UIF-BEV: An underlying information fusion framework for bird's-eye-view semantic segmentation

Yilong Ren, Lening Wang, Minda Li, Han Jiang, Chunmian Lin, Haiyang Yu, Zhiyong Cui

***Abstract*—Semantic segmentation based on Bird's-eye-view (BEV) is crucial for autonomous driving. However, current methods for voxel-uplifting-based depth estimation often result in flattened ground, and transformer-based methods lack model interpretability, resulting in information loss and false transformations during image and multi-camera fusion. To tackle this issue, we propose UIF-BEV, an end-to-end framework that fuses underlying information for BEV semantic segmentation. In UIF-BEV, we construct a fusion encoder to combine the camera's underlying information and vehicle motion features across continuous frames, enabling multi-view conversion and image fusion. Additionally, we propose directional attention and tracking attention modules to enhance recognition accuracy and perception prediction for moving vehicles with varying speeds, taking into account their unsynchronized perspectives and timing. To generate segmentation results, we design a bi-directional overlapping attention decoding block that fuses multi-features. Experimental results using the nuScenes dataset demonstrate the effectiveness of UIF-BEV. It significantly improves the stitching effect of image edges and cross-views in semantic segmentation, while also reducing deformation errors caused by image transformations. Furthermore, UIF-BEV outperforms all benchmarks. Ablation experiments confirm the efficacy of each component in the framework. UIF-BEV presents a promising solution for real-time BEV map reconstruction and holds potential for various applications in the field of computer vision and autonomous driving. Our code can be publicly available at https://github.com/LeningWang/UIF-BEV.**

***Index Terms*—autonomous driving, bird's-eye-view, semantic segmentation, underlying information fusion**

I. INTRODUCTION

This work was is supported by the National Key Research and Development Project of China (2021YFB1600503) and the Zhejiang Provincial Natural Science Foundation of China (LD24F020008) and the National Natural Science Foundation of China (52202378). *(Corresponding author:Haiyang Yu, Zhiyong Cui)*

Yilong Ren, Haiyang Yu are with State Key Lab of Intelligent Transportation System, Beihang University, Beijing 100191, China, and also with the School of Transportation Science and Engineering, Beihang University, Beijing, 100191, China, and also with Hefei Innovation Research Institute, Beihang University, Hefei 230012, China, and also with Research Institute for Frontier Science, Beihang University, Beijing 100191, China, and also with Zhongguancun Laboratory, Beijing 100094, China. E-mail: yilongren@buaa.edu.cn; hyyu@buaa.edu.cn

Lening Wang, Han Jiang, Chunmian Lin, Zhiyong Cui are with State Key Lab of Intelligent Transportation System, Beihang University, Beijing 100191, China, and also with the School of Transportation Science and Engineering, Beihang University, Beijing, 100191, China. E-mail: leningwang@buaa.edu.cn; buaajh@buaa.edu.cn; cmlin@buaa.edu.cn; zhiyongc@buaa.edu.cn

Minda Li is with Navigation College, Dalian Maritime University, Dalian 116026, China. E-mail: lmd040827@dlmu.edu.cn

AS one of the fundamental tasks for autonomous driving, semantic segmentation can assign category labels to each pixel in a camera image and provide a high-level understanding of the scene for autonomous vehicles [1], [2], [3], [4]. Benefiting from its powerful representation for road scenes, bird's-eye-view (BEV) semantic segmentation has received broad attention from both the academic and industrial communities in recent years [5], [6], [7], [8]. BEV possesses a comprehensive mapping system that includes intricate details about the geographical positioning and semantic segmentation of the environment. This mapping system can be swiftly created in real-time, using cost-effective methods, thus rendering it highly suitable for facilitating autonomous driving functionalities like planning and prediction [9], [10], [11], [12].

To date, significant efforts have been dedicated to the development of suitable BEV semantic segmentation models. Earlier work mainly focused on converting images from the perspective view (PV) to the BEV view. In [13] and [14], the authors employ the inverse perspective mapping (IPM) scheme, which uses geometry transformation matrices to rectify perspective relationships and thereby achieve image perspective transformations. However, the lack of an estimate of object depth limits the effectiveness of IPMs in guiding semantic segmentation tasks. To mitigate this limitation, voxel-uplifting-based depth estimation methods have been developed. These methods initiate by partitioning the captured PV images into numerous small voxel blocks and subsequently estimating the depth for each individual block. This process effectively transforms 2D images into a pseudo 3D point cloud representation. Subsequently, these pseudo 3D depth-enhanced images are re-projected from a BEV perspective onto the ground plane, yielding precise BEV image information. For example, the Lift, Splat, Shoot (LSS) method [15] segments images into voxels and extracts 2D features from each voxel to construct depth information within a 3D space. Similarly, the subsequent CaDNN [16] uses a similar approach to improve depth distribution prediction, effectively extracting 3D information from 2D space. However, these methods, due to their reliance on row-wise scanning or convolutional neural network model architecture, are susceptible to overfitting issues. In such scenarios, the model tends to excessively memorize fine-grained details within the depth estimation of each image layer. Consequently, when vehicles traverse uneven terrain or experience a shift in the camera's position, misalignment can occur between the images and the ground pitch angle. This

misalignment, in turn, diminishes the model's capacity for generalization and results in oscillation errors.

With the penetration of transformer, researchers have turned to end-to-end deep learning methods to address these inherent limitations. For example, BEVerse [17] leverages collaborative perception from surround-view cameras to enhance both efficiency and practical application. BEVSegFormer [18] encodes multiple images through a shared backbone network and introduces deformable attention to enhance surround-view image perception and transform images into the BEV view. BEVFormer [19] proposes a spatial cross-attention [20] mechanism to aggregate spatial information, aligns BEV positions, and utilizes a similar self-attention approach to fuse historical BEV information, thereby enhancing object perception accuracy and semantic segmentation. Though promising, these methods often overlook the impact of moving vehicles with varying speeds, leading to troublesome alignment time fusions. Additionally, there is not yet a model that adequately considers the association of features between multiple images and underlying information in a panorama. Existing methods based on simple multi-view cross-correlation fail to account for feature matching with temporal continuity, and the adoption of singular up-sampling operations has also resulted in incomplete information capture and feature loss.

To address these gaps, we introduce UIF-BEV, an underlying information fusion framework for BEV semantic segmentation. Unlike existing models, our approach incorporates camera and vehicle motion data into image encoding. The underlying information covers a wide range of image data facets, including internal parameters, external parameters, camera installation details, and camera identifiers. It also encompasses essential vehicle-related data, such as vehicle pose and motion attributes, along with the amalgamation of various temporal transformations across successive frames. Additionally, we introduce directional and tracking attention modules, as well as bi-directional overlapping attention decoding blocks. This eliminates the assumption of flat ground and enhances object detection [21] at image edges, resulting in accurate BEV segmentation results.

Our contributions can be summarized as follows:

- We propose an end-to-end underlying information fusion framework that predicts and segments road scenes and moving vehicles around the ego-vehicle in the BEV perspective using surround cameras and vehicle underlying information. To the best of our knowledge, this is the first work to fuse vehicle underlying information into a BEV perception network.
- We design a camera underlying information (UI) encoder data architecture to fuse each camera's internal and external parameters and design an ego-vehicle motion information fusion encode under continuous frames to unify the BEV perspective for moving vehicles. Guided by UI block, directional and tracking attention modules are proposed to enable cross-time and cross-view information interaction.
- We introduce a bi-directional overlapping attention decoder block that efficiently fuses multi-features in both temporal and spatial domains, resulting in semantic segmentation results in the BEV perspective.
- We conduct extensive experiments on the nuScenes dataset and perform comprehensive ablation studies to demonstrate the efficiency and effectiveness of the proposed UIF-BEV.

The remainder of this paper is organized as follows: Section II reviews related works, Section III introduces the proposed UIF-BEV architecture, Section IV presents the experimental analysis, and Section V concludes the paper.

II. Related Work

The detection of surrounding vehicles and the semantic segmentation of road scenes from the BEV perspective are research areas that overlap with other fields, including 3D target detection, image transformation, semantic segmentation, and multi-modal fusion. Therefore, this section will discuss methods for transforming multiple images from PV to BEV (PV2BEV), as well as perception and semantic segmentation methods that incorporate multi-modal fusion from the BEV perspective.

A. PV2BEV

In the field of PV2BEV, the mainstream methods can be broadly classified into two categories: geometric projection-based approaches and deep learning-based approaches. Geometric projection-based methods utilize the geometric transformation relationship between the PV and the BEV. These methods often involve traditional deep learning techniques like Multilayer Perceptron (MLP) or employ image transformation learning methods based on transformers, which have gained popularity in recent years.

Geometric transformation-based PV2BEV methods. Among the geometric transformation methods, IPM [13] is considered a pioneering technique for image transformation. In PV2BEV research, some methods [22] propose a Convolutional Neural Networks (CNN) based approach that parameterizes a homography matrix with four geometric parameters. They extract semantic features from PV images, estimate the vertical vanishing point and horizon in the image. Pseudo-LiDAR++ [23] establishes a pseudo-LiDAR framework for stereo depth estimation, enabling depth estimation for distant objects. They also propose the use of cheaper but sparse LiDAR sensors to mitigate depth estimation bias. PatchNet [24] presents an image-based CNN detection method that utilizes pseudo point clouds to extract deep features and improve the performance of 3D object detection. VPOE [25] utilizes an inertial unit to mitigate the impact of camera motion (pitch and roll) and employs a convolutional neural network for vehicle position detection and three-dimensional positioning and transformation of the image. Palazzi et al. [26] propose a semantic-aware translation method that focuses on IPM to map detections in a vehicle's dashboard camera view to a broader BEV. TrafCam3D [27] introduces a detection method based on a single traffic camera to extract three-dimensional vehicle position and attitude. They establish a homography mapping

under a dual-view network framework, aiming to reduce the distortion caused by IPM.

Deep learning-based PV2BEV methods. Deep learning has been widely utilized as a powerful tool for learning complex mapping functions to achieve image transformation between PV and BEV perspectives. In this context, several methods have leveraged deep learning techniques to address the PV2BEV task. VED [28] employs a variational coding and decoding model to encode front-view image information and decompose it into coordinate images in the top-view perspective. It proposes a smaller embedded variational sampling vector to mitigate the impact of vehicle disturbances. VPN [29] addresses cross-view semantic segmentation by utilizing flat mapping to convert PV perspective to BEV perspective and introduces a view parsing network trained on 3D images. FishingNet [30] adopts multi-modal fusion to acquire and fuse data from various sensors, enabling the conversion of information from different perspectives and modalities into semantic segmentation under a unified perspective. PON [31] proposes an end-to-end deep learning architecture that directly estimates the BEV map from a monocular image using a single network structure. It introduces a semantic Bayesian framework for multi-camera and continuous-time accumulation of information. STA-ST [32] implicitly considers the geometric shape of the image, extracts multi-layer image information using feature pyramids, and leverages factorial 3D convolution to learn from the monocular view, achieving effective prediction in the BEV network image. PETR [33] encodes 3D position embedding information into 2D multiple view features to enable position perception and detection of 3D objects using position transformation. In the follow-up work PETRv2 [34], it extends the 3D position embedding to the temporal domain for spatio-temporal fusion analysis. Graph-DETR3D [35] introduces a graph neural network structure to aggregate image information for each queried object. Subsequent works such as ORA3D [36] focus on the overlapping area of images and utilize adversarial networks to improve model performance in this area. PolarDETR [37] proposes the concept of polar coordinate parameterization to redefine the object's boundary and enhances the network using the polar coordinate form.

B. Perception and Segmentation

From the BEV perspective, precise recognition and classification of vehicles and drivable areas are essential for ensuring reliable path planning and control of autonomous vehicles [38]. Several research studies have focused on this aspect, primarily leveraging surround view cameras, LiDAR point clouds, and employing various sensor fusion methods to achieve unified perception and segmentation in the BEV perspective [39].

Semantic segmentation from the BEV perspective. The SBEVNet [40] network utilizes a pair of stereoscopic images to estimate disparities and predict depth, which are then projected onto the BEV perspective. The U-Net [41] model is employed for semantic segmentation in the BEV view. BEVSegFormer [18] introduces a GANs-based transfer learning approach for depth refinement and subsequent projection onto the BEV segmentation perspective. HDMapNet [42] employs FC operators to encode camera views from different locations, constructing a local BEV map and predicting map information within the BEV perspective. In contrast, Monolayout [43] adopts an encoder-decoder structure that leverages generative adversarial network (GAN) learning for efficient BEV view translation.

Perception from the BEV perspective. Recognizing that images contain rich color and texture information, certain studies analyze 2D image features to predict 3D target detection. CenterNet [44] proposes a 3D voxel-converted feature encoding (VFE) layer, converting points within each voxel into a unified representation, and combines the sparse point cloud with a regional proposal network (RPN) for 3D object detection. M3D-RPN [45] enhances spatial perception through a depth-aware convolutional layer, facilitating independent 3D monocular detection region proposal network for improved 3D scene understanding. The D4LCN [46] network establishes a depth map by converting the estimated depth image into a pseudo point cloud representation, bridging the gap between images and point clouds for dynamic guidance model kernel learning. FCOS3D [47] proposes a fully convolutional detector architecture, redefining the centrality function based on 2D Gaussian distribution to correspond to the 3D target formulation, considering different feature levels for different object scales. Moreover, PGD [48] constructs a geometric relationship graph between predicted objects, uses probabilistic representation to analyze depth estimation uncertainty, and employs depth propagation for depth estimation tasks.

III. METHODOLOGY

We proposed the UIF-BEV model by fusing the underlying information of images and ego-vehicle to improve the accuracy of image conversion, and build direction and tracking attention module to enhance the ability of cross perception. In Fig. 1, we show the overall architecture of the proposed UIF-BEV model which mainly composes of four blocks: (1) A share backbone encoding network with the fusion of image underlying information is used to process the color, the internal and external underlying information of the image features. (2) An underlying vehicle information encoding network under continuous frames realizes fusion encoding of vehicle motion characteristics under time series. (3) A pair of directional attention module and tracking attention module are applied to the cross-temporal and cross-view respectively. (4) A pair of bi-directional overlapping attention decoder blocks, which perform cross-multi-sampling in a supervised learning and finally output semantic segmentation results under BEV. In this chapter, we will discuss and analyze each blocks in detail.

A. Mission Definition

To provide a clearer understanding of our work, we have defined our mission. First, to obtain image information of k frames at different times t by n surround-view cameras with different perspectives $Im_n^k = (G_n, T_n, N_n, I_n, O_n)_{n=1}^k$, where k ranges from 0 to 40 and n ranges from 0 to 6, both being integers, we can extract the following components:

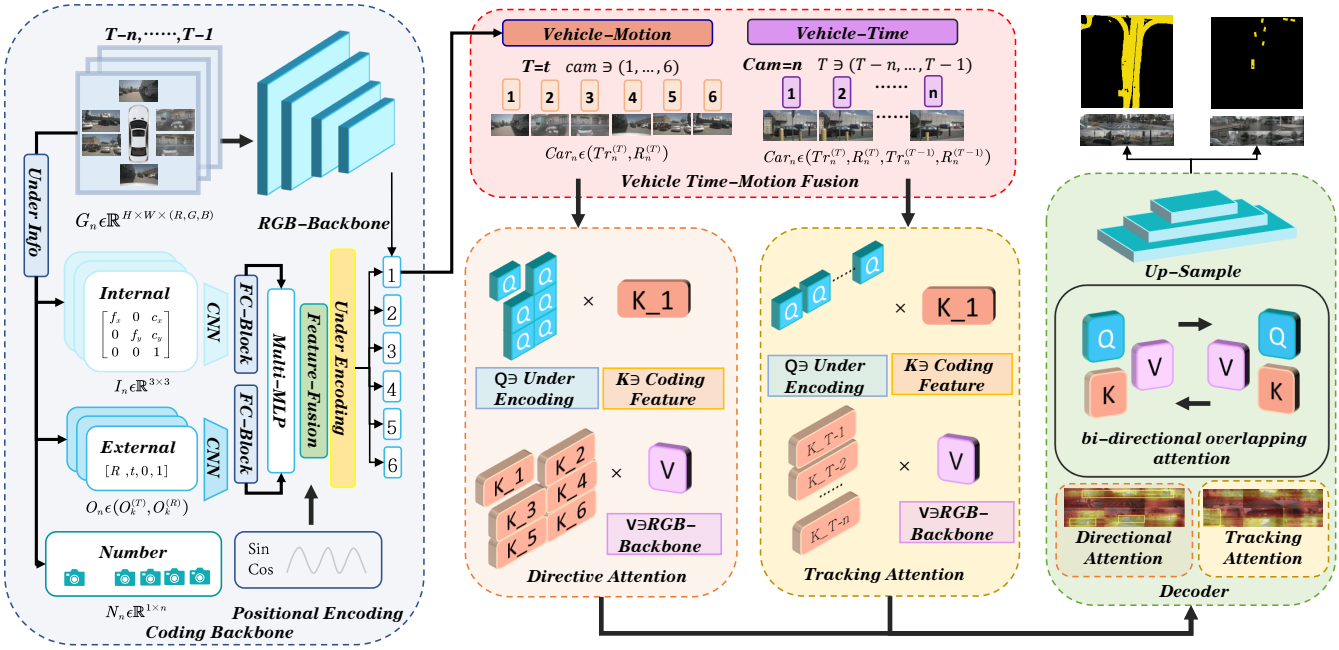


Fig. 1: The pipeline of the UIF-BEV model. The coding backbone includes image features and underlying information fusion. Vehicle time-motion fusion respectively corresponds to directional attention module and tracking attention module. A pair of bi-directional overlapping attention decoding blocks generate the final BEV feature.

- Color information: G_n represents the color information and is a real-valued tensor of size $H \times W \times (R,G,B)$.
- Time information: T_n represents the time information and is a real-valued vector of size $1 \times t$.
- Installation position number information: N_n represents the installation position number information of the camera on the vehicle and is a real-valued vector of size $1 \times n$.
- Internal and external parameter information: I_n and O_n represent the internal and external parameter information, respectively. Both are real-valued matrices, with I_n being of size 3×3 and O_n being of size 3×3 and 3×1 .

In addition, the motion position and state of the vehicle at continuous moments $C_n^k \in (T_n, R_n, Tr_n)_{n=1}^k$ are also needed, which includes the time information corresponding to the image $T_n \in \mathbb{R}^{1\times t}$, the ego-vehicle rotation parameter $R_n \in \mathbb{R}^{3\times 3}$, and the translation parameter $Tr_n \in \mathbb{R}^{1\times 3}$ with the ego-vehicle motion.

Furthermore, to enhance the comprehension of our work, we utilize these pieces of information as encoding blocks in each stage. This enables to capture the color information of the image $Im_n^k \in (G_n)_{n=1}^k$ across continuous frames $T_n \in \mathbb{R}^{1\times t}$, as well as the underlying layer parameters of the image $Im_n^k \supseteq (N_n, I_n, O_n)_{n=1}^k$ and the vehicle parameter information $C_n^k \supseteq (R_n, Tr_n)_{n=1}^k$, and other fused perception of underlying information guided by fusion coding $f \in \left\{Im_n^k, C_n^k\right\}$.

Finally, leveraging the proposed UIF-BEV framework, semantic segmentation results are generated, which include vehicle drivable area information $\hat{y}_r \in \{0,1\}^{h\times w\times c}$ and the location information of surrounding vehicles $\hat{y}_c \in \{0,1\}^{h\times w\times c}$, presented from a BEV perspective. These results offer a comprehensive understanding of the environment by identifying drivable areas and locating nearby vehicles.

B. Image UIF Encoder

Color images contain rich semantic information that can be utilized to obtain semantic segmentation results for drivable areas and vehicles from a BEV perspective. In the image processing of BEV semantic segmentation, a crucial task is to accomplish the transformation from the PV of the image to the BEV. Regardless of whether the homography method is used for image pose transformation or deep learning algorithms are directly employed to learn the feature relationship between PV and BEV, the deformation of the image itself during transformation can lead to a loss of features. In the case of a vehicle equipped with a surround-view camera, the traditional homography transformation becomes more complex when dealing with perspective transformation and image stitching from different positions. It becomes challenging to accurately deform, twist, and splice the images simultaneously. To address this issue, instead of directly learning complex function mapping relationships, it is essential to consider the inherent characteristics of camera pose transformation. Therefore, we propose a shared backbone encoding network that integrates image-underlying information fusion. Our network utilizes a shared image backbone to capture and encode the color information of the image. Additionally, we incorporate a MLP with CNN network and employ function fusion position encoding (PE) to enhance the effectiveness and interpretability of the model. This approach combines the pose conversion function with deep learning methods, thereby leveraging both

geometric transformation and advanced learning techniques for improved performance and interpretive capabilities. The structure of this module is illustrated in Fig. 2.

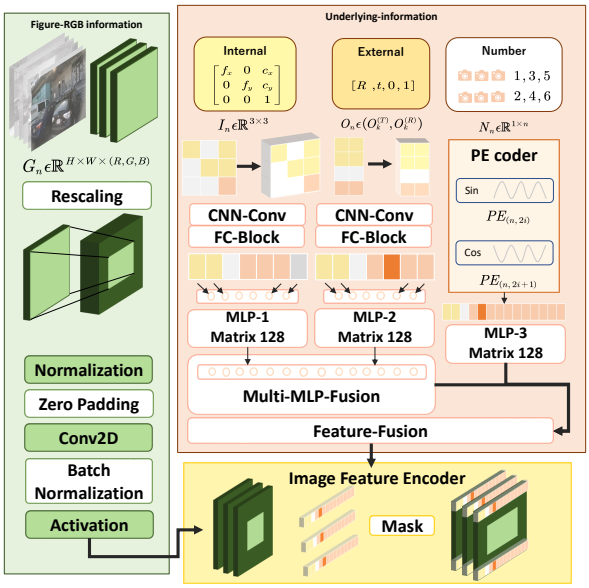


Fig. 2: Image underlying information fusion module integrates the underlying information of the vehicle, including vehicle motion, camera pose, and identification, seamlessly into the original image network.

Image color information backbone. We employ EfficientNet-B4 [49] as a shared pre-trained network for processing the color information of the images. This network is capable of extracting information and semantic features from each image. We input the information of each image, denoted as $G_n \in \mathbb{R}^{H\times W\times(R,G,B)}$, as a layer to achieve the encoding of multidimensional color feature representations.

Image underlying information encoding. To ensure effective transformation of photos captured by vehicle-mounted surround-view cameras with different positions and shooting angles into a unified perspective, we propose an image underlying information encoding network based on a kind of special MLP with CNN functions. Each camera within the surround-view camera system captures images denoted as G_n, which correspond to the underlying image parameter information captured by the n-th camera in frame k. These parameters are represented as $Im_n^k \supseteq (N_n, I_n, O_n)_{n=1}^k$. This information includes the camera's installation position number $N_n \in \mathbb{R}^{1\times n}$ within the vehicle and the underlying information specific to the camera itself (I_k), as well as external translation ($O_k^{(T)}$) and rotation ($O_k^{(R)}$) generated by the camera's installation position, angle, and other external parameters. Thus, we define the transformation function from the pose coordinates of each camera to the vehicle's pose coordinates as follows:

$$Im^{(In)} \Leftrightarrow I_k O_k^{(R)} \left(Im^{(Out)} - O_k^{(T)} \right) \tag{1}$$

Due to the variation in pose coordinates, internal parameters, and external parameters across different cameras, we aim to establish an integrated model for calculate the camera pose encoding ($Im^{(In)}$) and perform the transformation of internal and external parameters as follows:

$$Im_k^{(Out)} \Leftrightarrow Co = Im_k^{(In)} \left(I_k O_k^{(R)} \right)^{-1} + O_k^{(T)} \tag{2}$$

Considering the raw data types of camera parameters and the requirement for seamless fusion of multiple parameters, we initially developed a CNN to transform an image from one coordinate space to another type. This CNN convolves the raw matrix data and fully connects it into a 1xN-dimensional tensor, enabling effective fusion and processing of diverse information. Furthermore, to address the expectation of fusion and interaction among multiple input data, we incorporated a multi-MLP architecture immediately after the fully connected layer. This design aims to fuse and learn multi-scale underlying information features.

By utilizing the MLP with CNN model, we achieve the encoding of the pose information ($Im_k^{(In)}$) of the image itself, thereby transforming the image into $Im_k^{(Out)}$.

Furthermore, to incorporate the camera's installation position number information ($N_n \in \mathbb{R}^{1\times n}$) into the captured image, we encode the camera position and perform the transformation based on the coordinates of the final image ($Im^{(Out)}$) using PE as follows:

$$Im^{(Out)} = Im_k^{(Out)} + PE(N_n) \tag{3}$$

For the encoding process, we propose two PE methods to encode the camera's position as follows:

$$PE_{(n,2i)} = \sin\left(n / 10000^{2i/d} \right) \tag{4}$$

$$PE_{(n,2i+1)} = \cos\left(n / 10000^{2i/d} \right) \tag{5}$$

By leveraging the internal and external parameter information, we encode the image based on its underlying information, facilitating the conversion of camera pose coordinates. This encoding process allows for the transformation of both the vehicle's perspective imaging information and the encoded information into the BEV. Notably, this encoding method not only enables interpretable image pose transformation compared to traditional homography transformation methods but also introduces a novel camera underlying information encoder. By integrating the proposed MLP with CNN and PE encoders, the internal and external parameters of each camera are fused and encoded individually, further enhancing the model's effectiveness and interpretability.

To integrate the encoder into the shared backbone network for image color information, it is necessary to embed the encoding of the image's underlying information network within the established shared MLP model. Hence, we encode the direction vector of the underlying image information ($Im_n^{k\ (out)}$) into a 128-dimensional embedding $\delta_{k,i} \in \mathbb{R}^D$. Additionally, we embed the image information encoded by EfficientNet-B4 [49] into the image dimension ($\phi_{k,i}$) to establish specific correspondences between different views and timing. This enables the fusion and stitching of image color information

and the underlying encoding information, resulting in a tensor $\chi = (\phi_{k,i}, \delta_{k,i}), \chi^{H\times W\times D\times C}$.

C. Vehicle Time-Motion UIF

In previous research utilizing IPM and homography transformation, a common assumption of a flat ground surface is often made. However, this assumption can lead to significant error propagation and loss of imaging information during the image's viewing angle transformation. To tackle this challenge and improve the accuracy of edge detection and distant object perception, we introduce a vehicle time-motion feature fusion coding model based on continuous frames. Our approach integrates the encoding of fundamental information and time modality into the shared backbone encoding network, as detailed in Section 3.2, guiding the image's pose adaptation with the vehicle's motion.This integration aims to enhance the accuracy of drivable area segmentation and surrounding vehicle segmentation.

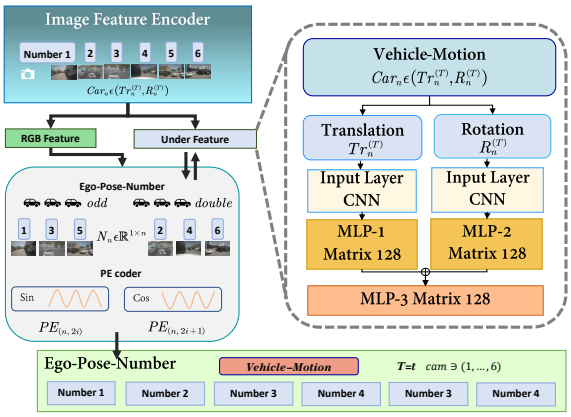


Fig. 3: Vehicle motion fusion module encodes the motion information of the vehicle into the backbone network and incorporates image parameters and position.

Vehicle motion fusion coding. The model structure module is depicted in Fig 3. When a vehicle is in motion, the movement of the vehicle itself can influence the captured imaging information from its cameras. Therefore, apart from adjusting and encoding the image parameters as discussed in Section 3.2, aligning all camera perspectives with the vehicle's motion becomes crucial. To achieve this goal, we propose motion pose encoding information for the ego-vehicle. This encoding scheme captures the vehicle's pose within the same frame for images with different orientations, leveraging the underlying information of the ego-vehicle.

In Section 3.2, we convert the perspective of $Im_n^{(In)}$ captured by n cameras within the same frame to $Im_n^{(Out)}$, which undergoes further processing for vehicle pose modal coding. This transformation results in images represented from the perspective of the vehicle's motion, denoted as $Im_n^{(Car)}$. By studying the specific influence of the vehicle's own parameters on image transformation, we employ the following transformation equation:

$$Im_n^{(Out)} \underset{T_n}{\Leftrightarrow} \left(R_n^{(T)} \left(Im_n^{(Car)} - Tr_n^{(T)} \right) \right) \tag{6}$$

Here, $Tr_n^{(T)}$ and $R_n^{(T)}$ represent the translation and rotation transformations, respectively, applied to the cameras' motion within the same frame.

Vehicle time fusion coding. The model structure module is shown in Fig.4. Considering the perception accuracy of highly occluded objects and distant objects, we propose to use continuous frames to fuse and encode a single-position camera in the vehicle pose mode. For a single camera positioned at different locations, we propose utilizing the temporal sequence information from the parameter information $C_n^k \in (T_n, R_n, Tr_n)_{n=1}^k$ of the primary vehicle in continuous frames as the basis for encoding the continuous images. Equation (7) is introduced to define this encoding process. Especially, the pose coding rules establish the connection between the vehicle's motion angle $Im_n^{(Car)}$ in the current frame and the vehicle's motion angle in the previous frame $Im^{(Car-t)}$.

In Equation (7), $Tr_n^{(T)}$ represents the vehicle translation transformation under the current frame, and $R_n^{(T)}$ represents the vehicle under the current frame rotation transformation, $Tr_n^{(T-1)}$ and $R_n^{(T-1)}$ respectively represent the translation of the vehicle in the previous frame transformation and rotation transformations. By examining the correlation between the changes in camera parameter coordinates, we construct a dedicated MLP model to learn the functional transformation described by Equation (8). This model encodes each image to align it with the ego-vehicle's pose in the continuous frames. Our approach considers the temporal relationship between the vehicle's motion in the current and previous frames and the corresponding image transformations. This enables us to analyze and encode the interplay between the image and position information from the previous and current frames effectively. As a result, the vehicle's position transformation can be simultaneously analyzed in both temporal and spatial

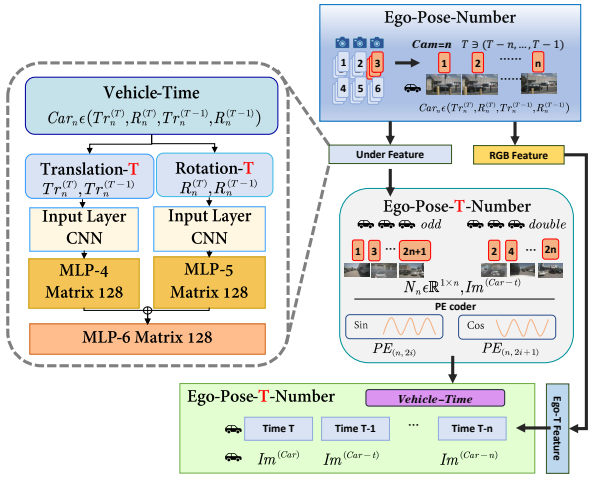


Fig. 4: Vehicle time fusion module integrates the temporal motion information of vehicles into the backbone network.

$$Im^{(Car)} \Leftrightarrow R_n^{(T-1)}\left(\left(R_n^{(T)}\left(Im^{(Car-t)} - Tr_n^{(T)}\right)\right) - Tr_n^{(T-1)}\right) \tag{7}$$

$$Im^{(Car-t)} \Leftrightarrow Ca = \left(Im^{(Car)}\left(R_n^{(T-1)}\right)^{-1} + Tr_n^{(T-1)}\right)\left(R_n^{(T)}\right)^{-1} + Tr_n^{(T)} \tag{8}$$

dimensions, providing valuable guidance for the image encoding process.

Based on the kinematics parameters of the vehicle, the spatial domain encoding is applied to multiple cameras within the same frame, while the temporal domain encoding is employed for cameras across consecutive frames. This encoding process involves considering attitude coordinates, internal parameters, and external parameters of the cameras. Initially, the cameras $Im^{(In)}$ located at different positions $n \in (0,6)$ are transformed based on their respective internal and external parameters. Subsequently, the transformed images $Im^{(Out)}$ undergo further transformation based on the final ego-vehicle coordinates. For each image with image information $Im_k^{(Out)}$, the image parameters are adjusted, resulting in the transformation of the image coordinates to $Im_k^{(Car)}$. Furthermore, to incorporate the information regarding the camera's installation position number $N_n \in \mathbb{R}^{1\times n}$ within the vehicle capturing the respective image, the position is also encoded. This encoding process is performed based on the coordinates of the image $Im_k^{(Car)}$, with the image's position encoded for each image denoted as $n \in (0,6)$, according to the following formula.

$$Im^{(Car-t)} = Im_k^{(Car-t)} + PE(N_n) \tag{9}$$

Likewise, following the aforementioned formula, we employ a mechanism for encoding the camera position utilizing function encoding. This encoding method is specifically designed to efficiently capture and represent the spatial information linked to the camera's position. Additionally, we employ a standard shared MLP as the fundamental model for encoding underlying information within images, such as vehicle motion information and camera number information. This approach enables the fusion encoding of vehicle poses effectively.

Finally, the direction vector $Im_n^{k\ (Car)}$ is encoded into a D-dimensional embedding $\psi_{k,i} \in \mathbb{R}^{D_2}$, while the direction vector $Im_n^{k\ (Car-t)}$ is encoded into a D-dimensional embedding $\zeta_{k,i} \in \mathbb{R}^{D_3}$, with D set to 128 for the encoding process. These embeddings are placed in the corresponding dimensions of the image, establishing an explicit connection between different views and timing. This integration captures the underlying information of the image, thereby incorporating the vehicle motion feature information across consecutive frames. The resulting fused and stitched representation is denoted as $\chi = \left(\phi_{k,i}, \delta_{k,i}, \psi_{k,i}, \zeta_{k,i}\right), \chi^{H\times W\times(D_1+D_2+D_3)\times C}$, where χ has dimensions of height (H), width (W), depth (D1+D2+D3), and the number of channels (C).

D. Direction and Tracking Attention

In the context of BEV perspective scene classification, the integration of multiple images and underlying information is crucial. We propose the utilization of direction and tracking attention mechanisms. These attention mechanisms are designed to handle UIF images captured at different times and positions, encoded as $\chi = \left(\phi_{k,i}, \delta_{k,i}, \psi_{k,i}, \zeta_{k,i}\right), \chi^{H\times W\times(D_1+D_2+D_3)\times C}$, as described in Sections 3.2 and 3.3 of this paper. The directional attention mechanism incorporates cross-view attention by ordering queries from different camera views at a single time under the UIF images. On the other hand, the tracking attention mechanism introduces a continuous frame to extract essential features from any single camera. We will provide a detailed explanation of the directional attention and tracking attention methods.

Directional attention module. Existing work mainly focused on building various attention methods based on a single camera installed on the head of a vehicle, such as MRSBEV [50]. Actually, self-driving vehicles equipped with surround-view cameras are the mainstream of autonomous driving in the future. Hence, it is necessary to set up attention blocks under cross-view, such as CoBEVT [51], BEVFormer [19], CVT [52] et.al. These methods all perform cross-attention based on multiple images. However, these existing methods mainly focus on information interaction among neighboring images or single image queries using multi-image position codes, thereby overlooking the possibility that the same vehicle and drivable area may appear in non-adjacent images simultaneously. Furthermore, these methods struggle to effectively handle the correlation features between non-overlapping areas in surround-view images, as their focus mainly resides in overlapping regions. Consequently, whether utilizing weight-sharing mechanisms in convolution or attention mechanisms for overlapping images, these studies do not sufficiently harness the unified integrated information from the surround-view cameras to extract the desired feature information effectively. To fully leverage the information from surround-view cameras, we propose a directional attention module. By incorporating directional attention, we can focus on non-overlapping areas across the surround-view cameras. The structure of the proposed module is illustrated in Fig. 5. This module enables us to effectively capture relevant information from these non-overlapping regions, enhancing the feature extraction process.

Our proposed directed attention module incorporates multiple cross-combined attention methods. Specifically, we utilize the image information extracted by the main module of image color information processing, as discussed in Section 3.2, and treat it as the value $V \supset \phi = [\phi_{1,i}, \phi_{2,i}, ..., \phi_{k,i}]$. Here, $\phi_{1,i}$, $\phi_{2,i}$,

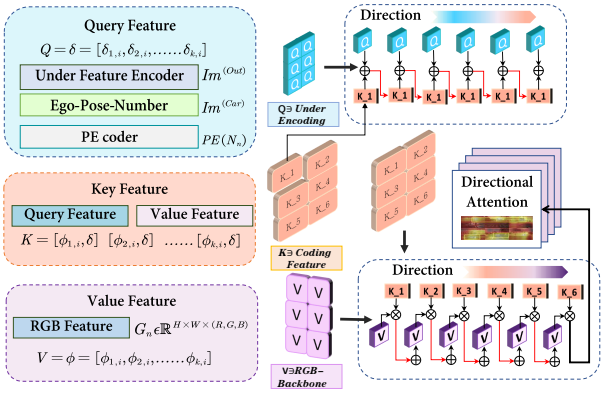


Fig. 5: Directional attention module aims to establish continuous attention among different images captured at the same moment, enabling efficient extraction of correlated image features.

..., $\phi_{k,i}$ represent the multi-image information extracted by the image backbone from the same frame. For the underlying information features $\delta_{1,i}, \delta_{2,i}, ...\delta_{k,i}$ encoded by the image underlying information encoding network, it is used as query $Q \supset \delta = [\delta_{1,i}, \delta_{2,i}, ...\delta_{k,i}]$. Regarding the key parameter in the attention mechanism, we propose to overlay and fuse the corresponding image feature $\phi_{k,i}$ tensor and utilize them as Key $K \supset [\phi_{1,i}, \delta], [\phi_{2,i}, \delta], ... [\phi_{k,i}, \delta]$. Each query and value will refer to multiple keys, and each key will engage with multiple queries and values through interactions. By employing this cross-view directed attention query mechanism, we can purposefully transfer and exchange information between different views, thereby obtaining more comprehensive and enriched feature representations. The specific formula can be expressed as:

$$Q = w_q \cdot [\delta_{1,i}, \delta_{2,i}, ... \delta_{k,i}] \tag{10}$$

$$K = \left\{ w_{k1} \cdot \left[\phi_{1,i}, \delta\right], w_{k2} \left[\phi_{2,i}, \delta\right], ... w_{ki} \left[\phi_{k,i}, \delta\right] \right\} \tag{11}$$

$$V = w_v \cdot \left[\phi_{1,i}, \phi_{2,i}, ... \phi_{k,i}\right] \tag{12}$$

In particular, we can systematically divide the computational form of directional attention into three key stages. Firstly, we perform a continuous matrix operation on each extracted Q and K.

$$At_D(K_1) = \left(\left(Q_1 \cdot K_1^T \right) \cdot Q_2 \right) \cdot Q_3 \ldots \cdot Q_6 \tag{13}$$

Subsequently, we perform operations on the resulting matrix $At_D(K_1)$ in relation to the corresponding V values.

$$At_D(V_1) = \left(\left(At_D(K_1) \cdot V_1 \right) \cdot V_2 \right) \cdot V_3 \ldots \cdot V_6 \tag{14}$$

Finally, we input the matrix $At_D(V_1)$ obtained from the continuous operations into the directional attention mechanism.

$$At_D(Q,K,V) = softmax\left(\frac{At_D(K_1)}{\sqrt{d_k}} \right) At_D(V_1) \tag{15}$$

The parameters required for the model to learn include $w_q, w_{k1}, \ldots, w_{ki}, w_v$. Through this proposed directional attention module, we are able to simultaneously learn perceptual features of specific targets.

Tracking attention module. The perception and prediction of highly occluded vehicles pose significant challenges to ensuring the safety of autonomous driving vehicles. Scholars have attempted to address this issue by utilizing the perception of the previous frame from the BEV perspective to align the position of the current frame, incorporating the attention mechanism to enhance the possibility of identifying highly occluded objects. However, existing approaches such as BEVformer [19] face limitations when there is a substantial speed difference between vehicles, particularly when the ego-vehicle's speed is slow and the surrounding vehicles' speed is fast, or when the surrounding vehicles' speed is slower than that of the ego-vehicle. Solely aligning the position of the ego-vehicle under these conditions does not provide an easy means of obtaining information about the corresponding vehicle from a cross-view perspective. Therefore, we propose a tracking attention module designed to perform attentive tracking across consecutive frames captured by multiple cameras. This module leverages the power of positional and temporal co-encoding to enhance tracking accuracy and robustness. This approach involves tracking the multi-camera view of the previous frame guided by vehicle motion encoding and tracking and focusing on the specific view of the previous frame guided by the current frame encoding. As a result, we can efficiently extract vehicle recognition information from consecutive frames captured from non-intersecting perspectives, allowing us to effectively leverage the directional attention module in the surround-view camera setup. The proposed model's structure is illustrated in the Fig. 6.

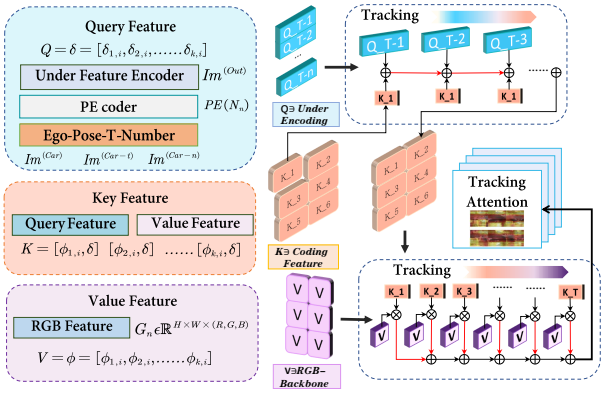


Fig. 6: Tracking attention module aims to apply attention to the imaging information captured at the same location across consecutive time frames, in order to extract correlated features from the sequential images.

The image color information in the context of consecutive frames is represented as the value $V_t \supset \phi = \left[\phi_{1,i}, \phi_{2,i}, \ldots \phi_{k,i}\right]$, where $\phi_{1,i}$, $\phi_{2,i}$, ..., $\phi_{k,i}$ correspond to consecutive frames captured by the same camera at different time instances. As for the continuous frame features encoded by the image underlying information, denoted as $\delta_{1,i}, \delta_{2,i}, ...\delta_{k,i}$, we use them as

$$At_T(V_1) = \left(At_T(K_1) \cdot V_1\right) \oplus \left(At_T(K_2) \cdot V_1\right) \oplus \ldots \oplus \left(At_T(K_6) \cdot V_1\right) \tag{17}$$

temporal queries, represented as $Q_t \supset \delta = [\delta_{1,i}, \delta_{2,i}, \ldots \delta_{k,i}]$. For the parameter keys in the attention mechanism, we propose to aggregate and fuse the temporal image feature tensors $\phi_{k,i}$ from the same pose, and use them as the keys, denoted as $K_t \supset [\phi_{1,i}, \delta], [\phi_{2,i}, \delta], \ldots [\phi_{k,i}, \delta]$. Queries and values from different time instances all point to different keys. Similarly, each key interacts with queries and values from consecutive time steps. Through the attention mechanism operating on continuous frames in a temporal sequence, the transmission and interaction of information is achieved, resulting in more enriched feature representations.

Similarly, we can systematically divide the computation form of the tracking attention into three key stages. First, we perform matrix operations on each extracted Q and K separately.

$$At_T(K_1) = \left(Q_1 \cdot K_1^T\right) \oplus \left(Q_2 \cdot K_1^T\right) \oplus \ldots \oplus \left(Q_6 \cdot K_1^T\right) \tag{16}$$

Subsequently, we perform operations on the resulting matrix $At_T(K_1)$ in relation to the corresponding V values as Equation (17).

Lastly, the matrix $At_T(V_1)$ obtained through consecutive and summation operations is input into the tracking attention mechanism.

$$At_T(Q,K,V) = softmax\left(\frac{At_T(K)}{\sqrt{d_k}}\right) At_T(V) \tag{18}$$

Based on these stages, we introduce a tracking attention module under temporal conditions, enabling the model to learn perceptual features of targets, such as highly occluded objects, across consecutive moments. Through the guidance of various underlying information encoding. The parameters $w_q, w_{k1}, \ldots, w_{ki}, w_v$ encapsulate the necessary learnable information for the model.

E. A pair of bi-directional overlapping attention decoding blocks

In the decoding phase of our model, we have devised a pair of decoding modules that incorporate bi-directional overlapping attention. This approach enables a secondary interaction between temporal and spatial data, allowing us to reconstruct the feature map to match the label size of the original BEV perspective. Our proposed decoding module comprises two main components: a bi-directional attention mechanism and an overlapping decoding block. The bi-directional attention mechanism allows the extracted network to benefit from multi-level features obtained through directional and tracking attention, guided by the encoding process described in Section 3.4. This mechanism ensures the application of bi-directional attention during the decoding process. Furthermore, the overlapping decoding block further enhances the effectiveness of the bi-directional attention mechanism. The final step involves upsampling the feature map in a supervised learning manner and generating the output of semantic segmentation. The proposed model structure, including the bi-directional attention mechanism and overlapping decoding block, is illustrated in the Fig. 7

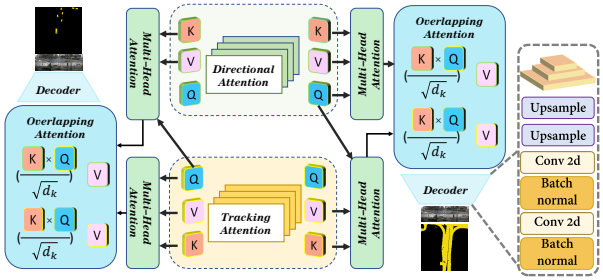


Fig. 7: The decoding blocks enable cross-fusion and sampling of information extracted from directional attention and tracking attention, allowing for the update of BEV features.

In the module design of the bi-directional attention mechanism, we aimed to effectively integrate the directional attention and the tracking attention mechanism described in Section 3.4. To achieve this, we introduced a bi-directional attention mechanism that facilitates the interaction between three models: A_C, A_D, and A_T, each corresponding to a different type of attention.

$$A_C = softmax\left(\frac{Q_t K^T}{\sqrt{d_k}}\right) \cdot V \tag{19}$$

As shown in the equation above, the A_C attention module incorporates a bi-directional interaction by utilizing the output value Q_t from the tracking attention and the output values K and V from the directional attention. This bi-directional interaction within the A_C attention module enables an enhanced integration of information, fostering a more comprehensive understanding of the dynamic environment[53].

$$A_D = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \tag{20}$$

$$A_T = softmax\left(\frac{Q_t K_t{}^T}{\sqrt{d_k}}\right) \cdot V_t \tag{21}$$

As illustrated in the preceding equations, a further tracking mechanism is employed for the images within the tracking attention and pointing attention. In the attention modules A_D and A_T, multiple values from the tracking attention (Q_t/Q) and multiple values $(K_t/K, V_t/V)$ from the directional attention are overlapped and interacted with each other. This enables multiple fusion interactions of information and facilitates the generation of multi-level output results.

In summery, to enhance the decoding and extraction of sampled data, we propose an overlapping sampling decoding mechanism. In the conventional decoding process, the model generates a single output result based on the conditional probability distribution, often resulting in limited diversity.

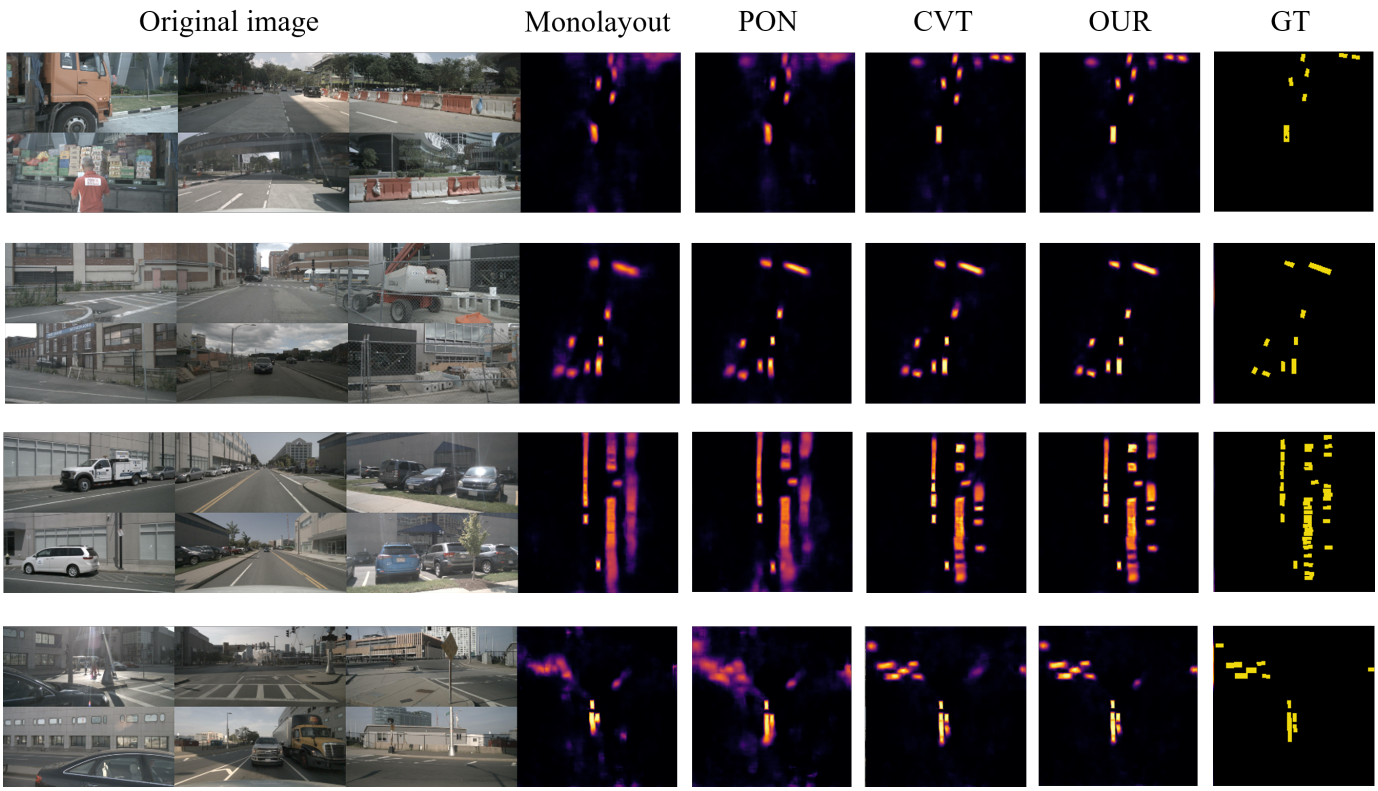


Fig. 8: Comparison of vehicle perception results. On the left side, raw image features extracted by the camera are displayed, while the right side showcases the generated results of Monolayout, PON, CVT, our proposed model, and the ground truth.

However, the overlapping sampling decoding mechanism increases the diversity of model generation while maintaining strong robustness.

Specifically, the multi-sampling decoding mechanism involves two steps. In the first step, we perform bi-directional overlapping attention between the directional attention mechanism and the tracking attention mechanism. Subsequently, the 3D spatial features are fed into the multi-level BEV Upsampling module, which includes a dropout layer, a 1×1 convolutional layer, and several stages of the BEV upsampling module. Each stage of the BEV upsampling module consists of a 3×3 convolutional block, a 1×1 convolutional block, and bilinear interpolation operations.

To enhance the fusion and attention perception of various information sources, such as image information, vehicle underlying information, spatial overlapping information, and time domain overlapping information, we propose a bi-directional overlapping attention mechanism. This mechanism performs self-encoding queries using multiple bi-directional attention mechanisms, resulting in multiple query results. Finally, these results are input into the multi-head attention mechanism for further cross-fusion. After multiple samplings, the overlapping model selects the final output result from multiple candidate results, achieving effective decoding.

IV. Experimental

A. Dataset

In this study, we evaluated our proposed model using the nuScenes dataset, which is a large-scale multimodal perception dataset for autonomous driving developed by nuTonomy [54]. This dataset is unique in its comprehensive coverage of underlying information, including surround-view camera data, vehicle pose information, and motion information. It consists of a significant amount of high-definition sensor data such as lidar, 6 surround view cameras, radar, GPS, IMU and provides high-precision vehicle position and attitude information. The nuScenes dataset serves as a valuable resource for researchers and engineers to develop smarter and safer autonomous driving systems. For research convenience, the nuScenes dataset includes comprehensive annotation information, encompassing details such as the position, size, speed, and motion trajectory of vehicles, pedestrians, and other obstacles. It also provides semantic segmentation and road topology information, enhancing its utility for various research purposes. Moreover, the dataset offers multiple challenging tasks, including lane line detection, object detection, and semantic segmentation.

B. Implementation Details

For the experimental implementation, we utilized an NVIDIA 3090 GPU and employed a PyTorch-based software framework. During model training, we utilized the Adam optimization algorithm with a learning rate of 1e-2 and a weight decay of 1e-7. We conducted semantic segmentation

Table I: Comparison of Vehicle Perception Results

Methods	Public Year	Modality	IoU(%)	mIoU(%)	mAP(%)
OFT	BMVC-2019	CAM (RGB)	30.1	30.7	34.2
LSS	ECCV-2020	CAM (RGB)	32.1	31.9	33.5
IPM	IVC-1991	CAM (RGB)	10.7	13.6	13.1
VPN	RA-L-2020	CAM (RGB)	24.7	25.7	27.9
MonoLayout	WACV-2020	CAM (RGB)	30.8	30.9	32.3
PON	CVPR-2020	CAM (RGB)	31.3	30.4	32.1
HDMapNet	ICRA-2022	CAM (RGB)	——	32.9	——
BEVerse	arXiv-2022	CAM (RGB+POS)	32.6	31.3	33.9
BEVSegFormer	WACV-2023	CAM (RGB)	——	——	——
BEVFormer	ECCV-2022	CAM (RGB+POS)	43.1	——	38.0
UniAD	CVPR-2023-Best Paper	CAM (RGB+POS)	——	——	——
FedBEVT	T-IV-2023	CAM (RGB+POS)	35.44	——	——
CVT	CVPR-2022	CAM (RGB+POS)	35.9	36.0	38.2
Ours	OURS	CAM (RGB)	35.04	34.97	37.3
Ours	OURS	CAM (RGB+POS)	35.87	35.93	38.1
Ours	OURS	CAM (RGB+POS)+Ego	**36.26**	**36.89**	**39.1**

separately for vehicles and lane lines. The model was trained for 120 epochs with a batch size of 4, and convergence was achieved after 39 hours of training. To ensure consistency, we uniformly resized each original image to dimensions of 224x448. Evaluation of the model was performed using both mean Intersection over Union (mIoU) and mean Average Precision (mAP) metrics. The experimental settings for the Backbone, Encoder, and Decoder are described below. We employed the pre-trained EfficientNet-B4[49] as our backbone model to extract multiple image features of different scales. The initial image input dimension was set to a tensor size of WxHxD, where D=128. Considering the quadratic increase in attention network complexity with grid size, we set W=H=25 to balance computational efficiency and information effectiveness. The Encoder in our model consists of multiple attention blocks that process the feature maps in each batch. These attention blocks help capture meaningful spatial relationships and dependencies within the features. The output BEV map has a resolution of 200x200, which effectively corresponds to the area centered around the vehicle.

C. Evaluation Metrics

The method we propose aims to accomplish two distinct tasks: vehicle drivable area segmentation and surrounding vehicle segmentation. Therefore, we will evaluate the performance of the proposed model using three commonly applied evaluation metrics. These metrics primarily include IoU, mIoU, and mAP.

IoU is a widely adopted metric that measures the overlap between the predicted segmentation mask and the ground truth mask. It is calculated as the ratio of the intersection to the union of the segmented region and the ground truth region. Higher IoU values indicate better segmentation accuracy.

$$IoU = \frac{TP}{TP+FN+FP} \tag{22}$$

mIoU is the average IoU calculated across all segmented regions. It provides a comprehensive measure of the overall segmentation performance. mIoU ranges from 0 to 1, with higher values indicating better segmentation quality.

$$MIoU = \frac{1}{n+1}\sum_{i=0}^{n}\frac{TP}{TP+FN+FP} \tag{23}$$

mAP is a widely used metric in object detection and segmentation tasks. It considers precision and recall values at various intersection over union thresholds and computes the average precision across all thresholds. mAP provides a robust evaluation of the segmentation performance, considering multiple IoU thresholds.

$$mAP = \frac{\sum_{i=0}^{n} AP_i}{n} \tag{24}$$

where, AP_1, AP_2, ..., AP_n are the Average Precision values for each individual class, and n is the total number of classes.

D. Experiment Results

This section presents a comprehensive quantitative evaluation of the proposed model. To effectively demonstrate the effectiveness and accuracy of our model, we conducted a comparative analysis with the most advanced related models, namely the OFT model [55], Lift-Splat model (LSS) [15], inverse perspective mapping model (IPM) [13], View Parsing Network model (VPN) [29], MonoLayout model [56], Pyramid Occupancy Networks model (PON) [31], BEVerse [17], HDMapNet [42], BEVSegFormer [18], BEVFormer [19], UniAD [57], FedBEVT[58], and Cross-view Transformers

model (CVT) [52] . The ground truth (GT) figures, which have been learned and contrasted by the network, are also presented on the far right of the images for visual comparison, thereby demonstrating the superior accuracy of our proposed model. This comparison was performed under nearly identical experimental conditions to ensure fair evaluations and enable the attainment of state-of-the-art results.

For the task of vehicle detection in the surrounding environment, our model relies on a substantial amount of multi-camera and vehicle underlying information data. To address this, we selected multiple models implemented on the nuScenes dataset for comparative analysis. The table I provides a summary of the evaluation metrics. Remarkably, our proposed model achieved the highest scores for all of these metrics. In particular, when compared to the latest and most superior methods, our model exhibited an efficiency improvement of approximately 0.89. We assessed the model's performance concerning the increment in distance from ego-vehicle, as depicted in Fig. 11. The measurement results are depicted in IoU accuracy, illustrating the changes in the model's performance and the trend of accuracy degradation with varying detection distance precision. Each entry showcases the average intersection accuracy for annotations at a minimum distance.

To substantiate the effectiveness of our proposed underlying information fusion module, we conducted experiments by progressively incorporating different layers of underlying image information. First, we introduced the image-based underlying information denoted as "pose or position (POS)," which aims to convey the inclusion of both internal and external image parameters, along with spatial location information. With the addition of POS camera pose encoding, we observed an improvement in IoU to 35.87%. Furthermore, we extended our experimentation by introducing the "Ego" information, which refers to the vehicle housing the sensors responsible for perceiving the environment around it. After the incorporation of Ego information, the IoU increased to 36.26%. These results effectively demonstrate the efficacy of our integration of underlying information, indicating that the inclusion of such information can significantly enhance perception outcomes. In summary, our experiments confirm that the incorporation of POS and Ego information leads to substantial improvements in IoU, providing compelling evidence for the enhanced performance achieved through our proposed underlying information fusion.

To provide significant visual evidence of the efficacy of our model, we included multiple images from the training dataset to showcase the impact of our proposed approach in Fig. 8. The visualized results provide robust support for our claims, illustrating that the method proposed in this article delivers superior prediction results for vehicle positions, particularly for vehicles located at relatively long distances within the images. Additionally, our model accurately detects the positions of surrounding parked vehicles with high speed differences, especially when the ego-vehicle is in motion. This achievement significantly improves perception outcomes.

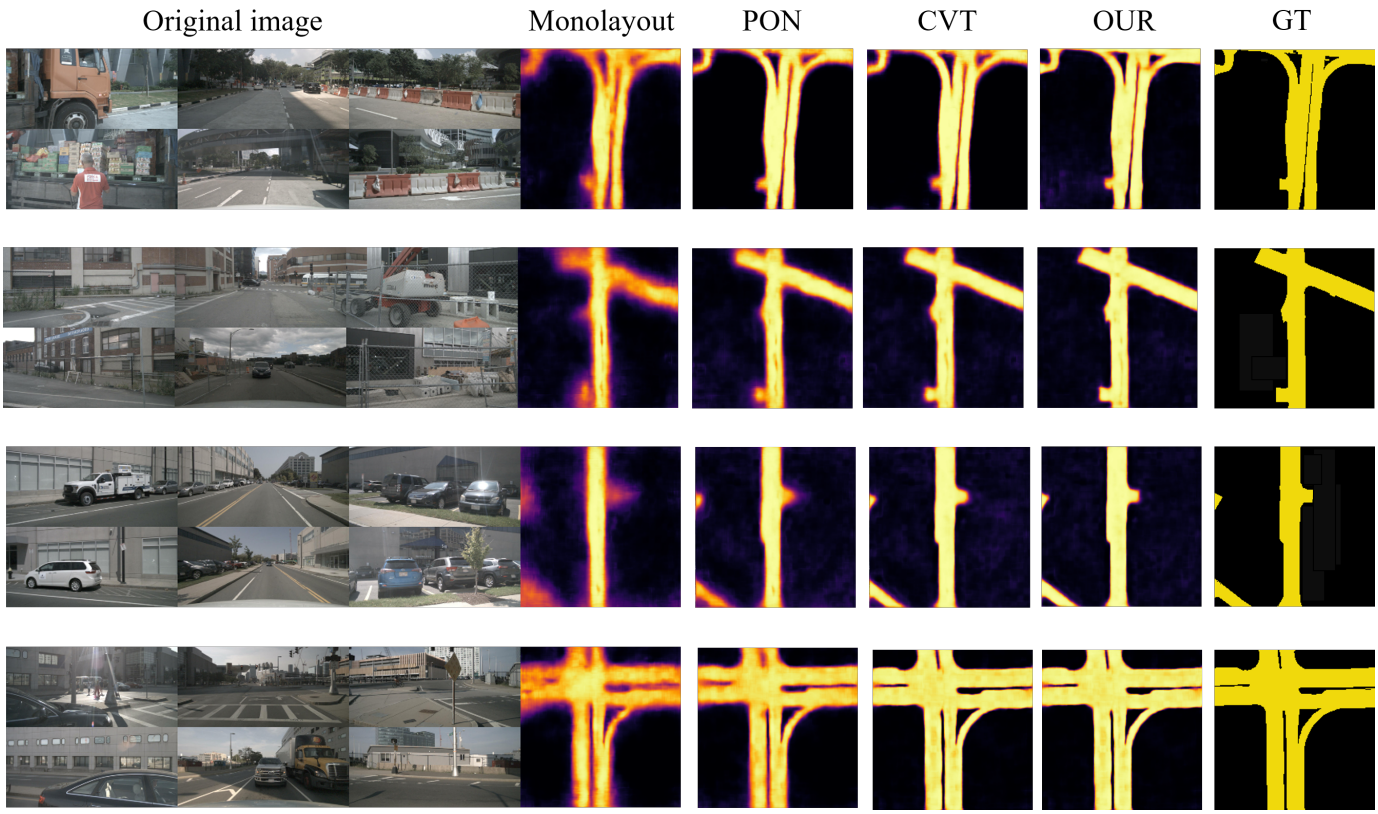


Fig. 9: Comparison of road perception results. On the left side, raw image features extracted by the camera are displayed, while the right side showcases the generated results of Monolayout, PON, CVT, our proposed model, and the ground truth.

The enhanced accuracy in perceiving the borders of these challenging scenarios can be attributed to the model we have proposed.

Besides, we conducted extensive experiments to accurately detect the drivable areas of vehicles. A comprehensive evaluation of our model's performance is presented in Table II, where our proposed approach outperforms all other models across various evaluation metrics. Compared to the traditional IPM model, our method leverages advanced deep learning techniques and incorporates attention mechanisms, resulting in an impressive performance boost of approximately 80%. Unlike the widely adopted LSS and OFT models, our approach introduces novel attention mechanisms that effectively capture the intricate edges and boundaries of drivable areas. Consequently, our method demonstrates a remarkable efficiency improvement of approximately 20% compared to these existing techniques. Furthermore, when compared to the latest state-of-the-art models, our model achieves a substantial 1.21 improvement in IoU score. This remarkable advancement is primarily attributed to our novel UIF framework, which facilitates effective perception, precise segmentation, and accurate edge recognition. By integrating these capabilities, our model attains superior accuracy levels and significantly outperforms contemporary approaches in drivable area detection. We evaluated the detection efficiency of the vehicle's drivable area in relation to the increment in distance from ego-vehicle, as shown in Fig. 12.

Table II: Comparison of Road Perception Results

Methods	Public Year	Modality	IoU(%)	mIoU(%)	mAP(%)
OFT	BMVC-2019	CAM (RGB)	71.3	73.8	74.5
LSS	ECCV-2020	CAM (RGB)	72.9	75.7	77.2
IPM	IVC-1991	CAM (RGB)	42.1	44.3	50.8
VPN	RA-L-2020	CAM (RGB)	57.4	59.7	60.4
MonoLayout	WACV-2020	CAM (RGB)	59.8	62.4	63.9
PON	CVPR-2020	CAM (RGB)	66.9	69.2	69.7
HDMapNet	ICRA-2022	CAM (RGB)	——	40.6	——
BEVerse	arXiv-2022	CAM (RGB+POS)	51.2	54.7	56.4
BEVSegFormer	WACV-2023	CAM (RGB)	——	51.7	——
BEVFormer	ECCV-2022	CAM (RGB+POS)	——	——	78.0
UniAD	CVPR-2023-Best Paper	CAM (RGB+POS)	69.1	——	——
FedBEVT	T-IV-2023	CAM (RGB+POS)	——	——	——
CVT	CVPR-2022	CAM (RGB+POS)	74.3	77.1	78.6
Ours	OURS	CAM (RGB)	73.72	75.96	76.44
Ours	OURS	CAM (RGB+POS)	74.86	76.98	78.42
Ours	OURS	CAM (RGB+POS)+Ego	**75.51**	**79.11**	**80.8**

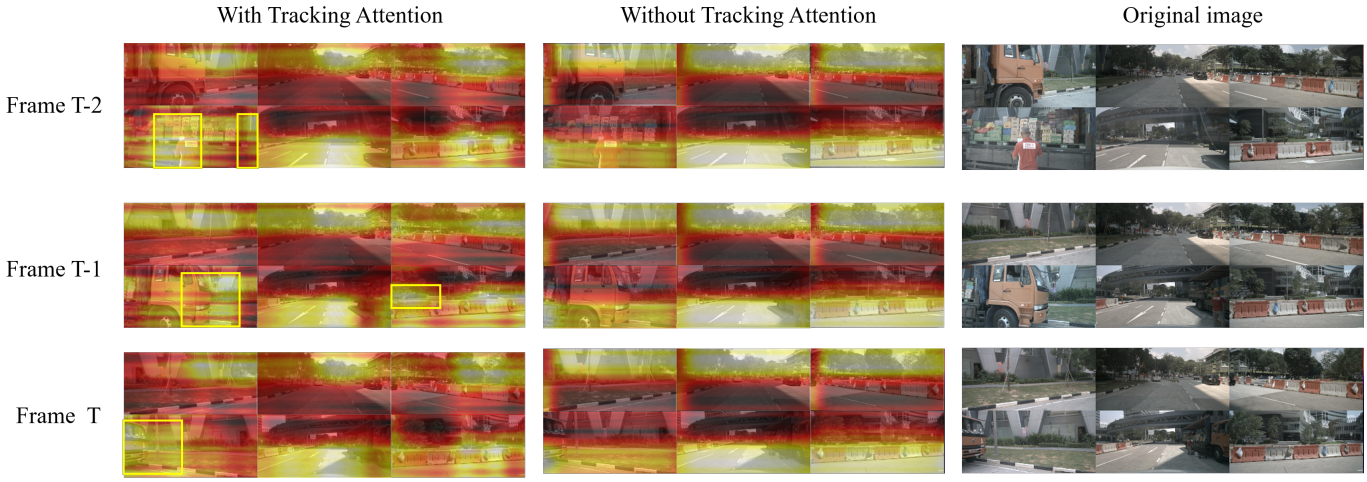


Fig. 10: The effectiveness of the tracking attention mechanism. For consecutive frames of images, the proposed tracking attention mechanism provides continuous tracking focus on objects. The visual representation with yellow and white regions showcased the attention effects.

Table III: Comparison of Ablation Experiment Results

Methods	component-time	image-attention	camera-aware	Type	IoU(%)	mIoU(%)	mAP(%)	param(M)
Including all components δ	✓	✓	✓	Vehicle	36.26	36.89	39.1	12.3
No camera-aware embedding δ	✓	✓	×	Vehicle	31.0	33.24	37.81	11.6
No image features ϕ in attention	✓	×	✓	Vehicle	33.2	34.42	36.21	10.3
No T camera-aware embedding δ	×	✓	×	Vehicle	30.9	32.67	35.72	9.7
No T image features ϕ in attention	×	×	✓	Vehicle	31.27	34.02	35.43	9.1
Including all components δ	✓	✓	✓	Road	75.51	79.11	80.8	13.2
No camera-aware embedding δ	✓	✓	×	Road	73.96	74.88	75.92	11.9
No image features ϕ in attention	✓	×	✓	Road	74.54	77.71	78.37	10.1
No T camera-aware embedding δ	×	✓	×	Road	71.39	71.78	71.02	9.2
No T image features ϕ in attention	×	×	✓	Road	69.92	72.11	71.97	8.9

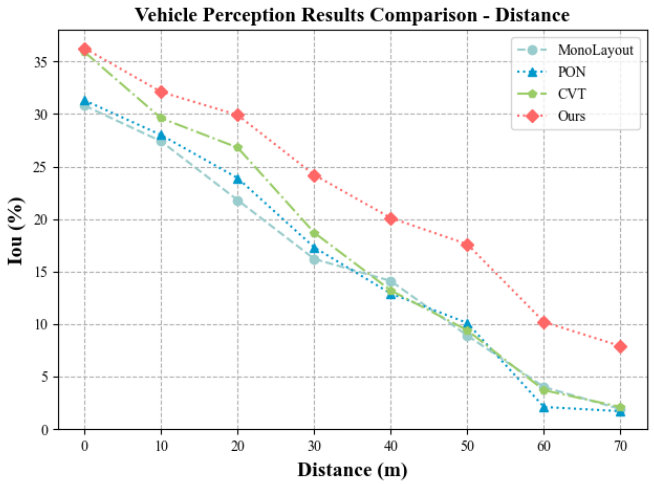


Fig. 11: The chart illustrates an experimental comparison of vehicle perception accuracy concerning the increase in distance from ego-vehicle.

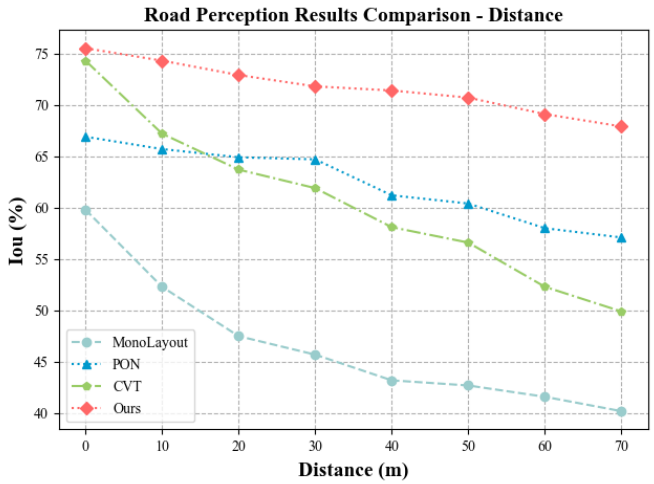


Fig. 12: The chart illustrates an experimental comparison of road perception accuracy concerning the increase in distance from ego-vehicle.

The measurement results are represented in Intersection-over-Union accuracy. Compared to vehicle detection, the efficiency of drivable area detection remains more stable.

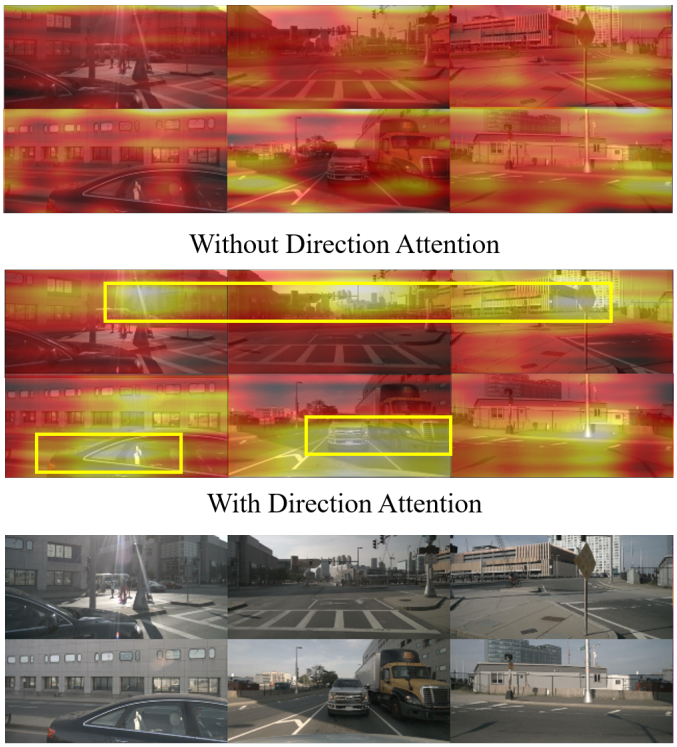
In the context of vehicle drivable area recognition, we have also validated the effectiveness of the proposed low-level information fusion module. By introducing the encoding of low-level information from POS images, we observed an improvement in IoU to 74.86%. Furthermore, with the incorporation of Ego information, the IoU increased to 75.51%. These results effectively demonstrate the efficacy of our integration of fundamental information, not only for vehicle perception but also for the perception of vehicle drivable areas. Additionally, they provide compelling evidence of enhanced model generalization performance.

To visually demonstrate the efficacy of our model, we have included multiple images from the test dataset in Fig. 9. The visualization results clearly illustrate the superior predictive capabilities of our proposed approach, particularly in accurately perceiving vehicle boundary lines. This accurate perception greatly enhances the segmentation of the drivable area. In contrast, MonoLayout, PON and CVT methods showcased on the left side of the figure adequately describes and recognizes the main body of the drivable area. However, it still exhibits significant blurriness at the edges of the drivable area and at the intersections of the two drivable regions. Such blurriness poses a considerable risk to the visual stability of autonomous vehicles. Conversely, our proposed UIF framework and the incorporation of multiple attention mechanisms based on continuous frames effectively capture these edge features. This enables a precise delineation of the drivable area, ensuring safe and reliable vehicle motion. So, the improved segmentation performance can be attributed to the effectiveness of our proposed model.

To evaluate the real-time performance of the model, we conducted experiments to measure frames per second (FPS), which is valuable for model deployment. Considering the impact of different hardware devices on the deployed model, we compared our model with methods such as CVT and PON

Table IV: Comparison of the real-time performance

Methods	Public Year	Modality	FPS	Capability
LSS	ECCV-2020	CAM (RGB)	27	8.4
PON	CVPR-2020	CAM (RGB)	21	8.4
CVT	CVPR-2022	CAM (RGB+POS)	32	8.4
Ours	OURS	CAM (RGB+POS)+Ego	**34**	8.4



Without Direction Attention

With Direction Attention

Original image

Fig. 13: The effectiveness of the directional attention mechanism. Visual representations illustrate the yellow and white regions, indicating the focal points of attention. For individual frames, the proposed directional attention mechanism specifically focuses on the features in the intersecting areas of the image.

in the same proposed hardware environment. It is essential to note that the efficiency of hardware devices significantly affects real-time performance. By employing higher-speed hardware, the deployment efficiency of the algorithm can be further improved. As demonstrated in Table IV, our model achieves superior recognition accuracy while demonstrating the optimal operational speed, reaching 34 FPS. This enhancement significantly improves the feasibility of model deployment.

E. Ablation Experiments

To assess the individual contributions of each component in our model to the overall performance, we conducted an ablation experiment on the nuScenes dataset. The results are summarized in Table III. Upon removing the camera pose encoding, the IoU drops to 31.0%. Similarly, eliminating the timing and image feature encoding in the attention mechanism results in an IoU of 33.0%. Removing both the time and camera pose encoding leads to a decrease in IoU to 30.9%. When the timing and image feature encoding are removed from the key in the attention mechanism, the IoU drops to 31.27%. From these findings, it can be inferred that the camera-aware embedding and the temporal attention module play crucial roles in enhancing the model's performance. Similarly, considering the performance in vehicle detection efficiency concerning the transformation in distance from ego-vehicle, as depicted in Fig. 14, similar ablation experiments were conducted to analyze the relationship between distance and detection accuracy.

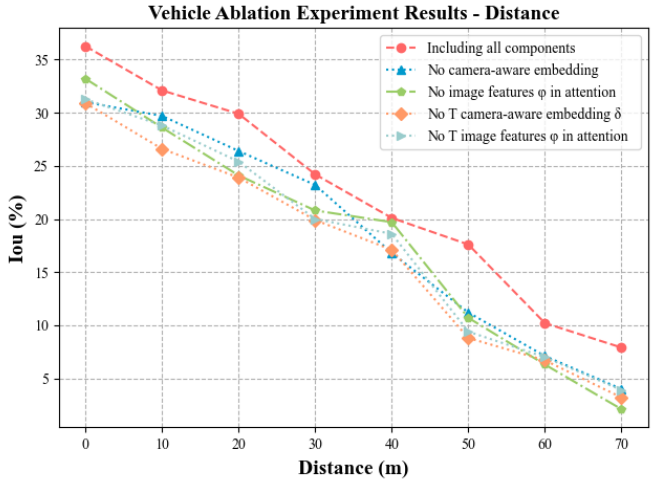


Fig. 14: The ablation experiments demonstrate a comparison of vehicle perception accuracy concerning the increase in distance from ego-vehicle.

In the ablation experiment conducted on road segmentation, we observe that eliminating the camera pose encoding results in an IoU of 73.96%. Removing the timing and image feature encoding in the key of attention leads to an IoU drop to 74.54%. Simultaneously removing both time and camera pose encoding results in an IoU of 71.39%. Lastly, when the timing and image feature encoding are eliminated from the key in the attention mechanism, the IoU drops to 69.92%. These results collectively underscore the effectiveness of our proposed continuous frame and underlying information fusion components.

To provide a more intuitive understanding of the effectiveness of the two proposed attention mechanisms, the perceptual visualization results of tracking attention are compared and presented in Fig. 10. We extracted continuous frames of attention feature data. Similarly, it can be intuitively observed that the proposed tracking attention method plays a significant role in the continuous tracking of images across different frames. As shown in Fig. 15, the ablation experiments illustrate the impact of distance transformation from ego-vehicle on road detection efficiency.

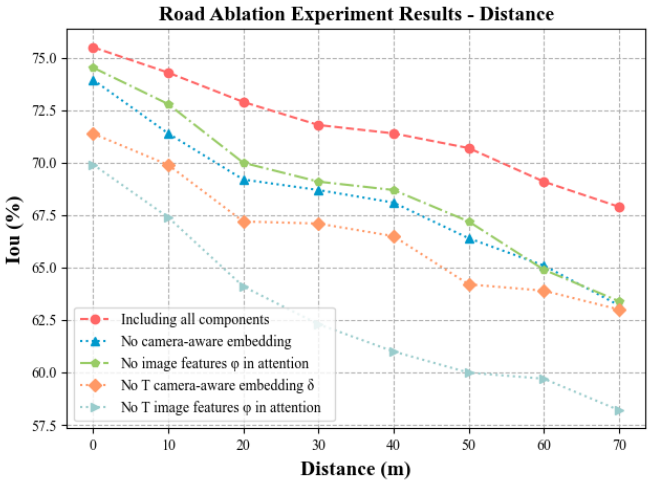


Fig. 15: The ablation experiments demonstrate a comparison of road perception accuracy concerning the increase in distance from ego-vehicle.

Concentrating attention solely on the overlapping area plays a crucial role in enhancing recognition accuracy. However, channeling all attention exclusively to this overlapping region may result in a substantial loss of computational power and resource utilization, ultimately diminishing overall recognition efficiency. Hence, our proposed attention mechanism strategically prioritizes the overlapping region, without exclusively emphasizing perceptual accuracy within this area. Instead, it takes into account other pertinent regions, preventing the concentration of attention solely in this specific area. This directional attention mechanism, as illustrated in Fig. 13, highlights areas of focus represented by bright white and yellow regions. The effectiveness of the proposed directional attention component becomes evident, emphasizing overlapping regions in the images compared to scenarios where this component is not utilized. This enhancement has significantly contributed to improved recognition efficiency and accuracy.

V. Conclusion

In this paper, we introduce a novel multi-modal end-to-end learning framework that leverages underlying information fusion coding for achieving multi-modal fusion coding of multiple camera and vehicle data for BEV semantic segmentation. Our approach surpasses previously published state-of-the-art methods in terms of multimodal fusion coding by utilizing underlying information from consecutive frames. Furthermore, by incorporating a multi-layer BEV decoder that encodes sequential frames and underlying information into BEV semantic results, we achieve even greater performance improvements. We conducted extensive ablation experiments to validate the effectiveness of our proposed method. In future research, we will focus on exploring more accurate and efficient techniques based on transformers and multimodal fusion for BEV semantic segmentation in the context of autonomous driving.

References

[1] L. Chen, Y. Li, C. Huang, B. Li, Y. Xing, D. Tian, L. Li, Z. Hu, X. Na, Z. Li, S. Teng, C. Lv, J. Wang, D. Cao, N. Zheng, and F.-Y. Wang, "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1046–1056, 2023.

[2] J. Guo, L. Chen, L. Li, X. Na, L. Vlacic, and F.-Y. Wang, "Advanced air mobility: An innovation for future diversified transportation and society," *IEEE Transactions on Intelligent Vehicles*, 2024.

[3] F.-Y. Wang and Y. Shen, "Parallel light fields: A perspective and a framework," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 542–544, 2024.

[4] L. Wang, Y. Ren, H. Jiang, P. Cai, D. Fu, T. Wang, Z. Cui, H. Yu, X. Wang, H. Zhou, H. Huang, and Y. Wang, "Accidentgpt: Accident analysis and prevention from v2x environmental perception with multi-modal large model," *arXiv preprint arXiv:2312.13156*, 2023.

[5] H. Yu, X. Liu, Y. Tian, Y. Wang, C. Gou, and F.-Y. Wang, "Sora-based parallel vision for smart sensing of intelligent vehicles: From foundation models to foundation intelligence," *IEEE Transactions on Intelligent Vehicles*, 2024.

[6] H. Yu, Y. Wang, Y. Tian, H. Zhang, W. Zheng, and F.-Y. Wang, "Social vision for intelligent vehicles: From computer vision to foundation vision," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 11, pp. 4474–4476, 2023.

[7] C. Chang, J. Zhang, K. Zhang, W. Zhong, X. Peng, S. Li, and L. Li, "Bev-v2x: Cooperative birds-eye-view fusion and grid occupancy prediction via v2x-based data sharing," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 11, pp. 4498–4514, 2023.

[8] F.-Y. Wang, Q. Miao, L. Li, Q. Ni, X. Li, J. Li, L. Fan, Y. Tian, and Q.-L. Han, "When does sora show: The beginning of tao to imaginative intelligence and scenarios engineering," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 4, pp. 809–815, 2024.

[9] S. Teng, L. Chen, Y. Ai, Y. Zhou, Z. Xuanyuan, and X. Hu, "Hierarchical interpretable imitation learning for end-to-end autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 673–683, Jan 2023.

[10] L. Luo, S.-Y. Cao, Z. Sheng, and H.-L. Shen, "Lidar-based global localization using histogram of orientations of principal normals," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 771–782, Sep. 2022.

[11] C. Lin, D. Tian, X. Duan, J. Zhou, D. Zhao, and D. Cao, "Da-rdd: Toward domain adaptive road damage detection across different countries," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3091–3103, 2023.

[12] Y. Ren, H. Jiang, X. Feng, Y. Zhao, R. Liu, and H. Yu, "Acp-based modeling of the parallel vehicular crowd sensing system: Framework, components and an application example," *IEEE Transactions on Intelligent Vehicles*, 2022.

[13] H. Mallot, H. Bülthoff, J. Little, and S. Bohrer, "Inverse perspective mapping simplifies optical flow computation and obstacle detection," *Biological Cybernetics*, vol. 64, no. 3, pp. 177–185, Jan. 1991.

[14] J. Jeong and A. Kim, "Adaptive inverse perspective mapping for lane map generation with slam," in *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, Nov 2016. [Online]. Available: http://dx.doi.org/10.1109/urai.2016.7734016

[15] J. Philion and S. Fidler, *Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D*, Nov 2020, p. 194–210. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-58568-6_12

[16] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nov 2021. [Online]. Available: http://dx.doi.org/10.1109/cvpr46437.2021.00845

[17] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.

[18] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5935–5943.

[19] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 2022, pp. 1–18.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[21] H. Zhang, G. Luo, X. Wang, Y. Li, W. Ding, and F.-Y. Wang, "Sasan: Shape-adaptive set abstraction network for point-voxel 3d object detection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.

[22] S. A. Abbas and A. Zisserman, "A geometric approach to obtain a bird's eye view from an image," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Mar 2020. [Online]. Available: http://dx.doi.org/10.1109/iccvw.2019.00504

[23] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Weinberger, "Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving," Apr 2020.

[24] X. Ma, S. Liu, Z. Xia, H. Zhang, X. Zeng, and W. Ouyang, *Rethinking Pseudo-LiDAR Representation*, Nov 2020, p. 311–327. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-58601-0_19

[25] Y. Kim and D. Kum, "Deep learning based vehicle position and orientation estimation via inverse perspective mapping image." in *2019 IEEE Intelligent Vehicles Symposium (IV)*, Aug 2019. [Online]. Available: http://dx.doi.org/10.1109/ivs.2019.8814050

[26] A. Palazzi, G. Borghi, D. Abati, S. Calderara, and R. Cucchiara, *Learning to Map Vehicles into Bird's Eye View*, Oct 2017, p. 233–243. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-68560-1_21

[27] M. Zhu, S. Zhang, Y. Zhong, P. Lu, H. Peng, and J. Lenneman, "Monocular 3d vehicle detection using uncalibrated traffic cameras through homography." in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Dec 2021. [Online]. Available: http://dx.doi.org/10.1109/iros51168.2021.9636384

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

[29] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, p. 4867–4873, Jun 2020. [Online]. Available: http://dx.doi.org/10.1109/lra.2020.3004325

[30] N. Hendy, C. Sloan, F. Tian, P. Duan, N. Charchut, Y. Xie, C. Wang, and J. Philbin, "Fishing net: Future inference of semantic heatmaps in grids," 2020.

[31] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Aug 2020. [Online]. Available: http://dx.doi.org/10.1109/cvpr42600.2020.01115

[32] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Enabling spatio-temporal aggregation in birds-eye-view vehicle estimation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, Oct 2021. [Online]. Available: http://dx.doi.org/10.1109/icra48506.2021.9561169

[33] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection."

[34] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "Petrv2: A unified framework for 3d perception from multi-camera images," pp. 3239–3249, 2023.

[35] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*. PMLR, 2022, pp. 180–191.

[36] W. Roh, G. Chang, S. Moon, G. Nam, C. Kim, Y. Kim, S. Kim, and J. Kim, "Ora3d: Overlap region aware multi-view 3d object detection," 2022.

[37] S. Chen, X. Wang, T. Cheng, Q. Zhang, C. Huang, and W. Liu, "Polar parametrization for vision-based surround-view 3d detection," *arXiv preprint arXiv:2206.10965*, 2022.

[38] Y. Tian, X. Li, H. Zhang, C. Zhao, B. Li, X. Wang, X. Wang, and F.-Y. Wang, "Vistagpt: Generative parallel transformers for vehicles with intelligent systems for transport automation," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 9, pp. 4198–4207, 2023.

[39] Y. Tian, L. Huang, H. Yu, X. Wu, X. Li, K. Wang, Z. Wang, and F.-Y. Wang, "Context-aware dynamic feature extraction for 3d object detection in point clouds," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 773–10 785, 2022.

[40] D. Gupta, W. Pu, T. Tabor, and J. Schneider, "Sbevnet: End-to-end deep stereo layout estimation," May 2021.

[41] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, Sep 2015, p. 234–241. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-24574-4_28

[42] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "Hdmapnet: An online hd map construction and evaluation framework," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4628–4634.

[43] K. Mani, S. Daga, S. Garg, S. S. Narasimhan, M. Krishna, and K. M. Jatavallabhula, "Monolayout: Amodal scene layout from a single image," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1689–1697.

[44] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," in *arXiv preprint arXiv:1904.07850*, 2019.

[45] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," 2019.

[46] M. ng, Y. Huo, H. Yi, Z. Wang, J. Shi, Z. Lu, and P. Luo, "Learning depth-guided convolutions for monocular 3d object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul 2020. [Online]. Available: http://dx.doi.org/10.1109/cvprw50498.2020.00508

[47] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Nov 2021. [Online]. Available: http://dx.doi.org/10.1109/iccvw54120.2021.00107

[48] T. Wang, Z. Xinge, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *Conference on Robot Learning*. PMLR, 2022, pp. 1475–1485.

[49] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," May 2019.

[50] Z. Rao, H. Wang, L. Chen, Y. Lian, Y. Zhong, Z. Liu, and Y. Cai, "Monocular road scene bird's eye view prediction via big kernel-size encoder and spatial-channel transform module," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[51] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," pp. 989–1000, 2023.

[52] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," pp. 13 760–13 769, June 2022.

[53] T. Zhang, Y. Sun, Y. Wang, B. Li, Y. Tian, and F.-Y. Wang, "A survey of vehicle dynamics modeling methods for autonomous racing: Theoretical models, physical/virtual platforms, and perspectives," *IEEE Transactions on Intelligent Vehicles*, pp. 1–24, 2024.

[54] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Aug 2020. [Online]. Available: http://dx.doi.org/10.1109/cvpr42600.2020.01164

[55] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," 2018.

[56] K. Mani, S. Daga, S. Garg, N. S. Shankar, J. Krishna Murthy, and K. M. Krishna, "Mono lay out: Amodal scene layout from a single image," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1678–1686.

[57] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

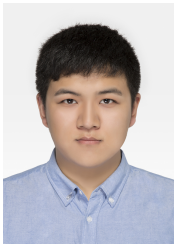
[58] R. Song, R. Xu, A. Festag, J. Ma, and A. Knoll, "Fedbevt: Federated learning bird's eye view perception transformer in road traffic systems," *IEEE Transactions on Intelligent Vehicles*, pp. 1–12, 2023.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3395272

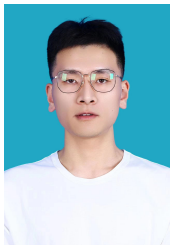
Yilong Ren received B.S. and Ph.D. degrees from Beihang University in 2010 and 2017, respectively. He is currently an Associate Professor at the School of Transportation Science and Engineering and the State Key Lab of Intelligent Transportation System, Beihang University. His research interests include vehicular communications, vehicular crowd sensing and traffic big data.

Lening Wang is currently pursuing the M.S. degree at the School of Transportation Science and Engineering and the State Key Lab of Intelligent Transportation System, Beihang University, Beijing, China. His current research interests include autonomous driving, computer vision, artificial intelligence, deep learning, and traffic big data.

Minda Li is currently pursuing the B.S. degree at the Navigation College of Dalian Maritime University, Dalian, China. His current research interests include autonomous driving, electrical engineering, artificial intelligence.



Han Jiang received his B.S. degree from Beihang University in 2017. He is currently pursuing his Ph.D. degree at the School of Transportation Science and Engineering and the State Key Lab of Intelligent Transportation System, Beihang University. His research interests include pervasive and mobile computing in intelligent transportation systems.



Chunmian Lin received the Ph.D degree in Electronic and Information from Beihang University, Beijing, China. He is currently a postdoctoral fellow in the School of Transportation Science and Engineering, Beihang University, Beijing, China. His current research interests include autonomous driving, image processing, computer vision, artificial intelligence and deep learning, particularly their applications in intelligent transportation systems.

Haiyang Yu received his Ph.D. degree in traffic environment and safety technology from Jilin University, China, in 2012. He is currently a Professor at the School of Transportation Science and Engineering and the State Key Lab of Intelligent Transportation System, Beihang University, China. His research interests include traffic big data, traffic control and intelligent vehicle infrastructure cooperative systems.

Zhiyong Cui received the B.S. degree in software engineering from Beihang University, Beijing, China, in 2012, the M.S. degree in software engineering and microelectronics from Peking University, Beijing, in 2015, and the Ph.D. degree in civil engineering from the University of Washington, Seattle, WA, USA, in 2021. He is currently an Professor with the School of Transportation Science and Engineering and the State Key Lab of Intelligent Transportation System, Beihang University. He was a University of Washington Data Science Postdoctoral Fellow with the eScience Institute, Seattle, WA, USA. His primary research interests include deep learning, machine learning, urban computing, traffic forecasting, connected vehicles, and transportation data science. He was the recipient of the IEEE ITSS Best Dissertation Award in 2021 and Best Paper Award at the 2020 IEEE International Smart Cities Conference.