Supplementary Material: 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

Abstract

In this supplementary document, we first describe the implementation details of our proposed approach EAFormer and illustrate its architecture in detail. In Sec. 2, we provide additional quantitative and qualitative results to support the claims made in the main paper. Finally, we discuss the limitations of our approach in Sec. 3.

1. Implementation and Architecture Details

Our architecture consists of an EfficientNet-B4 [5] image backbone that outputs multi-scale image feature maps. For nuScenes [1], we process all 6 cameras, while Argoverse 2 (AV2) [6] provides 7 ring-cameras as well as a stereo-camera setup with two image streams.

Based on our ablation study, we have chosen the feature map outputs with scales of 1/4 and 1/16 as our default scales. For each feature map scale, the Epipolar Transformer Encoder generates new bird's eye features, which serve as input queries for the next Epipolar Transformer Encoder stage. Specifically, encoder's output for the feature map scale of 1/4 serves as input query for the next encoder stage, which attends to the queries for the features of scale of 1/16. The final output is upsampled by a decoder with three ASPP [2] blocks outputting the final semantic segmentation mask.

Four heads are used per Epipolar Transformer Encoder block with a dimensionality of 32. By default, the distance strength parameter is set to $\lambda = 1$. As for CVT [8], we use AdamW [3] with a one-cycle learning rate scheduling [4] and a target learning rate of 0.004. Additionally, weight decay regularization is utilized, and the default training duration is 30 epochs. For studies involving pixel augmentation, we use color jittering, image sharpening, and pixel dropout with a dropout rate of 0.5.

We project the epipolar line onto each view and compute the Epipolar Attention Fields (EAF) using the coordinates of the image feature maps and the bird's eye view feature grid. For two distinct spatial positions, the EAFs are visulized in Fig. 1. The EAF serve as attention weight W in the weighted attention computation:

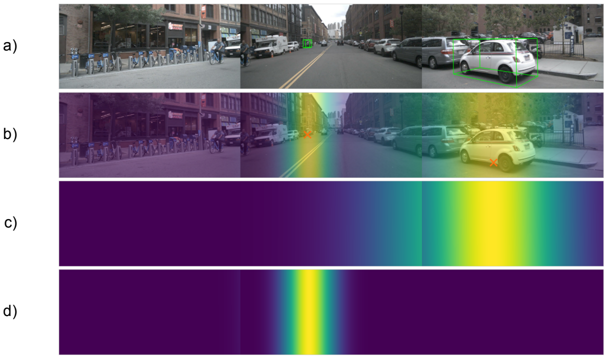

Figure 1. Visualization of Epipolar Attention Fields for specific grid locations. We depict the images of the front and front-side cameras. For two 3D object annotations (green) (a), we project the object center (red crosses) onto the ground and overlay the resulting Epipolar Attention Fields to the images (b). The figures (c) and (d) depict the individual heatmaps of the Epipolar Attention Field for each selected object location.

$$Attention(W, Q, K, V) = softmax\left(W \odot \frac{QK^\intercal}{\sqrt{d_k}}\right) V \tag{1}$$

The bird's eye view feature grid elements represent the queries Q used for attention computation. The keys and values correspond to the image features. It is important to note that the keys do not include any positional encodings. The attention weights W provide information on spatial relationships, which is accounted for by computing the Hadamard product between the scaled dot product and the EAF. In Fig. 2, we present a decomposed visualization of the key elements of our architecture.

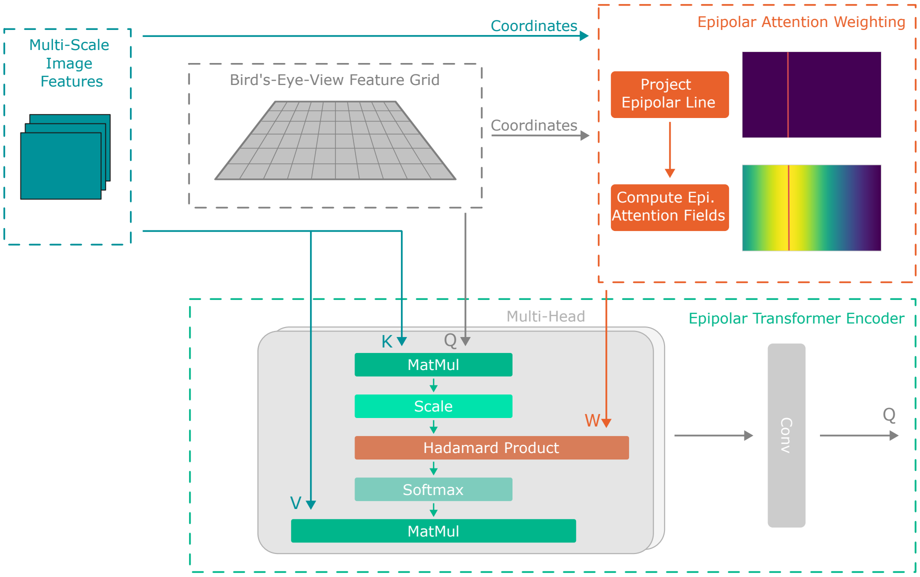

Figure 2. Simplified illustration of the key elements of our EAFormer architecture. The epipolar attention is computed iteratively for each scale of the image feature maps. First, we compute the epipolar lines for the all BEV coordinates and all views. Then, the Epipolar Attention Fields (EAF) are calculated based on the distance of each image feature coordinate to the epipolar line. We compute the attention between the input queries (Q) and the image features (V, K) for each Epipolar Transformer Encoder, weighted by the EAFs (W). After a convolution operation, the output is used as the input query for the next Epipolar Transformer Encoder.

Table 1. Intra-dataset evaluation for nuScenes. The zero-shot transfer experiments demonstrate the impact of domain shift and minor camera parameter deviations for two different cities within the same dataset. Initially, the methods are trained for 30 epochs on the source split, i.e., either Boston or Singapore. Then, without retraining, they are evaluated on the target split. We denote the zero-shot transfer from source dataset/split to target dataset/split as: source dataset → target dataset.

	nuScenes → nuScenes	Boston → Singapore	Singapore → Boston
CVT	36.69	17.38	17.68
EAFormer	38.76	18.89	18.68

2. Additional Experimental Results

In Tab. 2, results for the distance-based evaluation for the vehicle segmentation on AV2 are depicted. EAFormer shows almost consistently better segmentation performance for each depth interval. However, the performance gain for far range objects is not as evident as it is for nuScenes. Further experiments were performed by exchanging the front-view ring camera with the stereo cameras. The results are depicted in Tab. 3. The results show superior performance of EAFormer for the vehicle segmentation task when compared to CVT. However, the AV2 dataset is missing stereo camera calibration information for multiple scenes and these were filtered for these experiments. The dataset was reduced significantly and, thus, the results are not directly comparable to the evaluation with the standard camera configuration.

In addition to the cross-dataset evaluation, we observed that the camera parameters differ between the two cities in nuScenes, namely Boston and Singapore. In Tab. 1, we present the evaluation results for the zero-shot transfer, where we train on the one city and evaluate on the other. It is worth noting that the two cities are located on different continents and are subject to a domain shift. The results show that EAFormer generalizes slightly better than CVT.

Fig. 4 qualitatively shows the cross-dataset performance of EAFormer when trained on AV2 and evaluated (in a zero-shot transfer) on nuScenes. The figure illustrates the superior generalization capability of EAFormer compared to CVT.

Fig. 5 provides visualizations that illustrate the issue of the overlapping training and validation splits for the map annotation. The ground truth annotation also includes areas that may not be visible in the current frame. However, since the maps in the training and validation splits of nuScenes overlap, the network trained on the original split can predict non-visible streets or occluded junctions. This undermines the generalization capability of the networks trained on this data split. We also demonstrate the network's performance

Table 2. Distance-based evaluation (mIoU) for vehicle segmentation on Argoverse 2 (AV2). The standard ring camera configuration with 7 cameras is utilized.

	Epochs	0 - 10m	10 - 20m	20 - 30m	30 - 40m	40 - 50m	mIoU
CVT	12	72.43	59.02	42.48	29.55	20.71	38.00
EAFormer	12	74.26	60.12	42.76	30.08	20.25	38.66
CVT	30	73.73	59.47	42.74	30.25	21.4	38.47
EAFormer	30	74.44	60.38	43.54	31.5	22.0	39.6

Table 3. Distance-based evaluation (mIoU) for vehicle segmentation on Argoverse 2 (AV2) with stereo cameras. We replaced the front-ring camera with the stereo cameras for this experimental setup. Due to missing calibration data for the stereo cameras in the official dataset, the dataset size is reduced significantly compared to the standard camera configuration.

	Epochs	0 - 10m	10 - 20m	20 - 30m	30 - 40m	40 - 50m	mIoU
CVT	12	72.49	58.91	41.84	28.8	19.93	37.61
EAFormer	12	74.13	59.99	43.05	30.61	21.14	38.9
CVT	30	73.77	59.16	42.02	29.02	20.06	37.54
EAFormer	30	74.63	60.23	43.00	30.47	20.97	38.89

on the more realistic datasplit proposed by Yuan et al. [7]. As mentioned in our main paper, there is a significant decrease in performance. However, the semantic segmentation output clearly shows that the network cannot predict non-visible areas and therefore must focus more on visible image cues.

3. Limitations

Well-calibrated cameras are crucial for the final prediction performance of BEV perception approaches that rely on camera parameters. Although Epipolar Attention Fields account for parameter deviations, significant changes can still result in a major decrease in performance.

During training and for different camera settings, we must compute the essential matrix for the Epipolar Attention Fields online, which incurs additional computational overhead. For trainings involving data with non-changing camera parameters, the essential matrix can be pre-computed. Similarly, during deployment, it is important to account for changes in camera parameters that may be induced by online calibration. For each updated set of camera parameters, the essential matrix must also be updated for our approach.

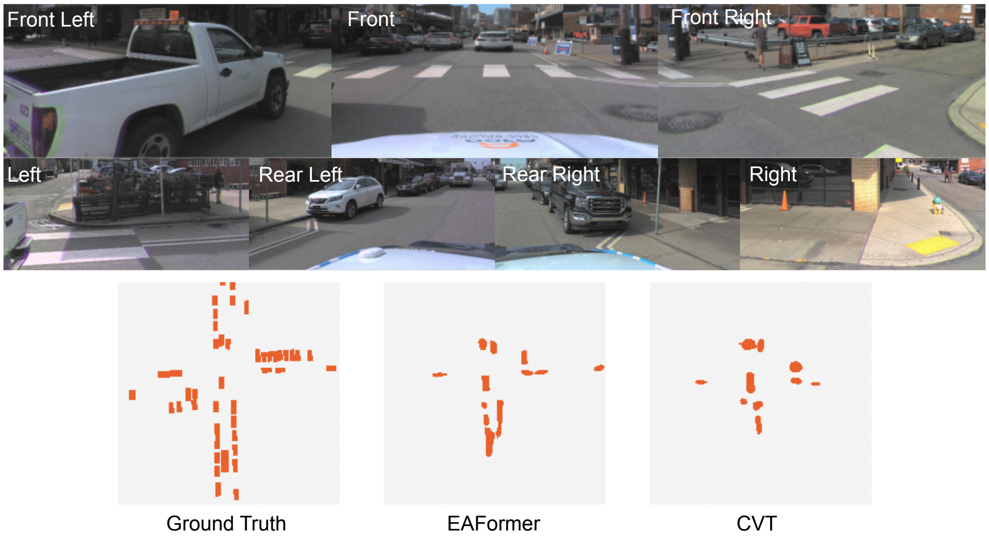

Figure 3. Visualization of zero-shot transfer performance for EAFormer and CVT. The models were trained on nuScenes and are evaluated on Argoverse 2 (AV2).

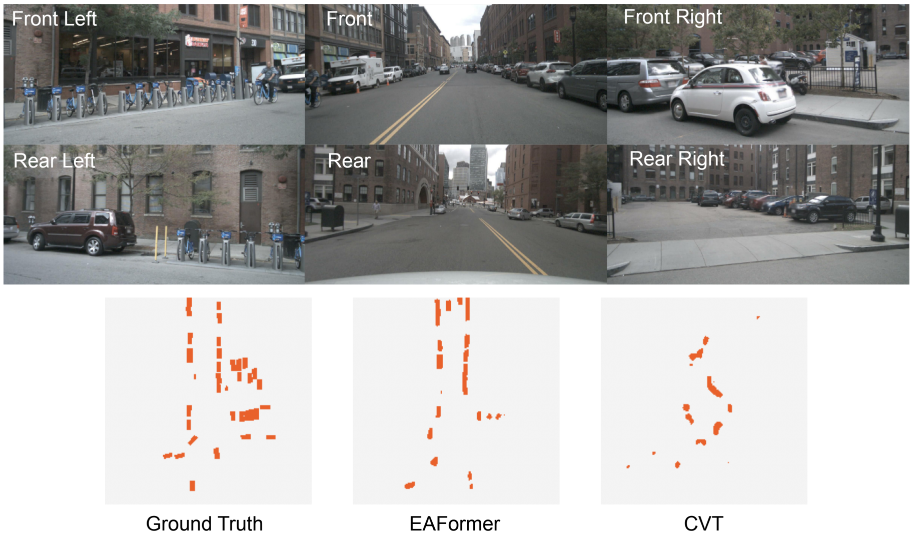

Figure 4. Visualization of zero-shot transfer performance for EAFormer and CVT. The models were trained on Argoverse 2 (AV2) and are evaluated on nuScenes.

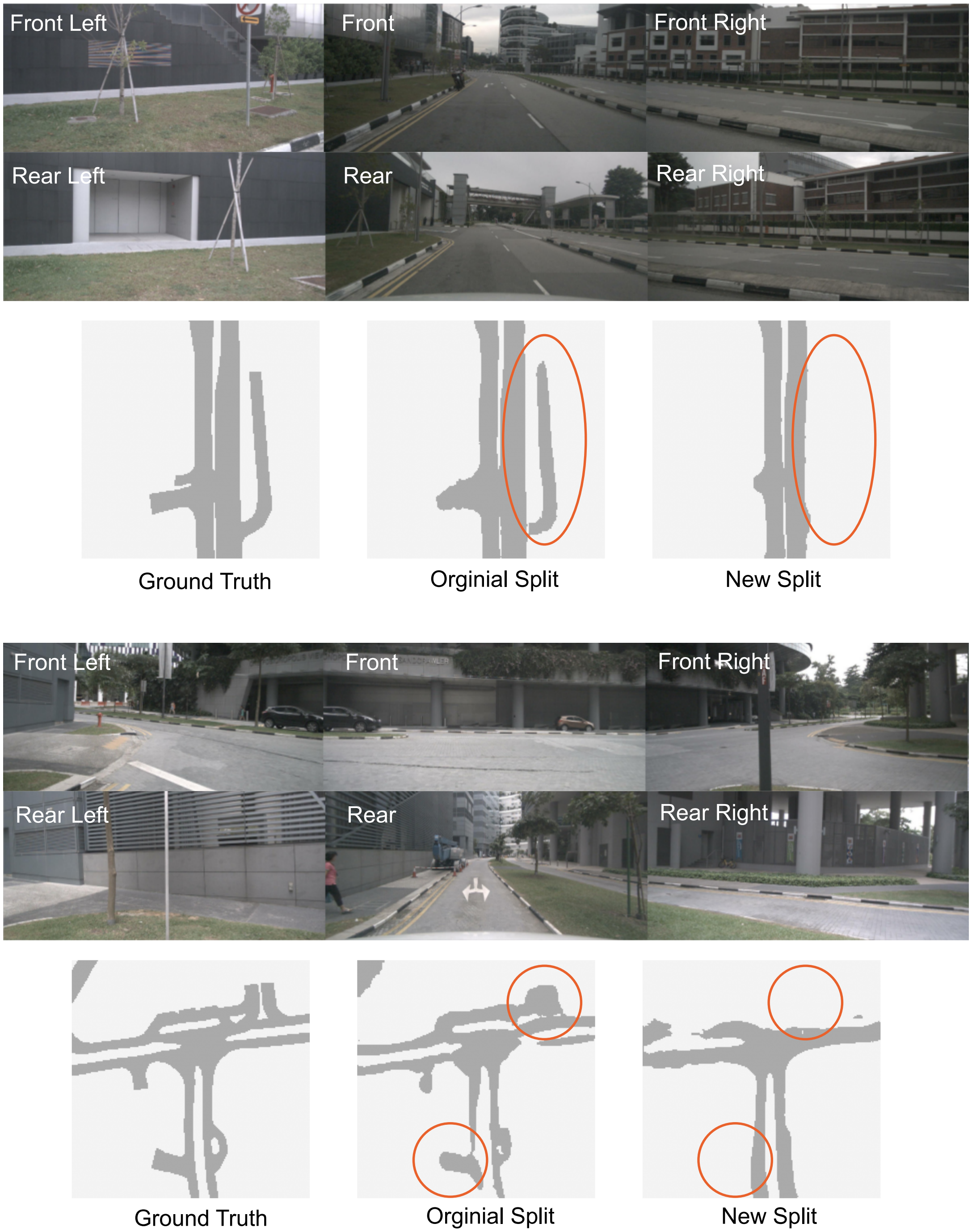

Figure 5. Additional visualizations of drivable area segmentation for networks trained on different splits. This figure shows further samples from the nuScenes validation datasets. Below each sample, we show the ground truth semantic mask (left), predictions of our network trained on the original split (middle) and on the disjoint split (right).

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice E. Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation withDeep Convolutional Nets, Atrous Convolution,and Fully Connected CRFs. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848, 2018. 1

[3] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2015. 1

[4] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 11006:369–386, 2019. 1

[5] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. of the Intl. Conf. on Machine Learning (ICML)*, 2019. 1

[6] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next Generation Datasets for Self-driving Perception and Forecasting. In *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2021. 1

[7] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2024. 3

[8] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1