Seq-BEV: Semantic Bird-Eye-View Map Generation in Full View Using Sequential Images for Autonomous Driving

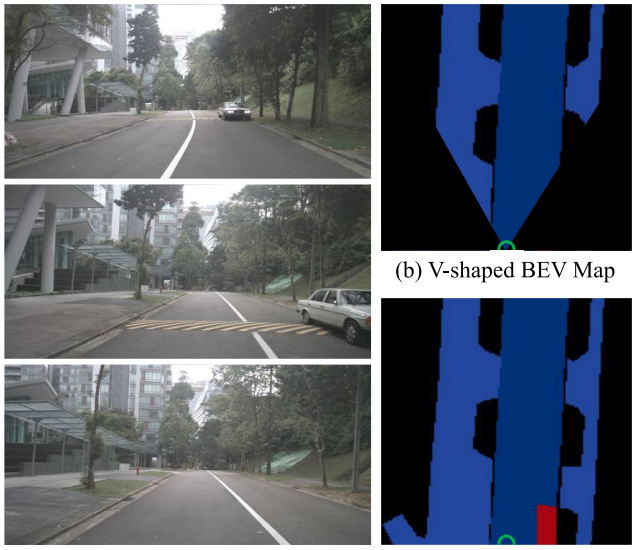
Shuang Gao, *Student Member, IEEE*, Qiang Wang, *Member, IEEE*, and Yuxiang Sun, *Member, IEEE*

***Abstract*— Semantic Bird-Eye-View (BEV) map is a straightforward data representation for environment perception. It can be used for downstream tasks, such as motion planning and trajectory prediction. However, taking as input a front-view image from a single camera, most existing methods can only provide V-shaped semantic BEV maps, which limits the field-of-view for the BEV maps. To provide a solution to this problem, we propose a novel end-to-end network to generate semantic BEV maps in full view by taking as input the equidistant sequential images. Specifically, we design a self-adapted sequence fusion module to fuse the features from different images in a distance sequence. In addition, a road-aware view transformation module is introduced to wrap the front-view feature map into BEV based on an attention mechanism. We also create a dataset with semantic labels in full BEV from the public nuScenes data. The experimental results demonstrate the effectiveness of our design and the superiority over the state-of-the-art methods.**

***Index Terms*— Semantic BEV maps, sequential images, view transformation, autonomous driving.**

I. Introduction

SEMANTIC scene understanding is a fundamental component for autonomous driving. Suitable formats of data representation for semantic scene understanding could facilitate downstream tasks, such as motion planning [1], [2], [3], [4] and trajectory prediction [5], [6]. Compared with semantic segmentation in front view, semantic maps in bird-eye-view (BEV) are more appropriate for the downstream tasks in autonomous driving due to the following reasons: 1) the distances and geometric relationships between the ego-vehicle and other road users can be explicitly indicated; 2) semantic BEV maps have no distortions that appear on front-view semantic segmentation maps. For example, the same object keeps the same size no matter how far the object is from the camera; 3) semantic BEV maps are high-level abstractions of the surrounding environment. So, using such maps to train control networks for autonomous driving in simulation environments, like CARLA [7], could alleviate the domain gap issue when deploying the networks in the real world.



(a) Equidistant Sequence Images (b) V-shaped BEV Map (c) Full BEV Map

Fig. 1. Comparison between V-shaped semantic BEV map and full BEV map. The corresponding equidistant sequence of RGB images is shown in (*a*), where three image frames, F_{i-2}, F_{i-1}, and F_i, are displayed from top to bottom. (*b*) is a V-shaped BEV map generated from the last frame, F_i, which is limited by the camera FOV. We can see that the car on the right side of the ego-vehicle cannot be seen in the V-shaped BEV map. (*c*) is the BEV map in the full view that enlarges the FOV by utilizing the supplementary information captured from the other two frames. Note that the green half-circle represents the position of the ego-vehicle on the BEV maps. The different colors represent different semantic classes.

To generate semantic BEV maps using front-view images, traditional methods involve a number of algorithms, such as geometric projection and semantic segmentation. Recently, deep learning-based end-to-end methods have shown great potential [8], [9], [10], [11], [12]. However, most existing methods that take as input the front-view image from a single camera suffer from the limited Field of View (FOV) issue. As illustrated in Fig. 1, due to the lack of visual information that is outside the FOV of the front-view camera, the BEV

Received 17 September 2023; revised 22 October 2024 and 10 April 2025; accepted 14 May 2025. Date of publication 25 June 2025; date of current version 6 August 2025. This work was supported in part by Hong Kong Research Grants Council under Grant 15222523, in part by the National Natural Science Foundation of China under Grant 61876054, and in part by the City University of Hong Kong under Grant 9610675. The Associate Editor for this article was Q. Wang. *(Corresponding author: Shuang Gao.)*

Shuang Gao is with the Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, China, and also with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: shuang.gao@connect.polyu.hk).

Qiang Wang is with the Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: wangqiang@hit.edu.cn).

Yuxiang Sun is with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong (e-mail: yx.sun@cityu.edu.hk).

Digital Object Identifier 10.1109/TITS.2025.3578070

map generated from a single front-view image is limited by the camera FOV, leading to a V-shaped BEV map, in which a large portion of the map is occupied by invalid pixels, leading to incomplete use of the map. In contrast, the full-view BEV map provides a wider FoV and all of the pixels are utilized. To get a full-view map, some works [13], [14], [15], [16] use images from multiple cameras around the ego-vehicle. However, those methods usually need to simultaneously process 6 or more images, which increases time and computing costs. Furthermore, our method enhances robustness by relying solely on a front-view camera, ensuring a complete front-view BEV map even if side cameras fail. This resilience benefits real-world deployment, where hardware failures or occlusions may occur. Additionally, multi-camera systems require precise calibration, which can degrade over time due to vibrations and temperature changes, leading to potential performance degradation. Furthermore, our single-camera approach is more cost-effective than multi-camera setups. There are also some attempts [16], [17] using the images from a temporal sequence to get more views of the environment. But in practice, the ego-vehicle could slow down or stop for a while on roads to wait for pedestrians or traffic lights. In such cases, the camera may repeatedly capture redundant images for the same scene, and existing methods could fail to get a larger view or full view when receiving these redundant images.

To provide a solution to this issue, we propose a novel network, Seq-BEV, which takes as input the equidistant sequential images and directly outputs semantic BEV maps in full view (i.e., 180° field-of-view). Equidistant sequential images are defined as those sampled at consistent distance intervals based on the ego vehicle's travel distance rather than temporal intervals. Our network is end-to-end trainable. It is composed of a sequence fusion module and a road-aware view transformation module. The former fuses the equidistant sequential images during the feature extraction process, and the latter transforms the front-view features into BEV with an attention mechanism.

To the best of our knowledge, our network is the first solution to use equidistant sequential images to get a semantic BEV map in full view. To train and evaluate the proposed network, we create a dataset with semantic BEV map labels in full view from the nuScenes dataset [18]. The experimental results demonstrate our effectiveness and superiority. The contributions of this work are summarized as follows:

1) We propose a novel semantic BEV map generation network that takes as input a set of equidistant sequential images sampled at uniform distance intervals and outputs a semantic BEV map in full view.
2) We design a new self-adapted sequence fusion module to fuse the features from different images, which provides complementary information to get more views.
3) We provide a new method for view transformation by first extracting the attention of road planes and then projecting attention-based features to BEV.
4) We create a dataset with semantic ground-truth labels in full view from the nuScenes dataset to train and test our method. Our code and dataset are publicly available.[1]

[1]Our code and dataset: https://github.com/lab-sun/Seq-BEV

The remainder of this paper is structured as follows. Section II reviews related work. Section III describes our network in detail. The experimental results are discussed in Section IV. Finally, we conclude our work and discuss several promising research directions in Section V.

II. Related Work

A. Semantic Scene Understanding in Front View

Most research work in semantic scene understanding focuses on classifying each pixel in an image captured by a front-view camera into individual classes. These works are known as semantic image segmentation. The Fully Convolutional Network (FCN) [19] is the milestone for deep learning-based semantic segmentation, which introduces the encoder-decoder structure. However, the spatial resolution of the feature map is reduced dramatically by the encoder, posing a risk of information loss. Zhao et al. [20] designed a pyramid pooling module in Pyramid Scene Parsing Network (PSPNet) to exploit the global context information. Different from PSPNet, DeepLab family [21], [22] uses atrous convolution to capture larger-scale context information with a relatively low amount of parameters. Recently, attention-based methods, such as Vision Transformer (ViT) [23], have been widely investigated in semantic segmentation because the attention operation could also increase the receptive field and capture the global salient feature. Based on ViT, Strudel et al. [24] proposed Segmenter, which is capable of modeling global context at the early stage of the network. Cheng et al. [25] designed Mask2Former, which is a generic framework for addressing segmentation tasks with masked attention. Segment Anything Model (SAM) [26] is proposed as a foundational model increasing the generalization of segmentation. Besides using only RGB images, there are also semantic segmentation methods using multi-modal images [27], [28], [29], [30], [31], [32], [33], such as RGB-Depth or RGB-Thermal images.

B. Semantic Scene Understanding in BEV

Different from semantic image segmentation, producing semantic BEV maps is a generation task. Lu et al. [8] proposed VED, conducting the view transformation through the variational encoder–decoder, but this model loses the spatial information due to the bottleneck of the network. Reiher et al. [34] employed a method that integrates information from cameras oriented in various directions to generate a complete 360° BEV map directly from semantically segmented images. View Parsing Network (VPN) [10] makes use of Multilayer Perceptron (MLP) to convert the view and cope with the shortage of the semantic BEV dataset by using the simulation environment. Roddick and Cipolla [13] made use of the internal and external parameters of the camera in the PON and projected the front view into the bird-eye view with the designed dense transformer layers. The rise of the Transformer framework provides a new insight into this view-transforming problem. TIM [35] approached this problem as a sequence-to-sequence prediction task, using cross-attention to transform image columns to BEV polar rays. Gong et al. [36], based on

the TIM, proposed GitNet and exploited the geometric prior in their method. Zhao et al. [37] proposed TaDe, decomposing the BEV map generation into a BEV understanding stage and a view transformation stage.

Some 3-D detection tasks also involve view transformation. For example, based on the Transformer framework, BEVFormer [17] introduces spatiotemporal transformer, fusing the information in the temporal sequential images from 6 cameras facing different directions. Later, the authors improved the network and developed BEVFormer V2 [16] with perspective space supervision. Li et al. [38] proposed BEVDepth, leveraging explicit depth information to supervise the view transformation.

C. Sequence Fusion

We aim to predict a BEV semantic map in full view from the equidistant sequential images. However, most of the related literature studies on the temporal sequence, so temporal sequence fusion methods are also reviewed. Using 2-D Convolutional Neural Networks (CNN) is a common way to process images. Some methods utilize the 2-D CNN to extract the spatial features of each individual image in a sequence and then conduct the temporal connection in the early-fusion or late-fusion step [39], [40]. Although the 2-D CNN achieves great success in image processing, these methods fail to fuse the low-level features during feature extraction and perform poorly in temporal modeling. Compared to the 2-D convolution, 3-D CNN is suitable for simultaneously learning spatial and temporal features. Ying et al. [41] proposed the D3Dnet, using the deformable 3-D convolution to incorporate spatio-temporal information for video super-resolution. However, the network size of 3-D CNN methods is larger than the 2-D counterparts, making it computationally inefficient and prone to over-fitting [42]. Long Short-Term Memory (LSTM) [43] is a commonly used method in some video-analysis work [44], [45] to aggregate long-term memory in a sequence, but suffer from a slow computation speed.

To efficiently model the sequence input in both spatial and temporal dimensions, Lin et al. [46] designed Temporal Shift Module (TSM), which shifts the information along the temporal dimension in the process of feature extraction. Mixed Temporal Convolutional kernels (MixTConv) [47] combined the concept of depthwise convolution and TSM and got an impressive result in action recognition. Later, Yang et al. [48] integrated the modified TSM into the transformer framework for video instance segmentation.

D. Difference From Previous Works

In this work, we propose an end-to-end road-aware semantic BEV generation network, Seq-BEV, that takes as input equidistant sequential images and outputs a semantic BEV map in full view.

Our network differs from previous works in three aspects: 1) The implementation of the view transformation in our Seq-BEV is more straightforward because we integrate the attention mechanism and the learnable Inverse Perspective Mapping (IPM) algorithm [49], which is in line with the traditional geometric view transformation pipeline; 2) As more environmental information is required to produce a full-BEV map, we resort to equidistant sequential images that are uniform in distance to complement the FOV limit. To process the equidistant sequential images, we fuse the sequential information from the individual image by shifting the channels in the sequential dimension, and we design a self-adapted shifting algorithm to make the sequence fusion suit for the training process; 3) We adopt a two-stream structure to extract the spatial and sequential features separately. This two-stream structure ensures the completeness of the spatial and sequential information.

III. The Proposed Network

A. The Overall Architecture

Our motivation is to generate semantic BEV maps in full view using sequential images that are sampled at equal distances. The structure of our network is illustrated in Fig. 2. As we can see, our network has two streams that respectively take as input single and sequential images at the same time. The two kinds of inputs are respectively fed into the spatial and sequential encoders, where the low-level features and high-level features are extracted at different stages of the encoder.

To complement the vision information from the equidistant sequential images and get the full view, we design a self-adapted sequence fusion module in the sequential encoder. This is a parameter-free module that can directly manipulate the feature tensor. The spatial low-level feature and the sequential low-level feature are fused into the S&S fusion feature via convolution operation. The S&S refers to spatial and sequential. To be specific, the feature maps extracted from the spatial and sequential encoders are first concatenated along the channel dimension. To effectively fuse the spatial and sequential low-level features while preserving their original size, a 3×3 convolution is then applied. This operation enhances the integration of spatial and temporal information, contributing to more effective feature fusion. The road-aware view transformation module takes as input the S&S fusion feature and computes the road layout attention under explicit supervision before projecting it into the road BEV feature. We concatenate the high-level features and the road BEV feature, producing the fused feature, and then send it to the decoder to get the semantic BEV map in full view.

B. The Two-Stream Encoder

We design a two-stream encoder to extract the spatial and sequential features, respectively. In such a way, the spatial integrity of the images can be preserved when performing the equidistant sequence fusion. The structures of the spatial and sequential encoders are similar. The only difference is that there is a sequence fusion module in the sequential encoder. So, in the following text, we do not particularly distinguish the spatial encoder from the sequential encoder and briefly term both of them as encoder.

We use the DeepLab V3+ [22] to extract the features and choose MobileNet V2 [50] as the backbone. MobileNet V2 is a lightweight network that requires relatively few computation

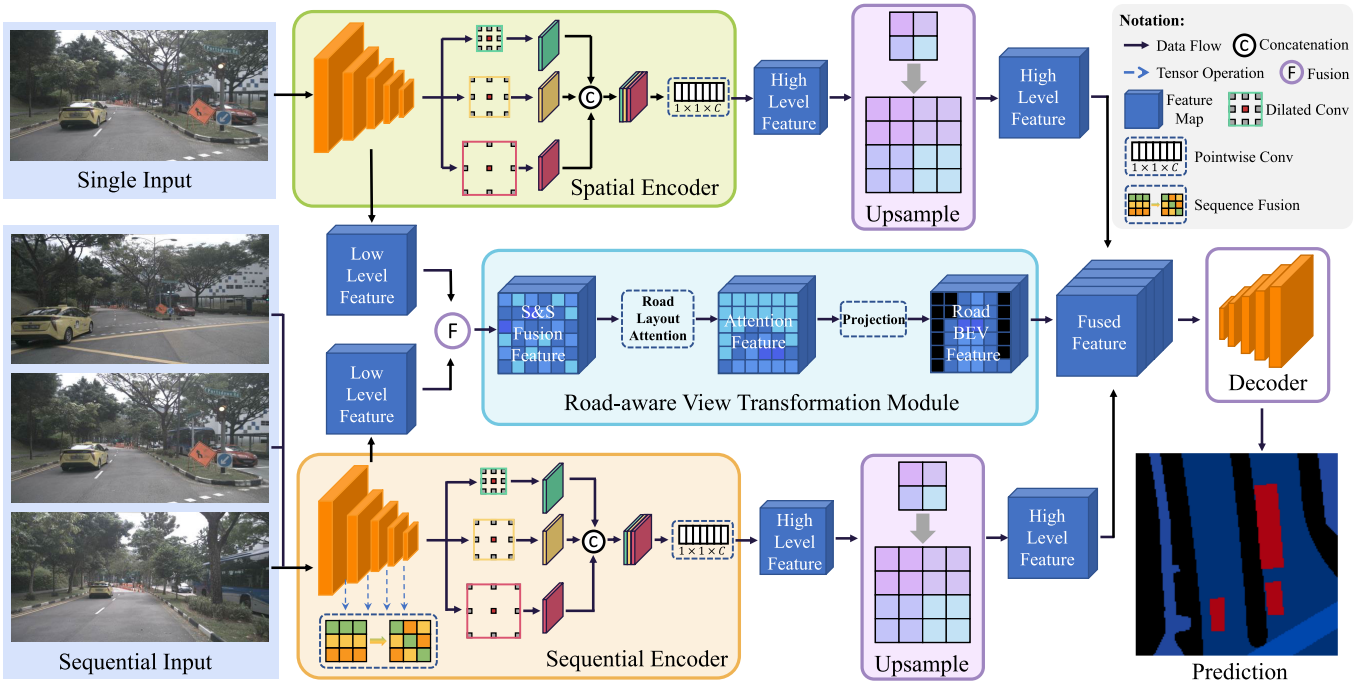


Fig. 2. The overall architecture of our proposed Seq-BEV network. It mainly consists of a spatial encoder, a sequential encoder, and a road-aware view transformation module. We feed the single and equidistant sequential images into the spatial encoder and sequential encoder, respectively. In both the encoders, we take out the low-level features and high-level features. The former encodes the structural features, which are fed into the road-aware view transformation module as the input to find the road layout attention. The latter encodes the higher-level semantic information generated by a set of dilated convolutions with various dilation rates. In addition, the sequence fusion module in the sequential encoder fuses the information from different images. The road-aware view transformation module first extracts the road layout attention and then projects the attention-contained feature into the bird-eye view to get the road BEV feature. Finally, the high-level features and road BEV feature are concatenated as the fused feature, which is fed into the decoder to recover the map resolution.

resources. In the encoder, we take out the low-level features $\mathcal{F}_L$ from the low stage of the backbone and the high-level features $\mathcal{F}_H$ from the high stage of the backbone. The low-level features keep the higher resolution and richer spatial information but contain relatively less semantic information. The spatial information covered in the low-level features, such as geometrical structure, could be helpful for our road layout attention extraction.

In contrast, the high-level features encode more semantic information, and these high-level features could be invariant in scale because they pass through the multi-scale dilated convolutions. In the front-view image, the same-sized objects may have various scales due to the different distances from the camera. So, the high-level features would be more suitable for sensing objects on roads. It should be noted that the low- and high-level features are extracted from the $3rd$ and $17th$ layers of the backbone, respectively. The $3rd$ layer corresponds to the one following the first downsampling operation, while the $17th$ layer represents the final layer of the feature extractor.

C. The Sequence Fusion Module

We design a self-adapted sequence fusion module to fuse the complementary information from the equidistant sequential images and get a semantic BEV map in full view. This module fuses the sequential features by applying varying degrees of grouping and shifting operations, based on the number of training iterations. The input and output sizes (i.e., channel and resolution) of the self-adapted sequence fusion module are the same, thus, it can be inserted into existing networks seamlessly.

The self-adapted sequence fusion module is shown in Fig. 3. The sequential feature is a 4 dimension tensor, which is denoted as $\mathcal{F}_{seq} \in \mathbb{R}^{B\times S\times C\times H\times W}$, where B, S, C, H, W are the batch size, numbers of the images in the sequence, number of channels, height, and width, respectively. The example provided shows a sequence of three frames, with different colors representing each one. The fusing operation is inspired by TSM [46]. The channel dimension is divided into groups, each containing a subset of the image features. As shown in Fig. 3, we assign two channels per group for demonstration. The groups are then shifted according to a defined principle to fuse the features across frames. However, TSM relies on a fixed-size grouping (e.g., $1/8$ of the channel dimension), where the group size is a pre-defined parameter. Selecting an inappropriate group size can result in suboptimal outcomes.

In experimentation, we observed that as the network training progresses, its feature extraction capacity improves, and a larger shift portion becomes necessary for more effective fusion. In response to this observation, we introduced a self-adaptive mechanism to the sequence fusion process. First, we divide the channel dimension into n groups $\{G^1, G^2, \ldots, G^n\}$. The proposed self-adapted fusion method dynamically adjusts the number of channels within each group according to the training iterations. The number of channels

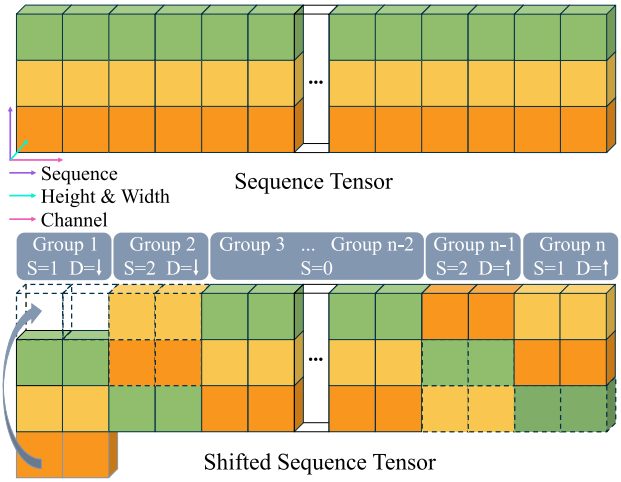


Fig. 3. The demonstration of the sequence fusion module. This module helps the network fuse the equidistant sequential information without any additional trainable parameters. The different sequence features are represented in different colors. Here, we assign two channels in each group as an example and move the groups with varying strides in either upward or downward directions. Taking the first group as an example, the vacancy due to the shifted operation is padded with the out-shifted elements. In practice, the number of channels of a group is determined by the current training iterations in a self-adapted manner.

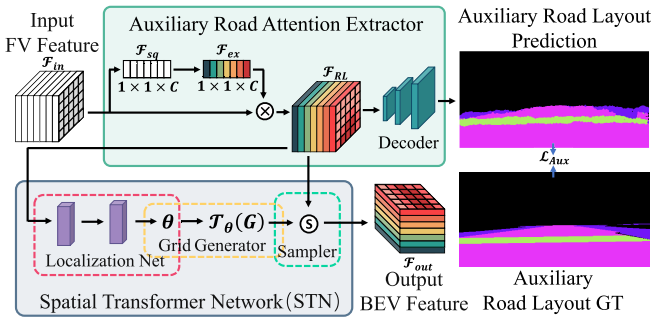


Fig. 4. The pipeline of the road-aware view transformation module. The module takes as input the front view low-level feature maps from both spatial and sequential encoders. The road attention is extracted under the auxiliary supervision to emphasize the road plane before performing the spatial transformation. The STN regresses a projection matrix and then wraps the road-attention-contained feature into the BEV feature. Conducting the view transformation with the road attention mechanism could alleviate the limitation caused by the flat road assumption.

in each group is determined by the following calculation:

$$\alpha = 1/\min\{\lfloor C/N \rfloor, \max(\lfloor \lambda I_t / I_c \rfloor, 1)\}, \tag{1}$$

$$m = \alpha C/N, \tag{2}$$

where α is the grouping coefficient, which is a dynamic parameter intended to adjust the channel number in a group according to the current number of iterations. N and C are the minimum number of groups and the total channel number of the current feature tensor, respectively. The total number and the current number of the iterations are denoted as I_t and I_c, respectively. λ is a parameter that indicates the reliability of the network feature extraction capability. The larger λ is, the fewer channels are assigned to each group. Here, we set λ as 0.5 empirically. $\lfloor \cdot \rfloor$ represents the round down operation. m is the current channel number in a group.

As illustrated in Fig. 3, after determining m, each group is shifted by varying strides in either the upward or downward direction. This module does not consume extra computation costs because the self-adapted grouping and shifting operations need no learnable parameters. We conduct the sequence fusion at the $2nd$, $4th$, $7th$, and $14th$ layer of the backbone network before the downsampling operations. It is worth noting that the shifting operation exchanges the channels across different frames, disrupting the spatial integrity of individual frames. To address this, we employ a spatial encoder to extract the feature map from each image, enhancing spatial modeling capabilities.

D. The Road-Aware View Transformation Module

Most current view transformation approaches employ data-driven methods to generate BEV maps, relying on the complex mapping relationships learned by deep neural networks. However, this process often lacks interpretability. To achieve a more reliable and explainable view transformation, we project front-view features onto the BEV plane using a learnable homography transformation, which is applied after extracting attention from the road surface. The use of a learnable homography enhances the interpretability of the transformation while incorporating road-aware features helps mitigate the distortion commonly associated with the flat-ground assumption [51].

To acquire an accurate road layout during the view transformation, we conduct a road attention extraction in an auxiliary supervision manner before projecting the front view feature into the bird-eye view. As shown in Fig. 4, the road-aware view transformation module consists of the auxiliary road attention extractor and the learning-based Spatial Transformer Network (STN).

The auxiliary road attention extractor employs SENet [52] to emphasize the informative components in the feature map. The input feature $\mathcal{F}_{in}$, which is produced by the low-level encoders, is fed into the extractor, and through a squeeze operation compresses the global spatial information into a $1 \times 1 \times C$ feature $\mathcal{F}_{sq}$, where C refers to the number of channels. The following excitation operation captures the channel-wise relations in $\mathcal{F}_{sq}$ via two fully connected (FC) layers and generates $\mathcal{F}_{ex}$. $\mathcal{F}_{ex}$ can be seen as a set of channel weights that indicates the salient features with a high score. Finally, the informative components are selected by multiplying $\mathcal{F}_{in}$ and $\mathcal{F}_{ex}$. The above steps can be formulated as:

$$\mathcal{F}_{sq} = \frac{1}{H' \times W'} \sum_{u=1}^{H'} \sum_{v=1}^{W'} \mathcal{F}_{in}(u, v), \tag{3}$$

$$\mathcal{F}_{ex} = FC(\mathcal{F}_{sq}, \mathbf{W}), \tag{4}$$

$$\mathcal{F}_{RL} = \mathcal{F}_{in} \otimes \mathcal{F}_{ex}, \tag{5}$$

where H' and W' are the height and width of the input feature map $\mathcal{F}_{in}$. The fully-connected operation is denoted as $FC(\cdot)$ and $\mathbf{W}$ is the learnable parameter. $\otimes$ represents element-wise multiplication. Fig. 5 qualitatively demonstrates sample salient features extracted by the attention mechanism. To get a more reliable road layout segmentation, we conduct this attention extraction under the auxiliary road layout supervision.

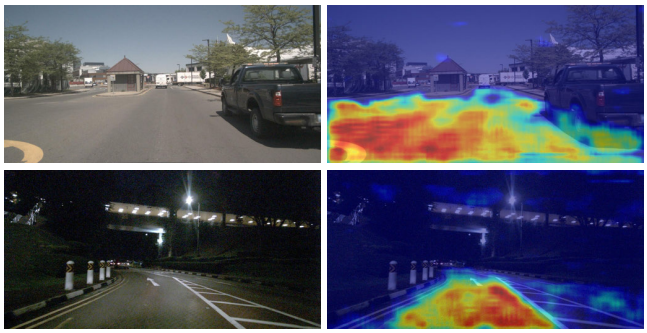


Fig. 5. Qualitative demonstrations of the attention extracted by the road attention extractor. (a) original images; (b) class activation map (CAM) visualization for the *Road* class. The region with warmer colors corresponds to the more attention-focused area.

The view transformation is implemented on the feature map $\mathcal{F}_{RL}$, which encodes the road plane attention. The STN [49] is employed to regress a 3×3 projection matrix θ via the localization net. Then, with the projection matrix, the grid generator creates a sampling grid before sending it to the sampler. The sampler samples $\mathcal{F}_{RL}$ at the sampling grid points. We refer readers to STN [49] for more details. Usually, the geometric projection methods suffer from the flat ground assumption, leading to distortions for objects above roads or distortions for roads that are not flat. But our proposed view transformation method focuses on the road plane through the attention mechanism before the projection, which alleviates the limitation of the flat road assumption.

E. Loss Functions

We use two losses in this work. One is the auxiliary loss $\mathcal{L}_{Aux}$, which supervises the road attention extractor. The other is the BEV loss $\mathcal{L}_{BEV}$, which enables the network to produce semantic BEV maps. Due to the class imbalance issue in the dataset, the Focal Loss [53] is used as $\mathcal{L}_{Aux}$ and $\mathcal{L}_{BEV}$. We train our network in an end-to-end manner. The overall loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{BEV} + \gamma \mathcal{L}_{Aux}, \tag{6}$$

where γ is the weighting parameter to balance the two losses. We empirically set γ as 1.

IV. Experimental Results and Discussions

A. The Dataset

We create our dataset from nuScenes [18], which includes 850 annotated scenes with 3-D object bounding boxes, semantic high-definition (HD) maps, etc. The ground-truth labels of our dataset include the semantic BEV maps and the road layout labels. The former is generated by projecting the 3-D bounding boxes to the HD map. The latter is created by overlaying the HD map segment on the front view image to get the road layout labels.

Compared to the other nuScenes-based BEV segmentation dataset, the one we created breaks the V-shape view limitation and is easier to load, saving much loading time to the network.

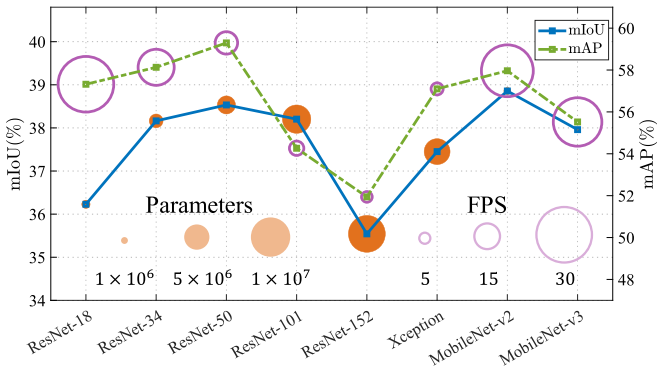


Fig. 6. Impacts caused by backbone selection in terms of mIoU and mAP. The blue solid line indicates the results measured by mIoU, while the green dotted line corresponds to the mAP measurements. In addition, the area of the solid orange circle reflects the number of parameters within the network for various backbone architectures. The area of the hollow purple circle represents the FPS performance of each respective backbone.

Note that the projections may not be correct if the ground is uneven. We divide the labeled 850 scenes into 548 training sets, 150 validation sets, and 148 test sets, excluding some incorrectly projected scenes (e.g., uneven roads). To achieve equal-distance sampling, three images with equal intervals are sampled using the ego-pose information provided by the dataset to construct a sequence. We set three dataset groups with different distance intervals for the experiment. Specifically, each group with the sequence comprises three images acquired at distances of 10, 20, and 30 meters, respectively, or at angular displacements exceeding 30°. The input images are normalized to the resolution of 256×512. The output semantic BEV grid maps are with the resolution of 150×150, each pixel encodes an area of $0.2 \times 0.2\ m^2$.

B. Training Details

We train our network with NVIDIA GeForce RTX 3090. To balance the memory consumption and the time cost, we set the batch size to 8. We train our network for 50 epochs using the AdamW optimizer. We initialize the learning rate as 5×10^{-4} and adopt the cosine annealing scheme [54] to adjust the learning rate during training. The warm-up strategy is employed for the learning rate adjustment. This strategy gradually increases the learning rate until the preset epoch for warm up ends (the $20th$ epoch in our network), and then decreases the learning rate according to the decay scheme. The momentum and weight decay are set to 0.9 and 5×10^{-4}, respectively. MobileNet V2 is used as our backbone and initialized with the pre-trained weight. The rest of our network parameters are initialized randomly. The sequence fusion is inserted into the $2nd$, $4th$, $7th$, and $14th$ block of the backbone, and the minimum number of groups for the self-adapt grouping N is set to 24.

C. Ablation Study

To verify the effectiveness of our network structure and to choose appropriate parameters for our network, we conduct several ablation experiments. The mean Intersection over

TABLE I

THE ABLATION STUDY RESULTS (%) ON DIFFERENT VARIANTS. THERE ARE TWO GROUPS OF TESTS, WHOSE INPUT IS THE SINGLE IMAGE (SGL) AND THE SEQUENTIAL IMAGES (SEQ), RESPECTIVELY. IN EACH GROUP, WE COMPARE THE RESULTS FROM THE THREE VARIANTS, WHICH ARE THE SEMANTIC SEGMENTATION BASELINE METHOD, THE PLAIN VIEW TRANSFORMATION VARIANT, AND THE ROAD-AWARE ONE. THOSE VARIANTS ARE DENOTED AS BASELINE, PLVT, AND RAVT IN THIS TABLE. THE BOLD FONT HIGHLIGHTS THE BEST RESULTS IN EACH GROUP

Variants	Background		Drivable Area		Ped. Crossing		WalkWay		Obstacle		Vehicle		Pedestrain		mIoU	mAP
	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP		
SGL-Baseline	56.36	69.67	64.67	76.91	16.75	34.73	31.54	52.64	10.67	23.27	15.85	36.18	**1.08**	**4.55**	28.13	42.56
SGL-PLVT	56.84	70.22	65.38	**77.47**	18.04	41.36	32.27	**53.85**	**11.26**	**36.24**	**16.52**	33.38	0.94	3.89	**28.75**	45.20
SGL-RAVT	**57.17**	**71.40**	**66.16**	77.28	**18.50**	**45.04**	**32.69**	52.33	11.09	36.13	15.07	**36.90**	0.43	2.71	28.73	**45.97**
SEQ-Baseline	62.96	73.98	70.86	82.51	26.28	57.08	39.31	58.75	16.79	39.05	17.73	40.69	0.51	4.57	33.49	50.95
SEQ-PLVT	62.36	74.45	70.82	81.25	27.02	56.37	39.28	60.13	17.74	40.68	17.68	42.94	0.81	3.49	33.67	51.33
SEQ-RAVT(Ours)	**66.59**	**76.85**	**74.89**	**85.19**	**35.76**	**65.27**	**43.73**	**64.72**	**22.27**	**52.43**	**27.63**	**54.32**	**1.12**	**7.24**	**38.86**	**57.96**

Union (mIoU) and the mean Average Precision (mAP) are employed to quantitatively evaluate the performance of our network.

1) Ablation on Backbone: We compare the performance of our network with different backbones in the encoders, including MobileNet V2 [50], MobileNet V3 [55], ResNet family [56], and Xception [57]. Similar to our proposed Seq-BEV, we modify each backbone to get the low-level feature and the high-level feature and also insert the multiple sequence fusion modules into the backbones.

Fig. 6 demonstrates the results, which shows the trade-off between the network performance and the number of parameters. The network runtime is assessed in terms of Frames Per Second (FPS) on the RTX 3090 GPU and represented visually by the hollow purple circle, whose area is inversely proportional to the number of network parameters. It is observed that MobileNet V2, which has the fewest parameters, achieves a frame rate of 28.04 FPS while delivering satisfactory performance. So, MobileNet V2 is chosen as our backbone.

2) Ablation on Different Variants: This ablation study is divided into two groups, which takes as input single (SGL) image and equidistant sequential (SEQ) images, respectively. Note that the self-adapted sequence fusion module is removed from the SGL group. For the first group, we employ the semantic segmentation network DeepLab V3+ [22] as the baseline. For the second group, we add the plain STN to the baseline as our view transformation module. We term this variant as PLVT. For the third group, we integrate the road-attention extractor with the view transformation to test the performance of the road-aware mechanism. We name this road-aware variant RAVT.

Tab. I displays the results. We can see that our proposed Seq-BEV (the SEQ-RAVT variant), which contains the sequential input fusion module and the road-aware view transformation module, achieves the best performance in terms of both mIoU and mAP. The data in the table leads to the conclusion that all the variants that take as input the sequential images get a superior performance against their counterparts. Comparing PLVT and RAVT, we can see that the prediction performance increases due to the incorporation of road attention.

3) Ablation on Sequence Fusion Module: We fuse the information of different images from a sequence by shifting the sequential channel in a feature map. Since this fusion module directly manipulates the feature tensor without any learnable parameters, it can be flexibly inserted into any position of a CNN. So, the insert position of this sequence fusion module needs to be chosen.

In our Seq-BEV, we select MobileNet V2 as our backbone network. The MobileNet V2 is stacked by the inverted residual blocks, which consist of an expansion layer, a depthwise layer, and a projection layer. In this ablation study, we test the performance of the Seq-BEV with the sequence fusion module inserted before different layers.

In addition, we design a self-adapted mechanism to group different channel numbers in the sequence dimension according to the training process. Here, we also compare the performance of our proposed self-adapted grouping strategy with that of the fixed grouping method. We set 3 fixed groups, dividing the channel dimension into 8, 16, and 24 groups, respectively.

Tab. III displays the results. From the table, we can see that the self-adapted grouping strategy improves the network performance in terms of mIoU. Moreover, the network achieves the best results when we insert the sequence fusion module before the depthwise layer. We conjecture the reason for this superior performance is the expansion operation in the expansion layer, which extends the dimension of the feature map. It can be seen as a process of data decompression. Thus, the feature map produced by the expansion layer could provide enough information for the sequence fusion.

4) Ablation on the Network Structure: In this ablation study, we first conduct experiments to determine the best way to combine the road BEV feature and the high-level features. Then, we adjust the input feature map of the road-aware view transformation module to choose the most effective Seq-BEV structure. In addition, an ablation study is conducted to determine the optimal loss weighting factor, denoted as γ, for appropriately balancing the BEV loss, $\mathcal{L}_{BEV}$, and the auxiliary loss, $\mathcal{L}_{Aux}$.

Element-wise addition and concatenation are two common ways to combine the separate features. We report the results

TABLE II

THE ABLATION STUDY RESULTS (%) ON THE COMBINATION METHOD OF THE ROAD BEV FEATURE AND THE HIGH-LEVEL FEATURE. WE APPLY ELEMENT-WISE ADDITION AND CONCATENATION TO COMBINE THE TWO FEATURES, RESPECTIVELY. IN ORDER TO MAINTAIN THE SAME CHANNEL SIZE OF THE FEATURE MAP PRODUCED BY THE SEPARATE METHODS, WE USE THE CONVOLUTION LAYER AFTER THE CONCATENATE OPERATION. THE BOLD FONT HIGHLIGHTS THE BEST RESULTS IN EACH COLUMN

Combine Method	Background		Drivable Area		Ped. Crossing		WalkWay		Obstacle		Vehicle		Pedestrain		mIoU	mAP
	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP		
Element-wise Add	65.21	75.48	73.83	83.41	30.07	64.89	42.38	61.86	17.36	**55.73**	17.87	**55.57**	0.25	**8.59**	35.28	57.93
Concat & Conv	**66.59**	**76.58**	**74.89**	**85.19**	**35.76**	**65.27**	**43.73**	**64.72**	**22.27**	52.43	**27.63**	54.32	**1.12**	7.24	**38.86**	**57.96**

TABLE III

THE ABLATION STUDY RESULTS (%) ON THE SEQUENCE FUSION MODULE. WE TEST THE NETWORKS WITH DIFFERENT INSERTION POSITIONS AND CHANNEL GROUPING SCHEMES AT THE SAME TIME. THE SEQUENCE FUSION MODULE IS RESPECTIVELY INSERTED BEFORE THE EXPANSION LAYER, DEPTHWISE LAYER, AND PROJECTION LAYER. TO TEST THE PERFORMANCE OF THE DESIGNED SELF-ADAPTED GROUPING MECHANISM, WE COMPARE IT WITH 3 FIXED GROUPING CONFIGURATIONS, CONSISTING OF 8, 16, AND 24 GROUPS

Grouping	Expansion Layer		Depthwise Layer		Projection Layer	
	mIoU	mAP	mIoU	mAP	mIoU	mAP
8-Groups	36.43	56.01	37.42	57.60	37.24	57.41
16-Groups	37.30	57.70	37.87	58.07	37.81	57.73
24-Groups	37.63	58.33	38.08	58.72	37.97	57.40
Self-adapted	37.75	56.93	38.86	57.96	38.58	57.44

of the network that adopts the two combination methods in Tab. II. Note that in order to keep the feature dimension unchanged, a convolution layer is applied after the concatenation operation. According to the results from Tab. II, the concatenation method gets the best performance in terms of mIoU and mAP. We also note that the element-wise addition method performs better in the segmentation of small objects on the road, like obstacles, vehicles, and pedestrians. The reason for this case may be that compared with the obvious road feature, those small object features become negligible ones during the convolution operation in the concatenation method, leading to inferior performance.

We use the low-level feature as the input of the road-aware view transformation module. The low-level feature maps preserve the high resolution and hence encode rich spatial information, such as geometrical structure, which may be suitable for road layout extraction. To verify this idea, we change the input of the road-aware view transformation module to the high-level feature or both the low-level and high-level features. Tab. IV shows the experiment results. We can see that the road-aware view transformation module conducted on the low-level feature has the higher mIoU with 38.86%, compared with the others. This is also true for the metric mAP. This result validates the applicability of low-level features to road attention extraction.

The entire network is trained under the supervision of both the BEV loss, $\mathcal{L}_{BEV}$, and an auxiliary loss, $\mathcal{L}_{Aux}$. A suitable weighting factor can balance the influence of these losses on the training process. Tab. V presents the network's performance across varying loss weights, based on which we assign a value of 1.0 to this factor.

5) Ablation on the Distance Intervals: The distance sequence employed in our network is designed to address blind spots resulting from the limited field of view. These distance intervals can be adjusted as long as the sequential images capture environmental details beyond the frame's visual range. To evaluate the effectiveness of the sequence fusion and determine the optimal distance configuration for processing inputs, we conducted an ablation study on various distance intervals. In this experiment, we utilized intervals of 10, 20, and 30 meters.

The experiment results are displayed in Tab. VI, demonstrating that the network achieves the highest performance when processing images at 10-meter intervals. These findings suggest that increasing the distance between images may lead to a decline in the accuracy of generating the full-view semantic BEV map, as larger intervals fail to capture the necessary environmental information in blind spots.

D. Comparative Results

1) The Quantitative Results: We evaluate the performance of our Seq-BEV with some state-of-the-art semantic BEV generation methods, including Cross-view Transformation (CVT) [58], Variational Encoder-Decoder Networks (VED) [8], MonoLayout [59], and View Parsing Network (VPN) [10]. Those methods originally take a single image as input. In order to compare them with our Seq-BEV, we modify those networks and enable them to predict the full BEV map with sequential images as well. To maintain the original network structure, we keep the original network unchanged to extract the spatial feature while we duplicate the encoders of those networks to fuse the sequential feature. Then, we feed the spatial and temporal features together into the decoder for the semantic BEV map generation. In addition, we also compare our Seq-BEV with some state-of-the-art BEV detection networks, such as BEVFormer [17], BEVdepth [38] and MatrixVT [60], by changing the detection head into the segmentation one. Keeping the original input settings as the same, the temporal sequential images from 6 vehicle-surrounding cameras are fed into those networks. However, the multi-view inputs slow down the training process. To trade off the computing resources and the training efficiency, we only use the BEVFormer-tiny for the comparison.

TABLE IV

THE ABLATION STUDY RESULTS (%) ON THE INPUT FEATURE OF THE ROAD-AWARE VIEW TRANSFORMATION MODULE. WE CONDUCT THIS EXPERIMENT BY SENDING THE HIGH-LEVEL FEATURE, LOW-LEVEL FEATURE, AND BOTH OF THEM TO THE ROAD LAYOUT ATTENTION EXTRACTOR. THE NETWORK THAT TAKES AS INPUT THE LOW-LEVEL FEATURE GETS THE BEST PERFORMANCE, WHICH IMPLIES THAT THE LOW-LEVEL FEATURE ENCODES RICH SPATIAL INFORMATION AND IS SUITABLE FOR THIS TASK. THE BOLD FONT HIGHLIGHTS THE BEST RESULTS IN EACH COLUMN

Input Feature	Background		Drivable Area		Ped. Crossing		WalkWay		Obstacle		Vehicle		Pedestrain		mIoU	mAP
	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP		
High-level	65.66	74.88	74.23	**85.62**	34.74	64.69	42.69	62.60	18.64	49.92	25.33	53.78	0.72	**9.35**	37.43	57.26
Low-level (Ours)	**66.59**	76.58	**74.89**	85.19	**35.76**	**65.27**	**43.73**	**64.72**	**22.27**	**52.43**	**27.63**	**54.32**	**1.12**	7.24	**38.86**	**57.96**
Both	65.78	**77.21**	74.38	84.31	34.28	62.00	43.40	63.34	19.50	49.98	27.62	52.82	0.96	7.48	37.99	56.73

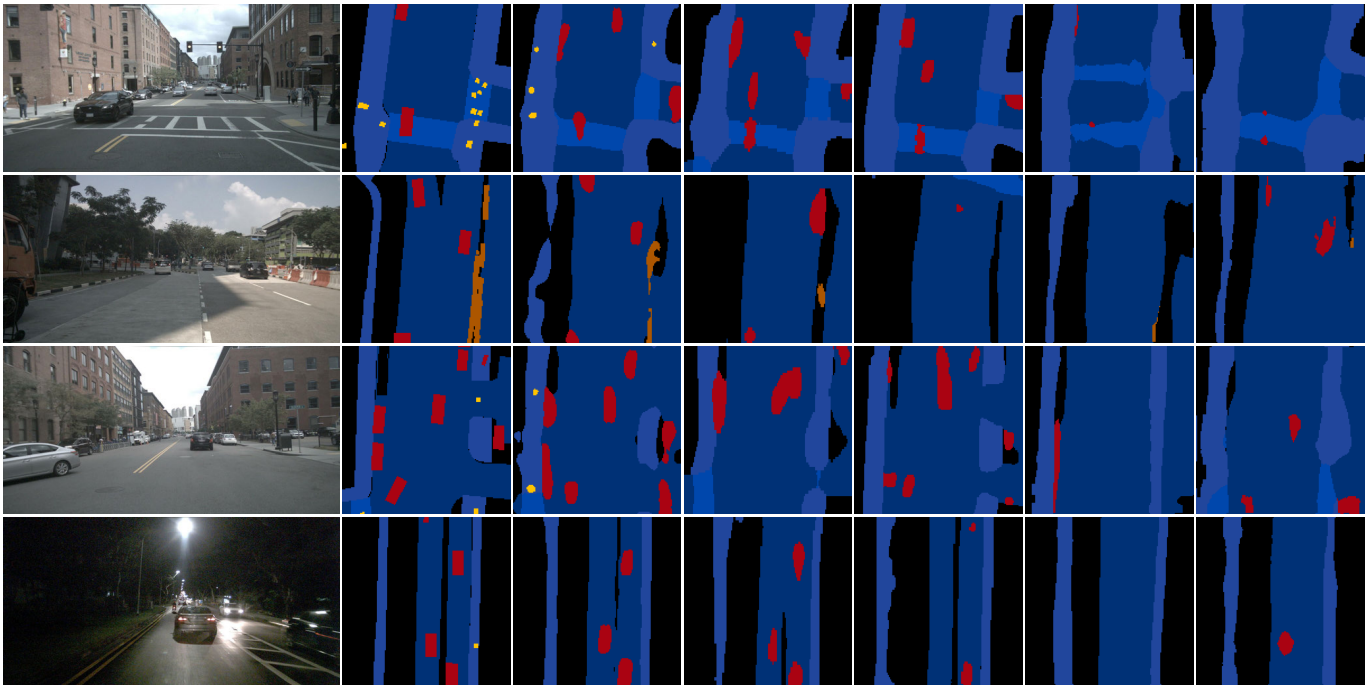


Fig. 7. Sample qualitative results for the full BEV semantic map generation networks. The results from the different networks testing with the same input are displayed in each row. Due to the space limit, we only show one image from a sequence here. The proposed Seq-BEV is more sensitive to small road objects and provides the more clear road layout. The comparative results demonstrate the superiority of our network.

TABLE V

THE ABLATION STUDY RESULTS (%) ON THE LOSS WEIGHT FACTORS γ. WE SET IT TO 0.5, 0.1, 1.0, 1.5, 2.0, AND 10.0. THE BOLD FONT HIGHLIGHTS THE BEST RESULTS

Weight	0.1	0.5	1.0	1.5	2.0	10
mIoU	38.38	38.50	**38.86**	38.33	38.53	38.01
mAP	58.82	59.72	57.96	58.31	57.53	**59.75**

The results are shown in Tab. VII. Testing with our own full BEV semantic map, the proposed Seq-BEV achieves 38.86% in mIoU and 57.69% in mAP, outperforming all the other methods for most categories. We find that our method is more effective in segmenting small objects, especially for the pedestrian category. It can also be seen that the performance of the original networks is better than the modified networks that take as input sequential images. The reason behind this result may be that the fusion of sequential information requires a special design to get better performance.

2) The Qualitative Demonstrations: Fig. 7 shows sample qualitative demonstrations. Due to space limitation, we only displayed the semantic BEV maps generated by the networks that achieved better quantitative results. We can see that our Seq-BEV produces a more accurate semantic full-BEV map. From the first two rows, compared with the other methods, our Seq-BEV is more sensitive to small objects, such as obstacles or pedestrians on the road. However, the predicted position of the small objects is not perfect because these categories account for a small proportion of the dataset. Moreover, the size of the vehicle predicted by Seq-BEV is closer to the real one. We credit this to the multi-scale dilated convolutions, which are used to process the high-level feature and make it invariant in scale. The last row demonstrates the semantic BEV prediction at nighttime, which indicates that Seq-BEV can still

TABLE VI

THE ABLATION STUDY RESULTS (%) ON THE DISTANCE INTERVALS. SEQ-BEV PROCESSES THE IMAGES AT SPECIFIC DISTANCE INTERVALS TO CAPTURE ENVIRONMENTAL DETAILS BEYOND THE FRAME'S VISUAL RANGE. IN THIS ANALYSIS, DISTANCE INTERVALS OF 10, 20, AND 30 METERS ARE USED TO IDENTIFY THE OPTIMAL CONFIGURATION. THE BOLD FONT HIGHLIGHTS THE BEST RESULTS IN EACH COLUMN

Distance Interval	Background		Drivable Area		Ped. Crossing		WalkWay		Obstacle		Vehicle		Pedestrain		mIoU	mAP
	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP		
10-meter	**66.59**	**76.85**	**74.89**	**85.19**	**35.76**	**65.27**	**43.73**	**64.72**	**22.27**	**52.43**	**27.63**	**54.32**	**1.12**	**7.24**	**38.86**	**57.96**
20-meter	65.15	75.40	73.48	84.34	33.00	62.10	40.23	61.01	16.03	41.91	25.20	51.64	0.52	4.68	36.23	54.44
30-meter	63.12	74.31	71.45	81.99	27.64	63.96	36.76	57.24	10.49	50.76	22.05	52.43	0.44	6.84	33.14	55.36

TABLE VII

THE COMPARATIVE RESULTS (%) COMPARED WITH THE STATE-OF-THE-ART METHODS. WE CONDUCT TWO GROUPS OF TESTS. ONE IS THE ORIGINAL NETWORK, WHICH TAKES AS INPUT A SINGLE IMAGE. THE OTHER IS THE MODIFIED NETWORK, WHICH TAKES AS INPUT SEQUENTIAL IMAGES. WE USE *SGL* AND *SEQ* TO DISTINGUISH THE TWO GROUPS. SOME BEV-BASED DETECTION METHODS ARE ALSO COMPARED BY ADDING THE SEGMENTATION HEAD. *FV* MEANS THAT THE FRONT-VIEW CAMERA IMAGES ARE TAKEN AS INPUT. *MV* MEANS THAT THE INPUT IS FROM MULTI-VIEW CAMERAS. THE BOLD FONT HIGHLIGHTS THE BEST RESULTS IN EACH COLUMN

Methods	Modality	Background		Drivable Area		Ped. Crossing		WalkWay		Obstacle		Vehicle		Pedestrain		mIoU	mAP
		IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP		
SGL-CVT (CVPR 2021)	FV	**67.27**	77.85	72.98	85.15	28.80	53.25	**46.69**	64.43	7.65	35.59	26.11	42.59	0	0	33.91	49.05
SGL-VED (R-AL 2019)	FV	65.91	76.56	74.08	83.98	33.02	57.42	44.30	**65.03**	12.61	49.81	16.82	42.41	0	0	35.25	53.60
SGL-MonoLayout (WACV 2020)	FV	53.84	64.43	62.58	78.15	2.82	16.97	26.99	43.88	0.82	1.58	8.60	23.56	0	0	22.13	32.65
SGL-VPN (R-AL 2019)	FV	63.35	71.74	71.97	83.75	22.93	55.24	38.19	64.56	14.12	41.40	21.06	44.91	0	0	33.09	51.65
SEQ-CVT (CVPR 2021)	FV	60.84	71.96	68.07	80.62	13.89	29.99	36.14	55.15	1.55	40.10	12.06	35.60	0	0	27.51	44.77
SEQ-VED (R-AL 2019)	FV	61.73	73.68	69.03	79.93	16.19	36.99	38.31	58.63	10.18	36.64	11.40	35.77	0.08	2.17	29.56	46.26
SEQ-MonoLayout (WACV 2020)	FV	53.40	63.06	62.23	78.37	2.85	19.44	25.79	43.46	0.04	4.10	8.40	24.49	0	0	21.82	33.28
SEQ-VPN (R-AL 2019)	FV	58.70	70.70	67.32	75.65	2.46	55.69	33.42	64.05	1.80	44.60	9.10	48.14	0	0	24.69	51.26
BEVFormer (ECCV 2022)	MV	50.73	66.72	63.45	74.05	17.11	43.39	26.5	45.23	5.5	21.7	14.17	37.56	0.02	0.2	25.35	41.26
BEVDepth (AAAI 2023)	MV	63.85	**77.87**	**75.53**	81.48	**36.71**	63.68	23.9	47.38	0	0	0.41	16.74	0	0	28.63	47.86
MatrixVT (ICCV 2023)	MV	54.07	61.37	63.71	80.20	24.67	57.41	27.67	48.33	6.76	33.12	6.13	33.67	0	0	26.14	44.87
SEQ-BEV (Ours)	FV	66.59	76.85	74.89	**85.19**	35.76	**65.27**	43.73	64.72	**22.27**	**52.43**	**27.63**	**54.32**	**1.12**	**7.24**	**38.86**	**57.96**

generate a clear and precise result under dark illumination conditions.

However, challenges may arise when the ego-vehicle is stationary while other vehicles remain in motion, such as waiting at an intersection. Since our sequential processing is triggered by changes in distance and the distance remains unchanged in such scenarios, our method is less robust to predict the BEV maps, especially in detecting dynamic objects. A possible solution to alleviate this limitation would be to integrate additional sensors, such as event cameras, to enhance the detection of dynamic objects.

V. CONCLUSION AND FUTURE WORK

We presented here a novel road-aware semantic BEV network, Seq-BEV, that takes as input the equidistant sequence images and directly outputs a semantic full-BEV map. In this work, we developed a self-adapted sequence fusion module and a road-aware view transformation module. We conducted extensive experiments to verify the effectiveness of the network structure and the designed modules. The proposed network was also compared with the other state-of-the-art methods and achieved superior performance. Although our Seq-BEV provides acceptable semantic BEV maps, the prediction of small objects is still not satisfactory. In the future, we plan to alleviate this issue by generating more training data with small objects using generative artificial intelligence technologies. We will also integrate other sensors, such event cameras, to enhance the detection of dynamic objects when the ego-vehicle is stationary. To make the BEV maps more practical in real applications, we plan to add more semantic classes, such as traffic signs, into the BEV maps in the future.

REFERENCES

[1] G. Du, Y. Zou, X. Zhang, Z. Li, and Q. Liu, "Hierarchical motion planning and tracking for autonomous vehicles using global heuristic based potential field and reinforcement learning based predictive control," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 8, pp. 8304–8323, Aug. 2023.

[2] Y. Yoon, C. Kim, H. Lee, D. Seo, and K. Yi, "Spatio-temporal corridor-based motion planning of lane change maneuver for autonomous driving in multi-vehicle traffic," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13163–13183, Oct. 2024.

[3] Y. Feng and Y. Sun, "PolarPoint-BEV: Bird-eye-view perception in polar points for explainable end-to-end autonomous driving," *IEEE Trans. Intell. Vehicles*, early access, Feb. 1, 2024, doi: 10.1109/TIV.2024.3361093.

[4] Y. Feng, Z. Feng, W. Hua, and Y. Sun, "Multimodal-XAD: Explainable autonomous driving based on multimodal environment descriptions," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 12, pp. 19469–19481, Dec. 2024.

[5] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-end vision-based autonomous driving via spatial–temporal feature learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2022, pp. 533–549.

[6] A. Feng, C. Han, J. Gong, Y. Yi, R. Qiu, and Y. Cheng, "Multi-scale learnable Gabor transform for pedestrian trajectory prediction from different perspectives," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13253–13263, Oct. 2024.

[7] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn.*, vol. 78, Aug. 2017, pp. 1–16.

[8] C. Lu, M. J. G. van de Molengraft, and G. Dubbelman, "Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 445–452, Apr. 2019.

[9] S. Gao, Q. Wang, and Y. Sun, "S2G2: Semi-supervised semantic bird-eye-view grid-map generation using a monocular camera for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 11974–11981, Oct. 2022.

[10] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4867–4873, Jul. 2020.

[11] S. Gao, Q. Wang, and Y. Sun, "Obstacle-sensitive semantic bird-eye-view map generation with boundary-aware loss for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2024, pp. 466–471.

[12] S. Gao, Q. Wang, D. Navarro-Alarcon, and Y. Sun, "Forecasting semantic bird-eye-view maps for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2024, pp. 509–514.

[13] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11135–11144.

[14] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "BEVSegFormer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5935–5943.

[15] Y. Jiang et al., "PolarFormer: Multi-camera 3D object detection with polar transformer," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, 2023, pp. 1042–1050.

[16] C. Yang et al., "BEVFormer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17830–17839.

[17] Z. Li et al., "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–18.

[18] H. Caesar et al., "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11618–11628.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[20] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[21] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2018, pp. 801–818.

[23] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[24] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7242–7252.

[25] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1280–1289.

[26] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Aug. 2023, pp. 4015–4026.

[27] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, Dec. 2023.

[28] H. Li, H. K. Chu, and Y. Sun, "Temporal consistency for RGB-thermal data-based semantic scene understanding," *IEEE Robot. Autom. Lett.*, vol. 9, no. 11, pp. 9757–9764, Nov. 2024.

[29] Z. Feng, Y. Guo, and Y. Sun, "Segmentation of road negative obstacles based on dual semantic-feature complementary fusion for autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 4, pp. 4687–4697, Apr. 2024.

[30] Y. Wang, H. K. Chu, and Y. Sun, "PEAFusion: Parameter-efficient adaptation for RGB-thermal fusion-based semantic segmentation," *Inf. Fusion*, vol. 120, Aug. 2025, Art. no. 103030.

[31] H. Li, H. K. Chu, and Y. Sun, "Improving RGB-thermal semantic scene understanding with synthetic data augmentation for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 10, no. 5, pp. 4452–4459, May 2025.

[32] W. Zhou, H. Wu, and Q. Jiang, "MDNet: Mamba-effective diffusion-distillation network for RGB-thermal urban dense prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 4, pp. 3222–3233, Apr. 2025.

[33] J. Liu, H. Liu, X. Li, J. Ren, and X. Xu, "MiLNet: Multiplex interactive learning network for RGB-T semantic segmentation," *IEEE Trans. Image Process.*, vol. 34, pp. 1686–1699, 2025.

[34] L. Reiher, B. Lampe, and L. Eckstein, "A Sim2Real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–7.

[35] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 9200–9206.

[36] S. Gong et al., "GitNet: Geometric prior-based transformation for birds-eye-view segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2022, pp. 396–411.

[37] T. Zhao et al., "Improving Bird's eye view semantic segmentation by task decomposition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 15512–15521.

[38] Y. Li et al., "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 2, pp. 1477–1485.

[39] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Rustenburg, South Africa, Jul. 2020, pp. 1–6.

[40] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Mach. Vis. Appl.*, vol. 32, no. 6, p. 121, Nov. 2021.

[41] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3D convolution for video super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 1500–1504, 2020.

[42] S. Mittal and Vibhu, "A survey of accelerator architectures for 3D convolution neural networks," *J. Syst. Archit.*, vol. 115, May 2021, Art. no. 102041.

[43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[44] J. Lin, S.-H. Zhong, and A. Fares, "Deep hierarchical LSTM networks with attention for video summarization," *Comput. Electr. Eng.*, vol. 97, Jan. 2022, Art. no. 107618.

[45] C. Bhuvaneshwari and A. Manjunathan, "Advanced gesture recognition system using long-term recurrent convolution network," *Mater. Today, Proc.*, vol. 21, pp. 731–733, Jun. 2020.

[46] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7082–7092.

[47] K. Shan, Y. Wang, Z. Tang, Y. Chen, and Y. Li, "MixTConv: Mixed temporal convolutional kernels for efficient action recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 1751–1756.

[48] S. Yang et al., "Temporally efficient vision transformer for video instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2875–2885.

[49] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 28, Dec. 2015, pp. 2017–2025.

[50] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[51] Y. Zhou, Y. Takeda, M. Tomizuka, and W. Zhan, "Automatic construction of lane-level HD maps for urban scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 6649–6656.
[52] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[53] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[54] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent.*, Jul. 2017.
[55] A. Howard et al., "Searching for mobilenetv3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1314–1324.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[57] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[58] W. Yang et al., "Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15531–15540.
[59] K. Mani, S. Daga, S. Garg, S. S. Narasimhan, M. Krishna, and K. M. Jatavallabhula, "Monolayout: Amodal scene layout from a single image," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Aug. 2020, pp. 1689–1697.
[60] H. Zhou, Z. Ge, Z. Li, and X. Zhang, "MatrixVT: Efficient multi-camera to bev transformation for 3D perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Jul. 2023, pp. 8548–8557.

Shuang Gao (Student Member, IEEE) received the B.S. degree from Harbin Institute of Technology, Harbin, Heilongjiang, China, in 2017. She is currently pursuing the joint Ph.D. degree with the Department of Control Science and Engineering, Harbin Institute of Technology, and the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong.

Her current research interests include autonomous driving, semantic segmentation, and deep learning.

Qiang Wang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in control science and engineering from Harbin Institute of Technology (HIT), Harbin, China, in 1998, 2000, and 2004, respectively.

Since 2008, he has been a Professor with the Department of Control Science and Engineering, HIT. His research interests include hyperspectral image denoising, signal/image processing, multi-sensor data fusion, wireless sensor networks, and intelligent detection technology.

Yuxiang Sun (Member, IEEE) received the bachelor's degree from Hefei University of Technology, Hefei, China, in 2009, the master's degree from the University of Science and Technology of China, Hefei, in 2012, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2017.

He is currently an Assistant Professor with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong. His current research interests include robotics and AI, autonomous driving, mobile robots, and autonomous navigation.

Prof. Sun serves as an Associate Editor for IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON INTELLIGENT VEHICLES, IEEE ROBOTICS AND AUTOMATION LETTERS, the IEEE International Conference on Robotics and Automation, and the IEEE/RSJ International Conference on Intelligent Robots and Systems.