on Argoverse 2 (AV2) and evaluated on nuScenes.

- PE가 없을때의 generalization 성능 강조