Regularizing Image Encoders to Generate Bird's-Eye View Representations for Autonomous Driving Tasks

Qiaoyi Deng[1], Satoshi Ikehata[1,2], Yusuke Sekikawa[3], Ikuro Sato[1,3]

[1]Institute of Science Tokyo, Japan

[2]National Institute of Informatics

[3]Denso IT Laboratory

Abstract: Bird's-Eye View (BEV) representations are critical for providing a unified spatial scene understanding to autonomous driving tasks. However, existing methods often struggle with a lack of transformation equivariance. This results in artifacts on BEV feature maps that degrade the performance of downstream tasks. To address this issue, we propose a regularization approach to enhance transformation equivariance through ego-vehicle and dynamic object motion transformations by aligning BEV features in the global coordinate system across consecutive frames and introduces a consistency loss to penalize feature misalignment. Experiments on the nuScenes dataset demonstrate that the proposed approach effectively reduces artifacts, stabilizes BEV representations, and improves the reliability of downstream tasks.

Keywords: Autonomous Driving, Bird's-Eye View (BEV), Feature Representation, Temporal-Spatial Consistency, Transformation Equivariance, Regularization

1. Introduction

BEV (Bird's-Eye View) representation in autonomous driving provides an ego-vehicle-centered, 360-degree top-down view of the vehicle and its surroundings, enhancing perception and decision-making for safe and efficient autonomous systems. The surroundings are represented as grid cells containing high-dimensional feature values that encode spatial and semantic information about the scene. Since BEV feature maps are centered on the ego-vehicle, the origin of the BEV coordinate system shifts with the ego-vehicle's updated global position during rotation or translation. As a result, features surrounding the ego-vehicle must move consistently within the BEV map to maintain their relative positions to the ego-vehicle.

However, existing approaches [3, 4] encounter the issue for lack of transformation equivariance when generating BEV features, leading to noise-like artifacts on BEV feature maps shown as Figure 1. The radial patterns centered on the ego vehicle cause the surrounding features to appear blurred or distorted. These artifacts indicate the feature positions in the BEV coordinate system do not align with their actual physical locations. And we suppose that this issue will degrade the performance of downstream tasks.

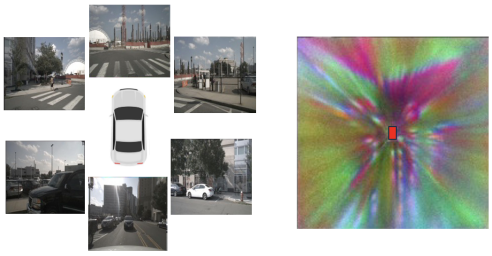

Figure 1. Legend of noise-like artifacts on BEV feature maps. The left panel displays raw data from six vehicle-mounted cameras, arranged as top-row (left-front, front, right-front) and bottom-row (right-rear, rear, left-rear) views centered on the ego vehicle. The right panel shows the BEV feature map generated by BEVFormer [3] visualized using PCA in RGB format. Red block in the middle indicates the ego-vehicle.

Going a step further, transformations occur over time as the ego-vehicle moves. To achieve transformation equivariance, the issue can be addressed by ensuring BEV feature alignment across consecutive frames (i.e. temporal consistency) under the ego-vehicle's motion transformations. Specially, for dynamic features, this alignment should consider both the ego-vehicle's motion and the objects' own motions, ensuring that features remain coherent and accurately represented in the global coordinate system over time.

To this end, we propose a method to improve temporal consistency for enhancing transformation equivariance by enforcing alignment of BEV features with the global coordinate system across frames through ego-vehicle and objects' motion transformations.

2. Method

We regularize the BEV feature maps to improve the temporal consistency for transformation equivariance in BEV representation. Our idea is applied to the partial pipeline of UniAD [2], as depicted in Figure 2, which employs the off-the-shelf BEV encoder from BEVFormer [3] to extract BEV features from multi-view images and then feeds them into decoders for downstream tasks. Unlike the original training pipeline of UniAD, the regularization term is calculated using the BEV features of consecutive frames and is added to the naïve loss functions for tracking and mapping [2].

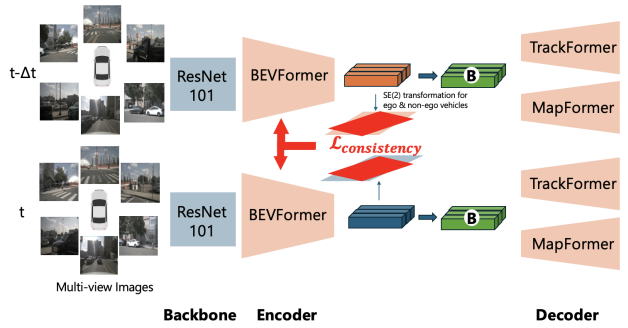

Figure 2. Pipeline for generating transformation equivariance preserved BEV features.

2.1. Data Pre-processing

The Figure 3 illustrates SE(2) transformations applied to align BEV features of both the ego-vehicle and dynamic objects in the global coordinate system. Raw images at $t - \Delta t$ are transformed to align with the global positions at time t.

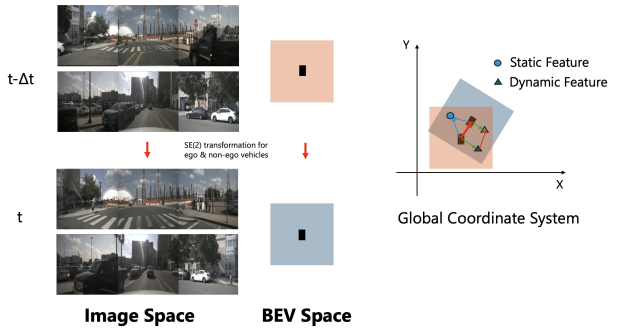

Figure 3. SE(2) transformations aligning BEV features of the ego-vehicle and dynamic objects in the global coordinate system.

Static Feature Alignment. Static features remain stationary in the global coordinate system. Let $\mathbf{P}$ denote the global position of a static feature. The ego-vehicle's motion between $t - \Delta t$ and t is represented by the SE(2) transformation matrix $\mathbf{E}_{\text{ego}}(t - \Delta t, t)$. The static feature's global position $\mathbf{P}$ remains unchanged:

$$\mathbf{P}' = \mathbf{P}.$$

The corresponding local positions in the ego-vehicle's reference frame are:

$$\mathbf{p} = \mathbf{E}_{\text{ego},t-\Delta t}^{-1} \cdot \mathbf{P}, \quad \mathbf{p}' = \mathbf{E}_{\text{ego},t}^{-1} \cdot \mathbf{P}'.$$

The static feature maps at $t - \Delta t$ and t are indexed by $\mathbf{p}$ and $\mathbf{p}'$, respectively:

$$F_{t-\Delta t}^{\text{static}}(\mathbf{p}) = F_{t-\Delta t}^{\text{static}}(\mathbf{E}_{\text{ego},t-\Delta t}^{-1} \cdot \mathbf{P}),$$

$$F_{t}^{\text{static}}(\mathbf{p}') = F_{t}^{\text{static}}(\mathbf{E}_{\text{ego},t}^{-1} \cdot \mathbf{P}').$$

The temporal alignment for static features can be expressed as:

$$F_{t-\Delta t}^{\text{static}}(\mathbf{p}) = F_{t}^{\text{static}}(\mathbf{p}')$$

ensuring spatial consistency in the global coordinate system.

Dynamic Feature Alignment. Dynamic features are influenced by both the ego-vehicle's motion and the motion of the dynamic object. Let $\mathbf{P}$ denote the global position of a dynamic object at time $t - \Delta t$, and $\mathbf{P}'$ denote its global position at time t. The position update is governed by the transformation matrix $\mathbf{T}_{\text{obj}}$, which describes the motion of the dynamic object in the global coordinate system:

$$\mathbf{P}' = \mathbf{T}_{\text{obj}} \cdot \mathbf{P}.$$

The transformation matrix $\mathbf{T}_{\text{obj}}$ is defined based on the object's pose at $t - \Delta t$ and t:

$$\mathbf{T}_{\text{obj}} = \mathbf{E}_{\text{obj},t} \cdot \mathbf{E}_{\text{obj},t-\Delta t}^{-1},$$

where $\mathbf{E}_{\text{obj},t}$ and $\mathbf{E}_{\text{obj},t-\Delta t}$ are the object's pose matrices in the global coordinate system at times t and $t - \Delta t$, respectively. The corresponding local positions of the dynamic object relative to the ego-vehicle's frame are given by:

$$\mathbf{p} = \mathbf{E}_{\text{ego},t-\Delta t}^{-1} \cdot \mathbf{P}, \quad \mathbf{p}' = \mathbf{E}_{\text{ego},t}^{-1} \cdot \mathbf{P}'.$$

The dynamic feature maps indexed by these positions are expressed as:

$$F_{t-\Delta t}^{\text{dynamic}}(\mathbf{p}) = F_{t-\Delta t}^{\text{dynamic}}(\mathbf{E}_{\text{ego},t-\Delta t}^{-1} \cdot \mathbf{P}),$$

$$F_{t}^{\text{dynamic}}(\mathbf{p}') = F_{t}^{\text{dynamic}}(\mathbf{E}_{\text{ego},t}^{-1} \cdot (\mathbf{T}_{\text{obj}} \cdot \mathbf{P})).$$

The temporal alignment for dynamic features can be expressed as:

$$F_{t-\Delta t}^{\text{dynamic}}(\mathbf{p}) = F_{t}^{\text{dynamic}}(\mathbf{p}')$$

ensuring spatial consistency in the global coordinate system. This process is applied to each dynamic object individually according to their unique transformation information.

2.2. Consistency Regularization

The consistency loss measures the discrepancy between the transformed features at time $t - \Delta t$ and the features at time t in the global coordinate system. The loss is computed over the overlapping region M_{overlap} between the two time steps and is expressed as:

$$\mathcal{L}_{\text{consistency}} = \sum_{\mathbf{p} \in M_{\text{overlap}}} \|F_{t-\Delta t}(\mathbf{p}) - F_t(\mathbf{p}')\|_2^2,$$

This formulation applies to both static and dynamic features.

3. Experiment

Experimental Setup. Our experiments are conducted on the nuScenes dataset [1], a large-scale benchmark for autonomous driving that includes 1,000 driving scenes with multi-view images captured by six cameras covering a 360° field of view. Each scene lasts approximately 20 seconds, with keyframes annotated at 2 Hz. The dataset provides annotations for ego-vehicle motion and 3D bounding boxes of 23 object categories.

The proposed method builds upon UniAD [2], a unified pipeline integrating perception, prediction, and planning tasks. Specifically, we focus on the perception task of 3D object detection, using the BEV encoder from BEVFormer [3] to transform multi-view image features into BEV features. The detection head is inherited from Deformable DETR [5], which predicts 3D bounding boxes based on the BEV features. Our pipeline incorporates the proposed consistency regularization during training, aiming to enhance the temporal stability and transformation equivariance of BEV feature maps.

Qualitative Results. Figure 4 demonstrates the impact of the proposed consistency regularization on BEV feature maps across consecutive frames ($t - \Delta t$ and t). In the naive model (left columns in each time step), noticeable artifacts appear in the BEV feature maps, particularly in the red-highlighted regions. These artifacts indicate instability in BEV feature alignment across frames, primarily caused by the loss of transformation equivariance. By incorporating the proposed consistency loss, the regularized model (right columns in each time step) produces BEV feature maps that are significantly smoother and more stable.

Quantitative Results. Table 1 compares consistency metrics between our regularized approach and the baseline UniAD. IDS (Identity Switches) measures ID mismatches, FRAG (Fragmentations) counts trajectory interruptions, TID (Track Initialization Duration) indicates the average time to initialize a track, and LGD (Longest Gap Duration) measures the longest time an object is lost. Lower values

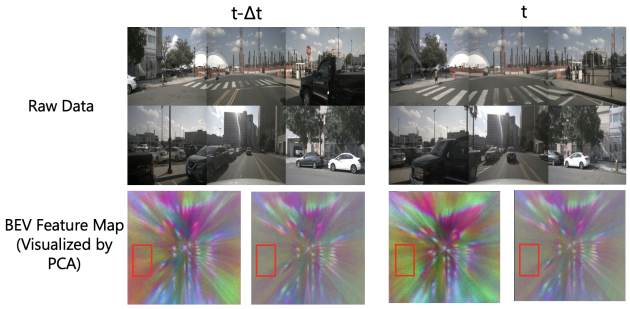

Figure 4. Each column in the figure corresponds to consecutive timestamps.

across these metrics indicate better tracking performance. Our method reduces IDS and FRAG, improving temporal consistency and trajectory continuity, while the decrease in TID reflects faster track initialization. Although LGD shows a slight increase, geometric consistency remains acceptable, demonstrating the effectiveness of our regularization in stabilizing BEV features and enhancing tracking performance.

Method	IDS↓	FRAG↓	TID(s)↓	LGD(s)↓
UniAD	768	1146	1.72	2.59
Ours	**720**	**998**	**1.60**	2.71

Table 1. Comparison of consistency evaluation metrics with naive method.

4. Future Work

While our method improves temporal consistency by incorporating transformation equivariance as a prior, we observed a slight trade-off with detection performance, as reflected in marginal declines in main metrics such as AMOTA and AMOTP. This highlights the need for more balanced regularization strategies that can simultaneously enhance temporal consistency and maintain high detection accuracy. Moreover, the development of BEV-specific data augmentation techniques tailored for end-to-end autonomous driving pipelines, particularly for planning tasks, remains a critical direction for practical deployment.

References

[1] H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[2] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, and L. Lu. UniAD: Planning-oriented

autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[3] H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, Z. Li, W. Wang, and J. Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3

[4] J. Philion and S. Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[5] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 3