electronics

MDPI

Article

Predicting Maps Using In-Vehicle Cameras for Data-Driven Intelligent Transport

Zhiguo Ma [1], Yutong Zhang [2] and Meng Han [1,*]

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China; 11921159@zju.edu.cn

[2] Innovation Center for Smart Medical Technologies & Devices, Binjiang Institute of Zhejiang University, Hangzhou 310053, China; yutongzhang@stu.sylu.edu.cn

* Correspondence: mhan@zju.edu.cn

Abstract: Bird's eye view (BEV) semantic maps have evolved into a crucial element of urban intelligent traffic management and monitoring, offering invaluable visual and significant data representations for informed intelligent city decision making. Nevertheless, current methodologies continue underutilizing the temporal information embedded within dynamic frames throughout the BEV feature transformation process. This limitation results in decreased accuracy when mapping high-speed moving objects, particularly in capturing their shape and dynamic trajectory. A framework is proposed for cross-view semantic segmentation to address this challenge, leveraging simulated environments as a starting point before applying it to real-life urban imaginative transportation scenarios. The view converter module is thoughtfully designed to collate information from multiple initial view observations captured from various angles and modes. This module outputs a top-down view semantic graph characterized by its object space layout to preserve beneficial temporal information in BEV transformation. The NuScenes dataset is used to evaluate model effectiveness. A novel application is also devised that harnesses transformer networks to map images and video sequences into top-down or comprehensive bird's-eye views. By combining physics-based and constraint-based formulations and conducting ablation studies, the approach has been substantiated, highlighting the significance of context above and below a given point in generating these maps. This innovative method has been thoroughly validated on the NuScenes dataset. Notably, it has yielded state-of-the-art instantaneous mapping results, with particular benefits observed for smaller dynamic category displays. The experimental findings include comparing axial attention with the state-of-the-art (SOTA) model, demonstrating the performance enhancement associated with temporal awareness.

Keywords: BEV; urban intelligent traffic management; view semantic graph

check for updates

Citation: Ma, Z.; Zhang, Y.; Han, M. Predicting Maps Using In-Vehicle Cameras for Data-Driven Intelligent Transport. *Electronics* **2023**, *12*, 5017. https://doi.org/10.3390/electronics12245017

Academic Editors: Felipe Jiménez and Fabio Grandi

Received: 30 October 2023
Revised: 11 December 2023
Accepted: 12 December 2023
Published: 15 December 2023

1. Introduction

Map prediction is crucial in intelligent transportation, especially in bird's eye view (BEV) map generation, which leverages in-vehicle cameras with real-time capabilities [1]. BEV map generation [2] is a pivotal task within intelligent transportation, offering vital data for environmental perception and path planning in autonomous driving systems. As shown in Figure 1, six more comprehensive perspectives are used to capture moving vehicles and passing pedestrians.

Two primary methods for BEV map generation exist, one based on target detection [3] and the other on semantic segmentation [4]. In the target detection approach, a deep neural network identifies obstacles in images and then maps their locations onto the BEV map using external camera and radar parameters [5]. On the other hand, the semantic segmentation method involves a deep neural network classifying images at the pixel level [6], then projecting the classification outcomes onto the BEV map using the same external parameters [7]. Each method has its merits and demerits. Target detection is more precise in locating obstacles but may overlook small or occluded objects [8]. In contrast, semantic segmentation is comprehensive but may need more clarity or misclassifications.

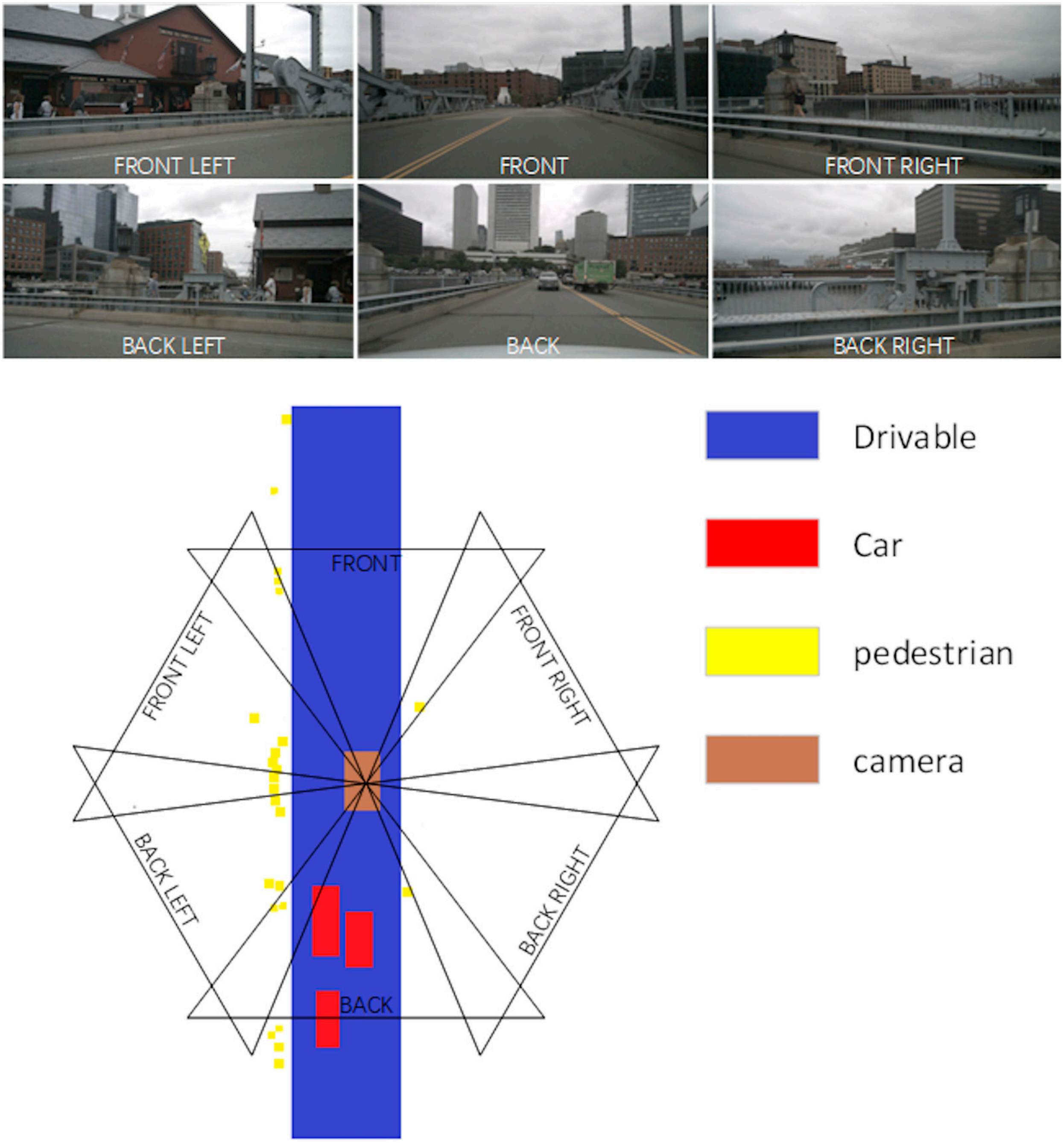

Figure 1. Own compilation based on using in-vehicle cameras to predict maps.

BEV is a technique that transforms images from the camera perspective to the bird's-eye view, which can be used for semantic segmentation, vehicle detection and tracking, and other tasks in intelligent transportation. BEV can provide a more intuitive and global scene representation, which helps to understand the traffic flow and behavior. The key challenge of BEV is how to deal with the semantic and positional uncertainty in the images, and how to perform effective transformation between different planes. The above content describes a method that uses a neural network model to solve these problems, which can predict the binary variables for each class, the transformation matrix between planes, and the semantic segmentation on the BEV surface from the input image. This method is end-to-end and does not require additional annotation or prior knowledge. This method has a wide research scope in intelligent transportation, and it can be extended to multiple cameras, dynamic scenes, and different road types. This method also has an impact on the early work on BEV, such as using geometric transformation, projection matrix, or depth estimation to achieve BEV. These works have some limitations, such as the dependence on the camera parameters, the assumption of the scene, or the demand for computational resources. This method tries to overcome these limitations, and improve the accuracy and robustness of BEV, to further promote the technology in urban multimodal transportation.

Recent research focusing on converting images into BEV maps is on the cutting edge [9–12], with relatively few reports available. This approach utilizes monocular or binocular cameras to capture road scenes and transforms them into a bird's-eye view through projection. This significantly enhances the perception capabilities of autonomous driving systems by providing more spatial information, including vehicle location, direction, speed, road geometry, and topology [13]. Image-to-BEV map research is broadly categorized into geometry-based and deep learning-based methods. Geometry-based methods necessitate prior knowledge of camera parameters and road plane equations for transformation based on perspective projection principles [14]. Deep learning-based methods, however, employ neural networks to learn transformation functions directly from images, eliminating the need for explicit camera parameters or road plane assumptions. Each method has its strengths and weaknesses, with simpler geometry-based approaches requiring accurate calibration and road plane assumptions. In contrast, deep learning-based methods [15] offer flexibility and robustness but demand substantial training data and computational resources. Nonetheless, these methods primarily cater to real-time in-vehicle navigation and have limitations in predicting small and medium-sized targets at long distances, dealing with unclear outlines, and addressing position drift. This limits their applicability in data-driven smart transportation systems [16]. To address this issue, we posit that converting image features into BEV features, particularly in the top-down process, does not sufficiently mine small-scale spatial information.

Our alignment model is fundamentally geared towards understanding the alignment relationship between the vertical scan lines present within the image and the polar coordinates in the BEV. We adopted a Transformer-based alignment model with a bidirectional transformer model which includes modifying the attention mechanism so that the model can better account for contextual information on both sides of the input sequence. This can be achieved by modifying the way the attention score is calculated so that the model considers both the first and second half of the input sequence when calculating contextual weights. This meticulously captures the pairwise interaction between a scanline and its corresponding polar coordinates within the BEV projection to achieve this objective. The Transformer architecture is particularly well suited for addressing image-to-BEV translation challenges. Its ability to reason about the complex interdependencies among objects, depth, and scene lighting results in globally consistent representations. Within our approach, this Transformer-based alignment model is seamlessly integrated into an end-to-end learning framework. This framework takes as input a monocular image and its internal matrix, enabling it to predict semantic BEV maps for both static and dynamic categories, as shown in Figure 1. Predicting semantic BEV maps for both static and dynamic categories is a challenging task because it involves dealing with various sources of uncertainty and complexity. Some of the possible challenges are (1) Static categories, such as buildings, roads, and sidewalks, may have different shapes, sizes, and orientations in the image plane, which require accurate estimation of their depth and boundaries to map them to the BEV surface. (2) Dynamic categories, such as cars, pedestrians, and cyclists, may have occlusions, motion blur, and varying poses in the image plane, which require robust detection and tracking to map them to the BEV surface. (3) The transformation between the image plane and the BEV surface may depend on the camera parameters, the scene geometry, and the road layout, which may not be known or consistent in different scenarios. (4) The semantic BEV map may have different resolutions, scales, and perspectives, which require adaptive and flexible representation and visualization.

2. Related Work

2.1. Map Generation in Intelligent Transportation

Map generation is a critical issue within intelligent transportation, with far-reaching navigation, planning, and safety implications [17]. Currently, various methods are employed for map generation [18–20], with the most prevalent ones involving remote-sensing satellites and vehicle-mounted cameras.

Remote sensing satellites utilize sensors on board the satellite, including optical cameras, radars, or laser scanners, to capture images and data from the ground [21]. Subsequently, these data are processed and analyzed to generate maps [22]. This approach offers distinct advantages, as it can cover expansive areas, deliver high-resolution and multi-spectral information, and remain unaffected by weather and lighting conditions. Nonetheless, it has drawbacks, such as extended data transmission and processing times [23], infrequent map updates, and the inability to depict specific details accurately and rapidly changing information, such as real-time road conditions and traffic flow [24].

In contrast, vehicle-mounted cameras strategically positioned on vehicles [25], including front and rear-view mirrors or dashboard cameras, capture real-time images and data from the road. These data are then processed and analyzed to create maps. The primary strength of the method lies in its capacity to deliver real-time road information [26], effectively reflecting dynamic changes in the map and offering detailed features like road signs, traffic signals, and lane demarcations. However, it is not without drawbacks, as image quality may fluctuate due to weather and lighting conditions, and it necessitates the participation of many vehicles to ensure data consistency and accuracy.

In this context, LiDAR data, as an emerging technology, introduces new possibilities for map generation. LiDAR can provide high-precision three-dimensional maps that are not limited by lighting and weather conditions, so it is widely used in BEV-based intelligent transportation scenarios. Machine learning algorithms can be trained on LiDAR data to identify and segment roads, vehicles, pedestrians, etc., allowing for more accurate and real-time map generation. This method overcomes some limitations of traditional methods while providing richer information and bringing new opportunities for the development of intelligent transportation systems.

2.2. BEV Maps Prediction

BEV map prediction is a methodology that employs bird's-eye view maps to estimate the position and orientation of three-dimensional objects [27]. This approach significantly enhances the perception capabilities of autonomous driving systems, contributing to improved safety and efficiency.

The fundamental principle behind BEV map prediction involves projecting three-dimensional objects onto a two-dimensional plane, simplifying the challenges associated with object detection and tracking [28]. The process of BEV map prediction encompasses several key steps: First, acquiring raw data from sensors like lidar, cameras, radar, and more, and performing preprocessing tasks such as filtering, calibration, and registration. Next, the system extracts features like edges, corner points, colors, textures, and others [29,30] from the raw data. These features are then used for object detection, identifying the category, location, and size of the object. The results of object detection are subsequently employed for object tracking, facilitating the estimation of the motion state and posture of the object. Finally, the tracked object data are projected onto the BEV map, and further processing steps like fusion, filtering, and optimization are carried out.

BEV map prediction boasts several advantages, notably its ability to effectively reduce data volume and computational complexity. It enhances the precision and robustness of detection and tracking processes, while also improving the understanding and visualization of the scene. However, the primary challenge in BEV map prediction lies in the necessity to coordinate and fuse information from various sensors. Additionally, addressing issues like occlusion, overlap, and noise is critical for ensuring the accuracy of the predictions.

2.3. Attention-Powered Image Translating

Attention-powered image translation is a methodology that leverages the attention mechanism to facilitate the transformation of images across various domains. The attention mechanism initially found its roots in natural language processing (NLP) to address the alignment problem within sequence-to-sequence (seq2seq) models. The core concept behind the attention mechanism is enabling the model to concentrate on the most pertinent sections of the input data when generating output, thereby improving the performance and interpretability of the model.

As deep learning has advanced, the attention mechanism has entered computer vision (CV) to handle various image-to-image conversion tasks, including style transfer, image restoration, and image super-resolution. Attention-powered image translation, a framework for image-to-image translation, capitalizes on the attention mechanism. It can produce an output image that aligns with the input image while adopting the desired style based on its content and the style of the target domain. One of the critical advantages of attention-powered image translation is its ability to perform image conversion across multiple domains without requiring paired data, while retaining the structure and intricate details of the input image throughout the process.

In image translation, the attention mechanism helps the model emphasize essential areas within the input image, facilitating the generation of an output image that aligns with the target domain. When applied to tasks like target detection and semantic segmentation, the attention mechanism aids the model in extracting local and global image features and enhancing feature interaction.

Several articles have effectively harnessed the attention mechanism for image translation, target detection, and semantic segmentation. "Attention-Guided Image-to-Image Translation with Adversarial Learning" [29]: This article introduces an image-to-image translation approach guided by attention, employing attention maps to direct the generator in generating an output image consistent with the target domain. Additionally, it uses a discriminator to supervise the attention map to enhance attention map quality and interpretability. "Attention-Aware Feature Pyramid Network for Object Detection" [30]: This article proposes a feature pyramid network based on attention perception. Incorporating attention modules across various feature pyramid levels strengthens the information flow and fusion between features, consequently boosting target detection performance. "Attention-Guided Semantic Segmentation with Cross-Attention and Self-Attention" [31]: This article outlines a semantic segmentation method guided by attention. It introduces a cross-attention module between the encoder and decoder to achieve alignment between the input image and the output segmentation map. Furthermore, it incorporates a self-attention module within the decoder to ensure internal consistency of the output segmented maps.

2.4. Small-Size Object Prediction in Computer Vision Tasks

Object detection of small targets is an important research direction in computer vision, which aims to identify small-sized objects from images or videos and give their locations and categories [32]. This task is significant in many practical applications, such as medical image analysis, driverless driving, and security monitoring. However, target detection of small targets also faces many challenges, such as low resolution [33], occlusion, background interference, and category imbalance. These factors lead to the performance of target detection of small targets being much lower than that of large targets [34], especially in complex scenes. In order to solve these problems, researchers have proposed many methods in recent years, including improving feature extraction, enhancing feature fusion [35], designing specialized loss functions, introducing attention mechanisms, and utilizing multi-scale information. However, these methods still cannot completely solve the difficulty of target detection of small targets because small targets themselves lack sufficient information and distinguishability. Therefore, target detection of small targets is still a direction worthy of further research.

Semantic segmentation is an essential task in computer vision, which aims to assign each pixel in an image to a category, thereby achieving a detailed understanding of the image content. Semantic segmentation of small objects is a challenging sub-problem of semantic segmentation [36], which involves detecting and segmenting objects with small sizes, large numbers, and different shapes in images, such as crowds, cells, and particles. The research on semantic segmentation of small targets has important theoretical significance and practical value. It can be applied to medical image analysis, remote sensing image interpretation, intelligent monitoring, and other fields. However, semantic segmentation of small targets also faces many difficulties, mainly including the size of small targets being much smaller than the receptive field [24], resulting in insufficient feature extraction and inaccurate classification; the distance between small targets is very close, and occlusion and confusion are prone to occur. It is challenging to distinguish boundaries; small targets have various shapes and need a unified expression, making it difficult to establish a practical model [37]; and the number of small targets is enormous, which increases computational complexity and memory consumption and reduces efficiency.

There currently needs to be a perfect solution to these problems. Most existing methods are based on deep learning, using convolutional neural networks (CNN) to extract image features and combining different strategies [38] to improve the detection and segmentation of small targets. For example, some methods capture targets of different sizes by designing multi-scale or pyramid-structured networks; some methods enhance the feature representation of small targets by introducing attention mechanisms or adaptive sampling; and some methods use conditional random fields (CRF) or generative adversarial networks (GAN) to optimize boundary details of small objects. However, these methods still have some limitations, such as sensitivity to hyperparameters, needing to be more robust to noise and illumination changes, and high demand for training data and computing resources. Therefore, semantic segmentation of small objects is still an open and exciting research direction worthy of further exploration and innovation.

3. Method

The input is N × M first-view observations sampled from a spatial location in a 3D environment, where N is different angles and M is different modalities (e.g., RGB images and depth images). The output is a top-down view semantic map, which is a map captured by a camera from a certain height from top to bottom, and each pixel is annotated with a semantic label. What we define is cross-view semantic segmentation, i.e., given the first-view observation as input, the algorithm must generate a semantic map from the top-down view. From semantic segmentation, the categories, locations, shapes and attributes of different objects in the environment, as well as the relationships between them, can be obtained. There are 14 semantic categories in this experiment, including roads, sidewalks, buildings, trees, cars, pedestrians, etc. At the same time, the relationship between the plane and the BEV is obtained through the perspective transformation module (VTM). The VTM transforms the feature map of the first perspective from the first perspective space to the top-down perspective feature space and fuses them into. A final feature map is then decoded by a decoder into a top-down semantic map.

3.1. Image Feature Translating

As shown in (1), Y represents the formulation scenario, φ represents the learning neural network model trained to resolve semantic and positional uncertainties, and I generates a semantic segmentation bird's-eye view of Y, a matrix $\mathtt{C} \in R^{3\times 3}$ including the input image $I \in R^{3\times H\times W}$. Next, we predict a set of binary variables $k \in K$ for each class, transformations between planes P^{I} and BEV surfaces P^{BEV}. A representation that only encodes semantics and depth is constructed in the image plane. We use an end-to-end method to perform semantic segmentation on BEV to achieve transformation from P^{I} to

P^{BEV}. I represents the image in intelligent transportation. Mapping the image to the BEV surface requires a mapping to determine the relationship between the pixels and the BEV aurora. There is a one-to-one correspondence between each random vertical scan line and the relevant ray. As shown in the Figure 2, the discretized static depth of the element on the vertical scan line of the image is at most r meters away from the camera. In the image sequence $S^I \in R^H$, H here represents the height of the column. Find the BEV ray $S^{\phi(BEV)} \in R^d$, and d here represents the radial direction from the camera. Distance, $S_i^{\phi(BEV)}$, represents radial elements.

$$P\left(Y^k \middle| I, c\right) = \phi(I, c) \tag{1}$$

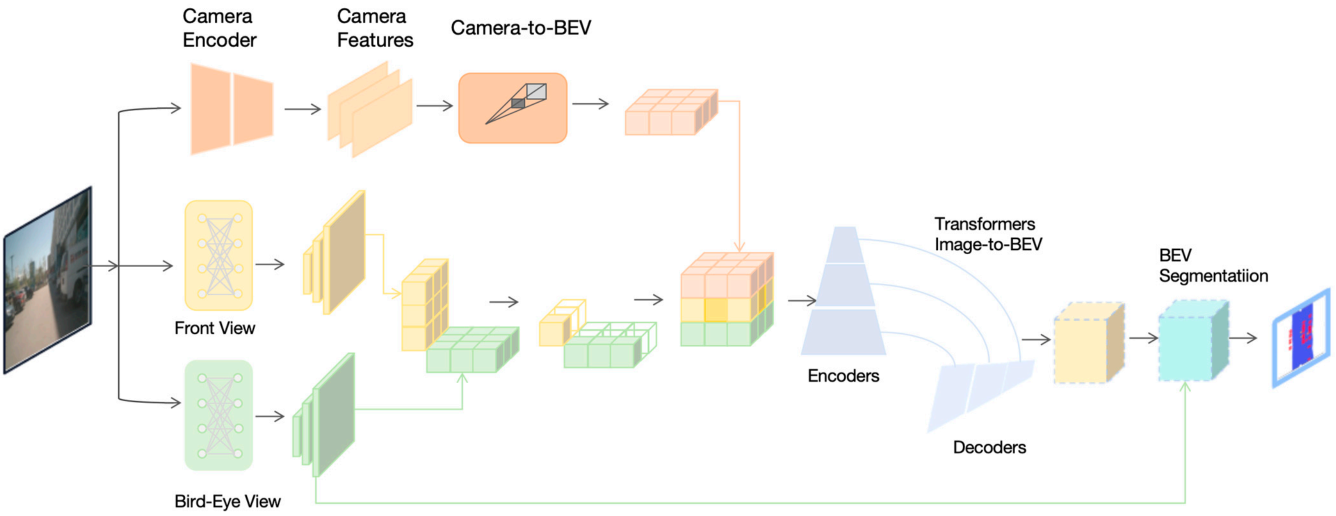

Figure 2. Own compilation framework for cross-view semantic segmentation, leveraging simulated environments as a starting point before applying it to real-life urban imaginative transportation scenarios.

This mapping can be viewed as the process of assigning the semantic objects in P^I to their position slots on the BEV plane and on the ray. This includes learning the alignment between input scan lines and output polar rays via an attention mechanism.

Semantically segmented image columns and their corresponding polar BEV ground truths, the relationship between columns and ground truth rays mentioned here is hard aligned (explicit assignment of each pixel to a unique semantic category, i.e., each pixel can only belong to a class), i.e., each pixel in the ray corresponds to a single semantic class in the image column, so we have to resolve the only uncertainty: the depth of each pixel.

However, we must assign features that help solve semantics and depth, so the challenging alignment task is unsuitable now. Instead, we need soft alignment (soft alignment methods allow each pixel to be assigned to multiple semantic categories with a certain probability and, therefore, can produce richer segmentation results, reflecting the uncertainty of the segmentation results), where each of the polar rays pixels is assigned a combination of elements in the image column, a context vector.

Specifically, when generating radial elements, we use convex combinations of elements in image columns S^I (they are used in many aspects such as weights, sparse representations, mixture distributions, and optimization problems. By rationally using convex combinations, we can better establish and solve problems in deep learning models.) and the radial position r_i of elements $S_i^{\phi(BEV)}$ along polar rays. $h \in R^{H \times C}$, where h represents the input sequence, and H represents the height of the image column. $y \epsilon R^{e \times c}$ represents a position query encoding the relative position along a polar ray of length e, where c represents the context. Generate context c based on input sequence h generates context based on input sequence h and query y. The input sequence h and query y are projected through $W_Q \epsilon R^{C \times D}$

and $W_k \epsilon R^{C\times D}$ to the corresponding representation of query Q and key K, as shown in (2) and (3):

$$Q_{(y_i)} = y_i \times W_Q \tag{2}$$

$$K_{(h_i)} = h_i \times W_K \tag{3}$$

Generate unstandardized alignment scores after projection $e_{i,j}$ in (4).

$$e_{i,j} = \frac{< Q(y_i), K(h_i) >}{\sqrt{D}} \tag{4}$$

Then, as shown in (5), we use Softmax to normalize the energy scalar to generate a probability distribution in memory:

$$\alpha_{i,j} = \frac{exp\left(e_{i,j}\right)}{\sum_{k=1}^{H} exp\left(e_{i,k}\right)} \tag{5}$$

The weighted sum of the final context k is computed using the context phase volume:

$$c_i = \sum_{j=1}^{H} \alpha_{i,j} K\left(h_j\right) \tag{6}$$

In this way, the generation context allows each radial groove r_i to independently collect information from the image column and represents the initial assignment from the image to its BEV corresponding position group. This approach is similar to boosting pixels based on depth. However, it is promoted to a depth distribution and is therefore able to overcome the common pitfalls of sparsity and elongated object frustums. This means that the image context available to each radial slot is decoupled from its camera distance.

Finally, in order to generate BEV features, $S_m^{\phi(BEV)}$ at radial position r performs a global operation on the specified context for all radial positions c = $\{c_1 \ldots c_m\}$.

3.2. Encoder and Decoder

From a spatial location in the 3D environment, we first sample N first-view observations from N angles and M and modalities (N = 6, M = 2) at even angles to capture all-round information in intelligent transportation. N first view observations are encoded by M encoders respectively for M corresponding modalities. The first view observations are encoded by M encoders for M corresponding modalities. These CNN-based encoders extract N spatial feature maps for the first graph input, and then all these feature maps are fed into the view transformer module (VTM).

The view converter module converts these view features from the first view space to the top-down view feature space and fuses them to obtain a final feature map, which already contains sufficient spatial information. Finally, convolutional decoder alignment is used for decoding to predict top-down view semantic maps.

However, the encoder-decoder architecture has succeeded in classic semantic segmentation; our experiments show that it performs poorly in cross-view semantic segmentation tasks. This is because the receptive field of view of the output spatial feature map is roughly aligned with the input spatial feature map in standard semantic segmentation architectures.

However, in cross-view segmentation, each pixel painted in the top-down view should consider all input first-view feature maps, not just the local receptive field region. After considering the shortcomings of current semantic segmentation structures, we designed the View Transformer Module (VTM) to learn the dependencies of all spatial positions between the first view feature map and the top-down view feature map. VTM does not change the shape of the input feature map so that it can be plugged into any existing encoder-decoder-type network architecture for classical semantic segmentation. It consists of the View Relationship Module (VRM) and View Fusion Module (VFM). The central plot of Figure 2 illustrates the entire process: the first view feature map is flattened while the channel dimensions remain unchanged. Then, we use the view relationship module R to learn the relationship between any two-pixel positions in the flattened first-view feature map and the flattened top-down view feature map.

$$f_t[m] = R_m(f[1], \dots, f[n], \dots f[HW]) \tag{7}$$

where $m, n \in [0, HW)$ are the indices along the flat dimension of the top-down view feature map $t \in R^{HW \times C}$ and the first view feature map $f \in R^{HW \times C}$, respectively, R_i models the relationship between the m-th pixel on the top-down view feature map and each pixel on the first-view feature map. Here, we simply use the multilayer perceptron (MLP) in the view relation module R. Afterwards, the top-down view feature map is reshaped back to H × W × C. Each first view input has its own VRM to obtain the top-down view feature map $t^m \in R^{H \times W \times C}$ based on its own observations.

We use VFM to fuse these top-down view feature maps to aggregate information from all observation inputs. View encoders and decoders. To balance efficiency and performance, we use ResNet-18 as the encoder. We remove the last Residual Block and Average Pool layers so that the resolution of the encoded feature map remains large, thus better preserving the details of the view. We adopt the pyramid pooling module used in [39] as the decoder. Regarding the view transformer module, for each view relationship module, we simply use a two-layer MLP. We chose this because a two-layer MLP does not bring much extra computation, so we can make our model follow the principle of being lightweight and efficient. The input and output dimensions of VRM are both HIWI, where HI and WI are the height and width of the intermediate feature map, respectively. As for the view fusion module, we simply sum all features to keep the shape consistent.

Simulation to reality. For generator G, we use the architecture of a 4-view VPN. We adopt the same architecture as in [40] for discriminator D. It has five convolutional layers, each followed by a leaky ReLU with a parameter of 0.2 (except the last layer). We extract semantic masks from real-world images using HRNet [41] pre-trained on the nuScenes [42] dataset.

3.3. Loss Design

Since the training signal provided to the predicted occupancy grid must resolve semantic and positional uncertainties, we use the same multi-scale Dice loss. The average dice loss for K classes at each scale u is:

$$L_u = 1 - \frac{1}{k} \textstyle\sum_{K=1}^{K} \frac{2 \sum_m^N \hat{y_l^k} y_m^k}{\sum_m^N \hat{y_l^k} + y_m^k + \sigma} \tag{8}$$

where y_m^k is the ground truth binary variable grid cell, and $\hat{y_l^k}$ is the predicted sigmoid output of the network and is a constant used to prevent division by zero.

During the training phase, we forward a set of input images from the source target domain $\{I_s\}$ to ζ and optimize them with the commonly used segmentation loss ζ_{seg}.

Then, we use ζ to extract the feature map of the image after passing through the Softmax layer from $\{I_t\}$ and F_i also uses the discriminator to distinguish whether F_t comes from the source domain. The loss function for optimizing ζ can be expressed as shown in (9):

$$\zeta(\{I_s\}, \{I_t\}) = \zeta_{seg}(\ \{I_s\} + \lambda_{adv}\zeta_{adv}(\{I_t\})) \tag{9}$$

where ζ_{seg} is the cross-entropy loss of semantic segmentation, and ζ_{adv} is designed to train ζ and interfere with the discriminator D. The loss function of the discriminator ζ_d is the cross-entropy loss of the binary source.

4. Experiments and Discussion

4.1. Experimental Settings and Dataset

We use the NuScenes dataset to evaluate the effectiveness of treating image-to-BEV conversion as a translation problem. The dataset contains 1000 20 s video clips shot in Boston and Singapore, annotated with 3D bounding boxes and vectorized road maps. Each data sample in NuScenes contains first-view RGB images taken from six directions (front, front right, back right, back, back left, front left), as well as different modalities. We selected 919 data samples without top-view masks for unsupervised training and 515 data samples with binary top-view masks for evaluation. We conducted ablation experiments on the lookback direction, the role of long-distance horizontal context, and the influence of polar position information in monotonic attention. For the generator ζ, we use the architecture of 4-view VPN. It has five convolutional layers, and each convolutional layer except the last layer is followed by a leaky ReLU with a parameter of 0.2.

Research in intelligent transportation has been pursuing more accurate vehicle, pedestrian, and road perception to improve the performance and safety of autonomous driving systems. This article relies on multiple sensors onboard the vehicle, including six ring-shaped RGB cameras. These cameras are mounted around the vehicle, providing an all-round view. Specifically, these cameras can capture images around the road in real time, forming a complete ring view to help vehicle systems perceive their surroundings. In order to better understand and utilize these image data, metrics like "IOU" (Intersection over Union) are often used to evaluate the performance of this BEV map generation algorithm. IOU is a metric used to measure the degree of overlap between an algorithm-generated map and a real map. In this case, IOU can measure the overlap between objects (such as vehicles and pedestrians) in the algorithm-generated BEV map and objects in the real world. A higher IOU value means that the map generated by the algorithm is closer and more accurate to the real situation. Therefore, the goal of this research is to develop a high-quality BEV map generation algorithm by utilizing the six ring-shaped RGB cameras on the vehicle to perceive the main view of the road surface in real time and use evaluation indicators such as IOU to measure its performance, thereby improving the performance of the autonomous driving system and environmental awareness and path planning capabilities. These works are significant to achieve safer and more efficient intelligent transportation systems.

The positioning accuracy can be measured by the overlap between the generated map and the real map, the intersection-over-union (IOU) ratio.

The P–R curve is a two-dimensional curve with precision and recall as the vertical and horizontal axis coordinates. The more convex and to the upper right, the better the effect. The overall trend is drawn by selecting the corresponding precision and recall rates at different thresholds. The higher the precision, the lower the recall. When the recall reaches 1, it corresponds to the positive sample with the lowest probability score. At this time, the number of positive samples is divided by all those greater than or equal to the threshold. The number of samples is the lowest precision value. The area enclosed by the P–R curve is the AP value. The larger the area, the higher the recognition accuracy. The higher the AP value, and vice versa.

In target detection, each category can draw a P–R curve based on recall and precision. AP is the area under the curve, and mAP is the average AP of all categories.

4.2. Comparison Results and Discussion

In this paper, we evaluate the performance of different semantic segmentation models in urban scenarios. We used three metrics: drivable (accuracy in drivable areas), car (accuracy in cars), and ped (accuracy in pedestrians). We compare four existing models: VPN, PON, STA-ST, and TIIM-ST, as well as our proposed new model. Our model outperforms other models on all metrics, indicating that our model is better able to capture detailed and semantic information in urban scenes.

Specifically, as shown in Table 1, our model achieves 78.2% on drivable, which is 3.7% higher than the best existing model TIIM-ST; 40% on car, which is 0.3% higher than TIIM-ST;

and on ped, it reached 10.2%, 0.7% higher than TIIM-ST. These results illustrate that our model has strong robustness and generalization ability when dealing with complex and diverse urban scenes. At the same time, Figure 3 can more intuitively show the superiority of our proposed method for small target prediction compared with existing methods.

Table 1. The implementation of accurate perception of vehicles, pedestrians, and roads in urban roads under intelligent transportation relies on the six ring RGB cameras of the vehicle to perceive the RGB main view of the road in real time, studies the BEV map generation algorithm, and evaluates the index IOU.

Model	Drivable	Car	Ped
VPN [43]	58.0	25.5	7.1
PON [44]	60.4	24.7	8.2
STA-ST [45]	70.7	36.0	8.6
TIIM-ST [46]	74.5	39.7	9.5
Our	78.2	40.0	10.2

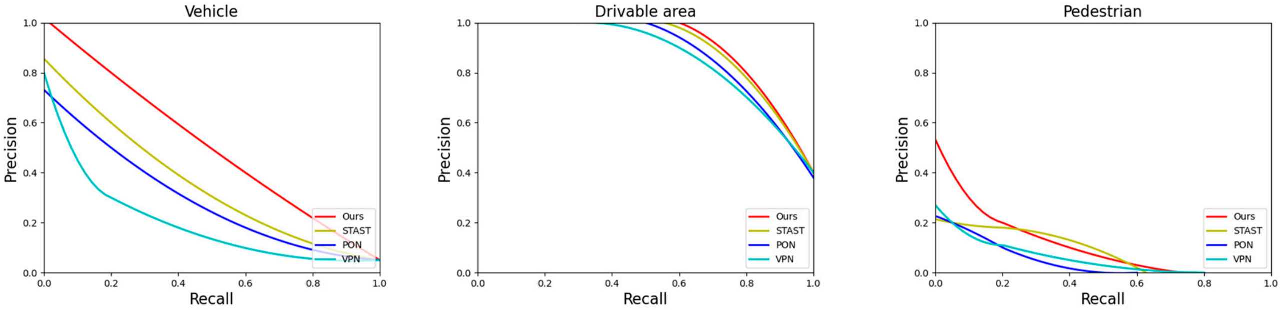

Figure 3. Study the BEV map generation algorithm and evaluate the index IOU (higher the better).

4.3. Qualitative Results and Discussion

As can be seen in Figures 4 and 5, from the day and night effect display pictures, our model can generate clear and realistic scene images under different lighting conditions. Our model can also mark the positions and trajectories of vehicles and pedestrians in images, as well as information such as roads and lane lines, to provide drivers with more visual references. In contrast, other models generated images that were either too blurry or had obvious distortion or artifacts that did not reflect the real scene well. Our method can adapt to different sensor configurations and data sources, such as lidar, cameras, radar, etc. Our method does not rely on specific sensor types or data formats but utilizes multi-modal data fusion technology to effectively integrate and represent data collected by different sensors. This enables our method to be tested on the NuScenes dataset and has good generalization ability and robustness.

Our method can handle complex traffic scenarios and multi-objective interaction problems, such as avoidance, overtaking, lane changing, etc., between vehicles. Our method not only considers the motion status and target of a single vehicle, but also considers the influence and feedback of surrounding vehicles and pedestrians. Our method uses a model based on attention mechanism and graph neural network to capture the relationships and dependencies between different targets and generate more accurate and safer trajectory predictions based on this information.

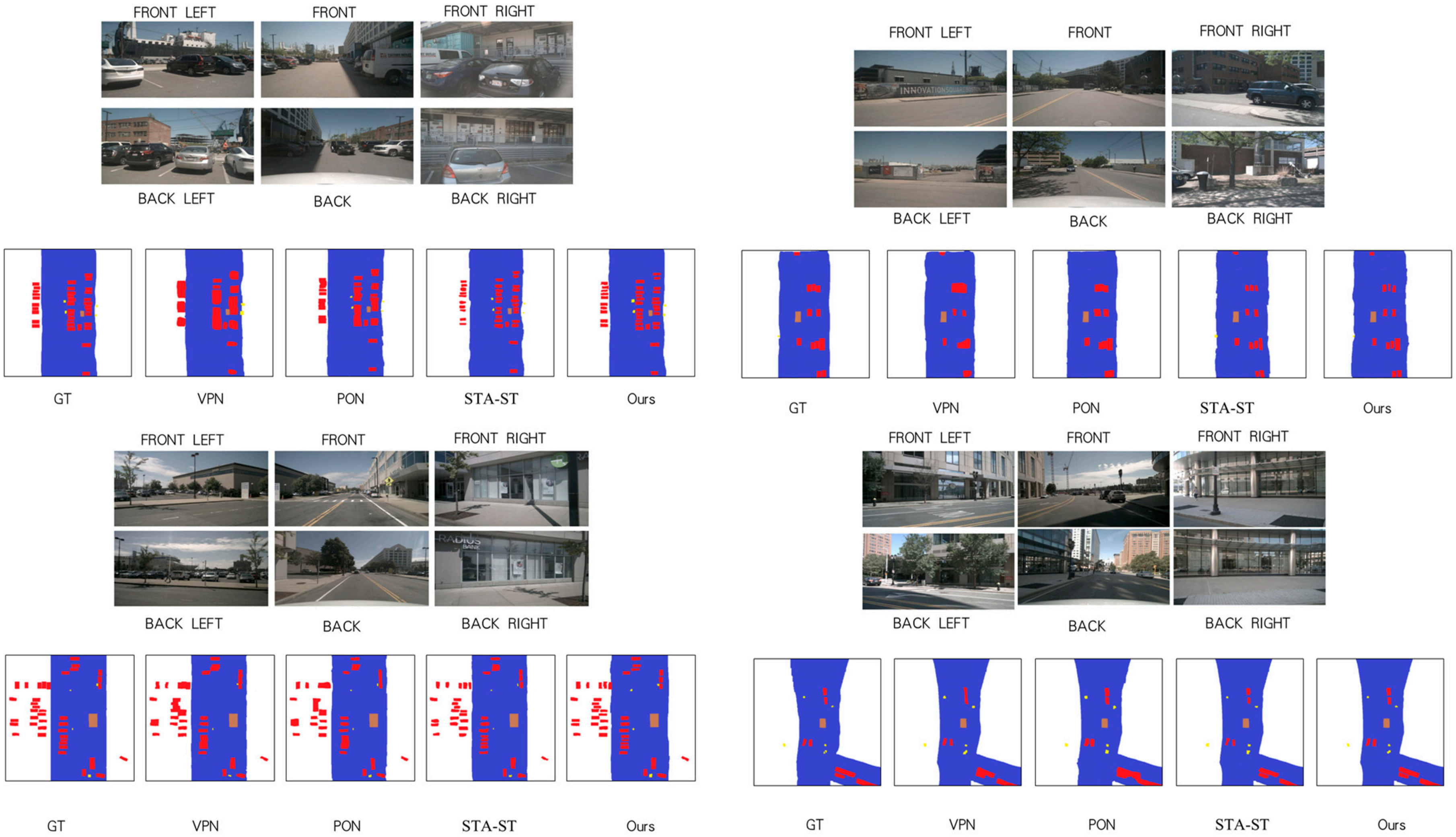

Figure 4. Own compilation qualitative results on the validation set of NuScenes during the day.

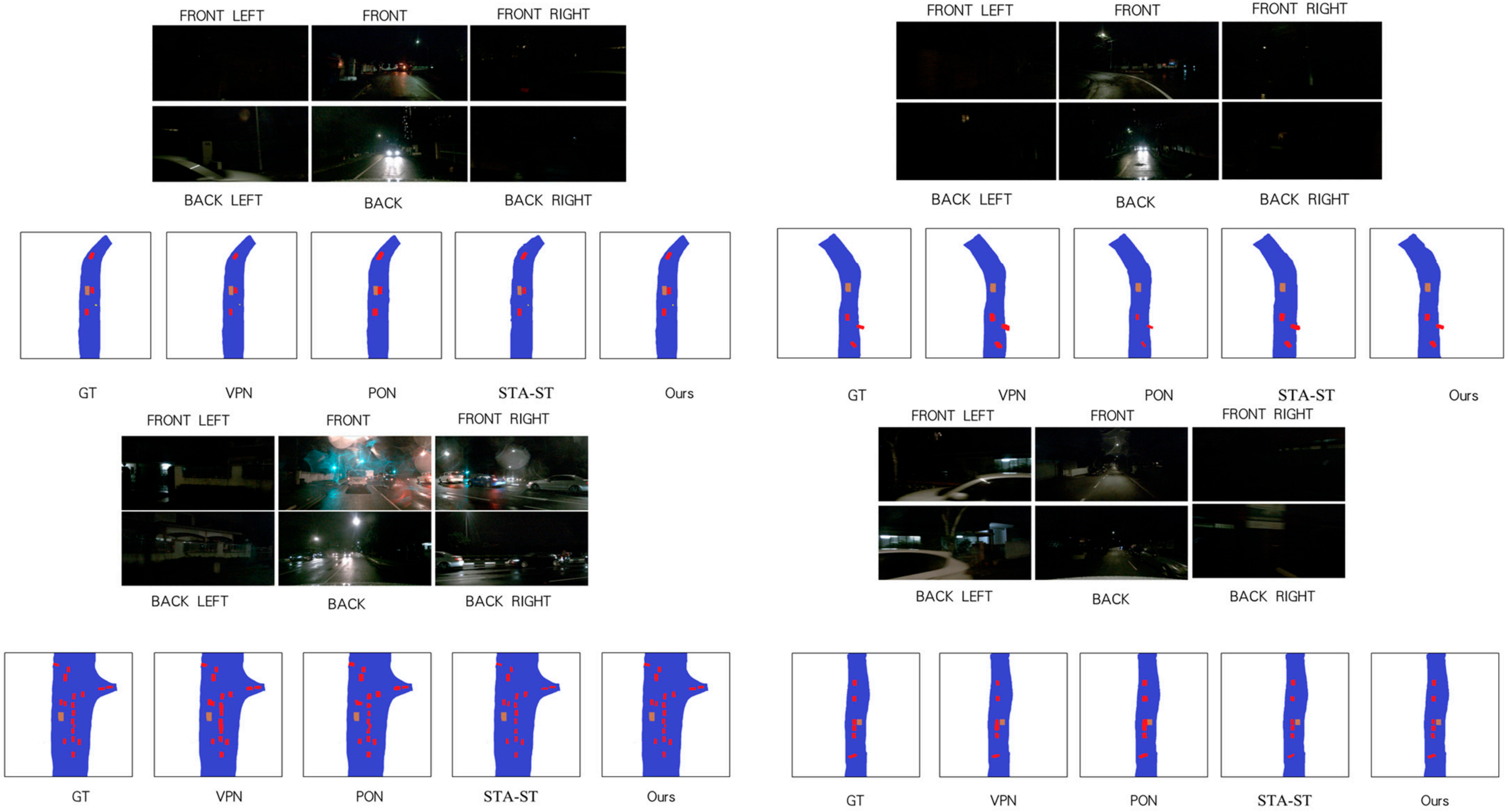

Figure 5. Own compilation qualitative results on the validation set of NuScenes at night.

4.4. Discussion

Our paper makes significant contributions in the following key areas:

1. We introduce a comprehensive framework featuring a View Parsing Network, which incorporates three essential modules: the generation of BEV graphs from images, the construction of constrained and data-efficient transformer networks based on physical principles, and the integration of formulas with monotonic attention inspired by the language domain. This framework excels in the accurate prediction of small objects at a distance and proves highly beneficial for tasks like vehicle and pedestrian detection, as well as traffic flow monitoring within intelligent transportation scenarios.
2. Leveraging the principles of physics, we develop an innovative transformer network that exhibits convolutional characteristics concerning the horizontal x-axis, all while maintaining exceptional spatial awareness. We seamlessly fuse our model with the monotonic attention mechanism, drawn from the domain of natural language processing. This combination facilitates more precise mapping by emphasizing information situated below a specific point in the image. The synergistic effect of these components yields optimal performance, substantially enhancing the capabilities of our model.
3. Our post-experimental results demonstrate our proficiency in generating BEV graphs from images represented as a set of 1D sequence-to-sequence transformations. We elucidate how axial attention significantly boosts performance by introducing temporal awareness and showcasing state-of-the-art results across an extensive dataset, the NuScenes dataset. Additionally, we apply domain adaptation techniques to enable the transfer of our model into real-world data without necessitating any supplementary annotations. Overall, our experiments underscore the superior performance of our model, particularly in the prediction of small objects.

4.5. Ablation Study

In this section, as shown in Table 2, we use ablation study to evaluate the effectiveness and robustness of our proposed BEV (Bird's Eye View) method for intelligent transportation applications. We compare our method with four other methods, namely: 1 (looking up): this method uses a simple perspective transformation to transform the image from front view to top view, and then uses a pre-trained CNN to perform semantic segmentation and object detection; 2 (looking both ways): this method uses a bi-directional perspective transformation to transform the image from front view and side view to top view, and then uses a pre-trained CNN to perform semantic segmentation and object detection; 3 (BEV-plane): this method uses a plane-based BEV generation method to transform the image from front view to top view, and then uses a pre-trained CNN to perform semantic segmentation and object detection; 4 (both-planes): this method uses a plane-and-curve-based BEV generation method to transform the image from front view and side view to top view, and then uses a pre-trained CNN to perform semantic segmentation and object detection.

Table 2. IOU (%) for ablation studies.

Model	Drivable	Car	Ped
1	69.2	31.7	6.5
2	73.3	34.6	7.9
3	74.8	35.9	8.4
4	76.9	38.4	10.5
Our	78.2	40	10.2

We use three metrics to evaluate the performance of different methods, namely the accuracy of drivable area (Drivable), car (Car), and pedestrian (Ped). From the table, we can see that our method achieves the highest accuracy for all three metrics, which are 78.2%, 40%, and 10.2%, respectively, while the other four methods have average accuracies of

73.5%, 35.2%, and 8.3%, respectively. These results show that our method can effectively utilize multimodal information, including image, radar, and laser, to generate high-resolution BEV images, and use deep neural networks to perform semantic segmentation and object detection. Our method can also adapt to different weather and lighting conditions, as well as different road types and structures. Our method has strong advantages and robustness for applying BEV in intelligent transportation systems.

4.6. Multiple-Object Prediction

Multiple-object prediction is the task of predicting the locations and categories of multiple objects in a scene, such as cars, pedestrians, bikes, etc. This task is important for intelligent transportation systems, as it can help to monitor and manage the traffic flow and safety. Based on the results shown in the table, we can see that our method outperforms the other four methods in this task. As shown in Table 3, our method achieves the highest mean accuracy of 25.8%, while the other four methods have mean accuracies of 17.4%, 19.1%, 23.7%, and 25.7%, respectively. Our method also achieves the highest or second-highest accuracy for most of the object categories, such as drivable, crossing, walkway, carpark, car, bus, etc. These results demonstrate that our method can effectively detect and classify multiple objects in various scenes and conditions, using a novel bird's eye view (BEV) representation.

Table 3. IOU (%) for multiple-object prediction.

Model	Drivable	Crossing	Walkway	Carpark	Car	Truck	Trailer	Bus	Con. Veh	Bike	Motorbike	Ped	Cone	Barrier	Mean
VPN	58	27.3	29.4	12.3	25.5	17.3	16.6	20	4.9	4.4	5.6	7.1	4.6	10.8	17.4
PON	60.4	28	31	18.4	24.7	16.3	16.6	20.8	12.3	9.4	7	8.2	5.7	8.1	19.1
STA-ST	70.7	31.1	32.4	33.5	36	22.8	13.6	29.2	12.1	12.1	8	8.6	6.9	14.2	23.7
TIIM-ST	74.5	36.6	35.9	31.3	39.7	26.3	13.9	32.8	14.2	14.7	7.6	9.5	7.6	14.7	25.7
Our	78.2	36.1	35.7	30.9	40	27	14.2	33.5	13.4	14.8	6.7	10.2	7.8	13.3	25.8

4.7. Dynamic and Static Timing Prediction

As shown in Figure 6, the PR graphs for static (Drivable) and dynamic elements (vehicle, Pedestrian, Large vehicle, Bicycle, Bus, Trailer, Motorcycle), which are obtained by adjusting different confidence thresholds, and different points on the coordinate system represent different recall and precision.

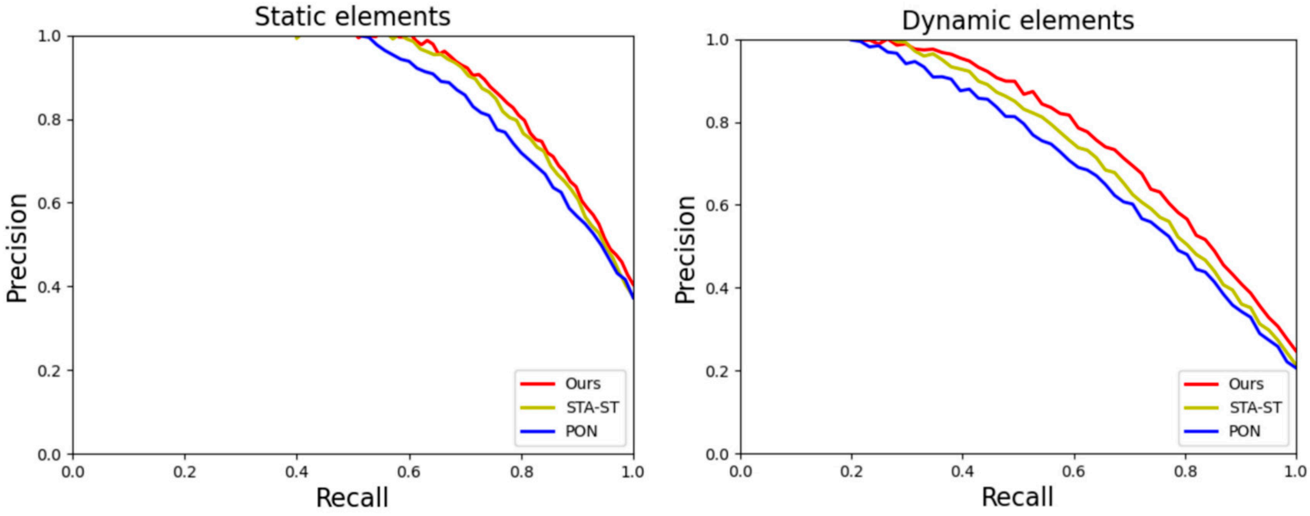

Figure 6. PR (Precision–Recall) plots for static elements (Drivable) and dynamic elements (vehicle, Pedestrian, Large vehicle, Bicycle, Bus, Trailer, Motorcycle).

The P–R curve is an important tool for evaluating the performance of a model, and by finding a balance between recall and precision, a more comprehensive understanding of the model's performance can be obtained. In the context of intelligent transportation, we are particularly interested in dynamic temporal prediction, the ability to accurately predict targets in motion in a traffic scenario. For a given recall, a higher detection accuracy means that the model is more accurate to the target while maintaining a high recall. In this experimental comparative analysis, we can clearly observe the performance of our method on the P–R curve. Our method is not only able to achieve a high recall, but also achieves superior detection accuracy at the corresponding recall. This suggests that our algorithm represents a better result for a given recall setting in terms of dynamic timing prediction.

As shown in the figure, this result is crucial for practical applications of ITS. Highly accurate dynamic timing prediction means that the motion trajectories of vehicles, pedestrians, bicycles, and other targets can be more accurately captured and predicted by the model, thus improving the overall safety and efficiency of transportation systems. Through comparative analysis, we are able to better understand the performance of different algorithms on the dynamic timing prediction task, which provides a strong support for the technological advancement in the field of intelligent transportation.

In BEV (Bird's Eye View)-based intelligent transportation systems, continuous time prediction of automotive elements in the scene is a key task. This prediction involves not only the position and speed of the target car, but also an accurate estimation of its shape, size and orientation. With the NuScenes dataset, we are able to validate and demonstrate our prediction models in a real traffic environment. As shown in Figure 7, we demonstrate a visual comparison of different models predicting car elements on 10 consecutive keyframes in the NuScenes dataset. This comparison clearly shows the effectiveness and limitations of different models in dealing with complex traffic scenarios. By integrating data from multiple viewpoints, our network is able to more accurately capture the dynamics of cars, including their trajectories and behavioral patterns.

In addition, our model makes it possible to take into account historical data and future trends when predicting the dynamics of automotive elements. In ITS applications, this approach not only improves prediction accuracy, but also generates smoother and more coherent motion trajectories, which are critical for understanding complex traffic situations and making accurate driving decisions. Accurate prediction not only improves the safety of self-driving vehicles, but also optimizes traffic flow, reduces congestion, and improves the efficiency of the entire transportation system. In conclusion, our approach provides strong technical support for efficient and safe traffic management in BEV-based intelligent transportation systems.

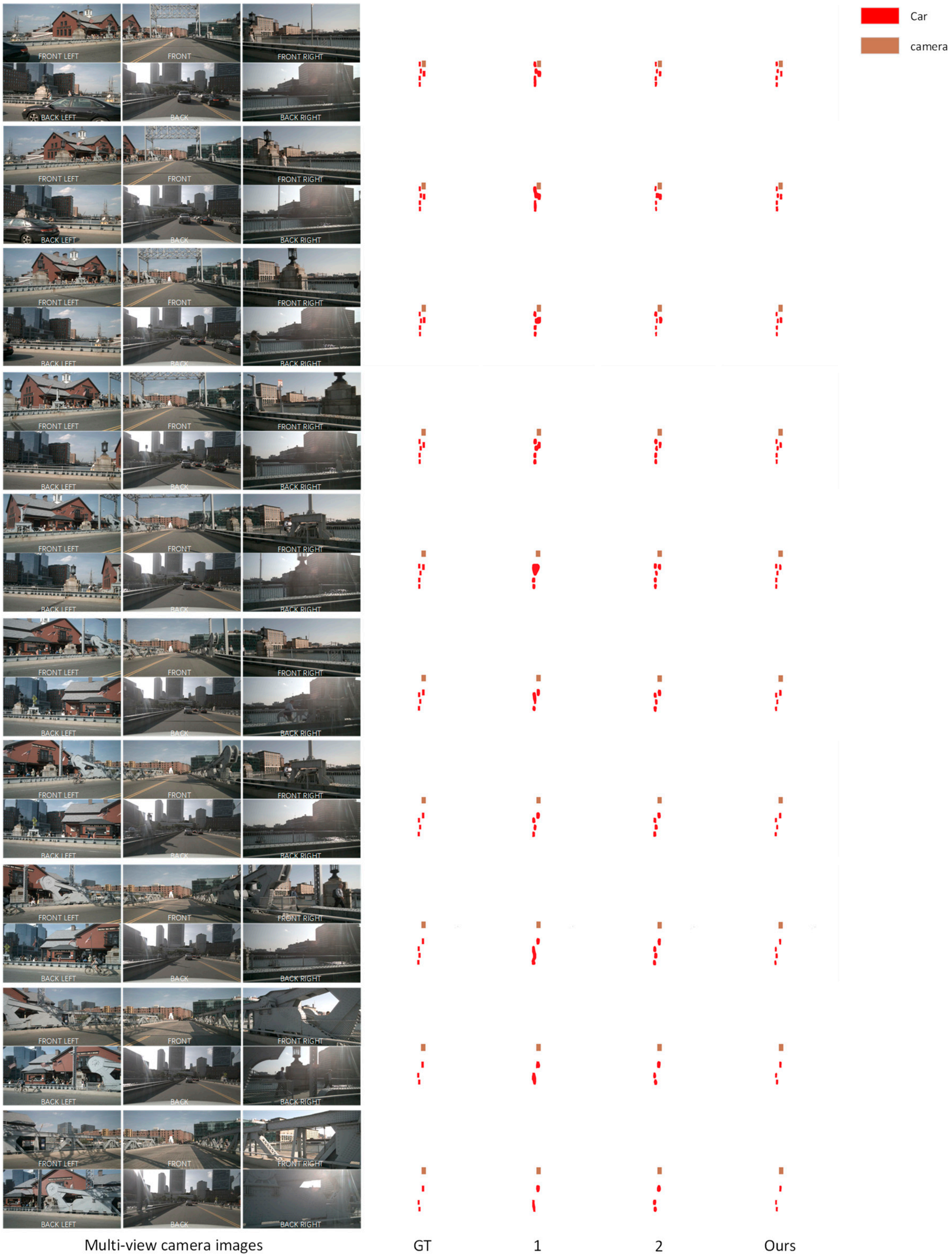

Figure 7. Visualization comparison of different models predicting automotive elements on 10 consecutive keyframes in the NuScenes dataset.

5. Conclusions

In our research, we propose an innovative Transformer network application designed to map images and video sequences into a bird's-eye map or top view of the environment, which is significant to intelligent transportation. Our approach adopts mathematical formulations of physical constraints, validates these formulations through resection studies, and combines recent advances in monotonic attention to confirm our intuition about which contextual information of points is more important in this map generation process. Experimentally, we demonstrate that our novel approach achieves state-of-the-art performance in instantly mapping a widely recognized dataset. This technology can be used in autonomous vehicles in intelligent transportation to help them better understand the road and environment around them.

Furthermore, we propose the concept of a cross-view semantic segmentation task to enhance environment awareness, which has substantial application prospects in the context of intelligent transportation. We introduce a neural architecture called View Parsing Network (VPN) to solve this task. Through experimental results, we demonstrate that VPN can be applied to mobile robots to enhance their perception of the surrounding environment by providing a lightweight and efficient top-down perspective semantic map. In many cases, object height information is unnecessary, so VPNs are lighter and more efficient than traditional 3D-based methods, which are more expensive in terms of data storage and computing requirements. This research provides more efficient and cost-effective solutions for intelligent transportation systems and helps improve environmental understanding and map generation, especially in autonomous vehicle technology.

To provide a basis for further work, we also plan to explore how to extend our approach to other scenes and transitions between viewpoints. Furthermore, we would like to investigate how to exploit multimodal information, such as sound or radar signals, to improve the quality and robustness of map generation and semantic segmentation. Finally, we would also like to explore how our approach can be combined with reinforcement learning or planning algorithms to realize more intelligent and adaptive mobile robot behavior in intelligent traffic scenarios.

The innovative prospects of applying Transformer networks in intelligent transportation systems offer great promise for shaping the future of autonomous vehicles and environmental perception. Integrating our proposed mapping and segmentation techniques into actual autonomous vehicle systems. This could lead to safer and more reliable transportation, as vehicles equipped with our technology will have a greater ability to interpret and adapt to complex environments. Furthermore, our approach can potentially be integrated into smart city infrastructure, helping to develop more efficient and adaptive traffic management systems. Future research can focus on extending the scope of our proposed cross-view semantic segmentation task to a wider range of environmental elements. This could include identifying dynamic objects such as pedestrians, cyclists, and other vehicles, thereby improving the overall situational awareness of autonomous systems. By solving real-world challenges and continually refining our approach based on feedback from field deployments, we ensure our technology meets the evolving needs of intelligent transportation systems. Overall, our research not only proposes cutting-edge technologies in the field of smart transportation, but also lays the foundation for a future in which advanced mapping and sensing systems play a key role in creating safer, more efficient, and greener transportation ecosystems.

Author Contributions: Conceptualization, Z.M. and M.H.; methodology, Z.M.; software, Y.Z.; validation, Y.Z. and Z.M.; formal analysis, Y.Z.; investigation, Y.Z.; resources, M.H.; data curation, Z.M.; writing—original draft preparation, Z.M.; writing—review and editing, M.H.; visualization, Y.Z.; supervision, M.H.; project administration, M.H.; funding acquisition, M.H. All authors have read and agreed to the published version of the manuscript.

Funding: National Natural Science Foundation of China under Grant (No. 62376246).

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qiu, H.; Liu, X.; Rallapalli, S.; Bency, A.J.; Chan, K.; Urgaonkar, R.; Manjunath, B.S.; Govindan, R.K. Kestrel: Video analytics for augmented multi-camera vehicle tracking. In Proceedings of the 2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI), Orlando, FL, USA, 17–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 48–59.
2. Xiong, X.; Liu, Y.; Yuan, T.; Wang, Y.; Wang, Y.; Zhao, H. Neural map prior for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 17535–17544.
3. Xu, Z.; Liu, Y.; Sun, Y.; Liu, M.; Wang, L. Road lane centerline graph detection with vehicle-mounted sensors by transformer for high-definition map creation. *arXiv* **2022**, arXiv:2209.07734.
4. Zhang, R.; Cao, S. Extending reliability of mmwave radar tracking and detection via fusion with camera. *IEEE Access* **2019**, *7*, 137065–137079. [CrossRef]
5. Ng, M.H.; Radia, K.; Chen, J.; Wang, D.; Gog, I.; Gonzalez, J.E. Bird's eye view semantic segmentation using geometry and semantic point cloud. *arXiv* **2020**, arXiv:2006.11436.
6. Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M.S. A review of semantic segmentation using deep neural networks. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 87–93. [CrossRef]
7. Chen, S.; Cheng, T.; Wang, X.; Meng, W.; Zhang, Q.; Liu, W. Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer. *arXiv* **2022**, arXiv:2206.04584.
8. Daily, M.J.; Harris, J.G.; Reiser, K. Detecting obstacles in range imagery. In Proceedings of the Image Understanding Workshop, Los Angeles, CA, USA, 23–25 February 1987; pp. 87–97.
9. Li, Y.; Ma, L.; Zhong, Z.; Liu, F.; Chapman, M.A.; Cao, D.; Li, J. Deep Learning for LiDAR Point Clouds in Autonomous Driving: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 3412–3432. [CrossRef] [PubMed]
10. Rahman, M.M.; Tan, Y.; Xue, J.; Lu, K. Recent advances in 3D object detection in the era of deep neural networks: A survey. *IEEE Trans. Image Process.* **2019**, *29*, 2947–2962. [CrossRef] [PubMed]
11. Chen, C.; Chen, C. *Mapping Scientific Frontiers*; Springer: Berlin/Heidelberg, Germany, 2003.
12. Mozaffari, S.; Al-Jarrah, O.Y.; Dianati, M.; Jennings, P.; Mouzakitis, A. Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 33–47. [CrossRef]
13. Homayounfar, N.; Ma, W.-C.; Liang, J.; Wu, X.; Fan, J.; Urtasun, R. Dagmapper: Learning to map by discovering lane topology. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2911–2920.
14. Fan, Y.; Feng, Z.; Mannan, A.; Khan, T.U.; Shen, C.; Saeed, S. Estimating tree position, diameter at breast height, and tree height in real-time using a mobile phone with RGB-D SLAM. *Remote Sens.* **2018**, *10*, 1845. [CrossRef]
15. Albuquerque, V.; Oliveira, A.; Barbosa, J.L.; Rodrigues, R.S.; Andrade, F.; Dias, M.S.; Ferreira, J.C. Smart Cities: Data-Driven Solutions to Understand Disruptive Problems in Transportation—The Lisbon Case Study. *Energies* **2021**, *14*, 3044. [CrossRef]
16. Wang, Z.; Huang, J.; Miao, K.; Lv, X.; Chen, Y.; Su, B.; Liu, L.; Han, M. Lightweight zero-knowledge authentication scheme for IoT embedded devices. *Comput. Netw.* **2023**, *236*, 110021. [CrossRef]
17. Neurohr, C.; Westhofen, L.; Butz, M.; Bollmann, M.H.; Eberle, U.; Galbas, R. Criticality analysis for the verification and validation of automated vehicles. *IEEE Access* **2021**, *9*, 18016–18041. [CrossRef]
18. Yu, J.; Gao, H.; Zhou, D.; Liu, J.; Gao, Q.; Ju, Z. Deep temporal model-based identity-aware hand detection for space human–robot interaction. *IEEE Trans. Cybern.* **2021**, *52*, 13738–13751. [CrossRef] [PubMed]
19. Yoo, J.H.; Kim, Y.; Kim, J.; Choi, J.W. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXVII. Springer: Berlin/Heidelberg, Germany, 2020; Volume 16, pp. 720–736.
20. Wang, H.; Cai, P.; Sun, Y.; Wang, L.; Liu, M. Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 13731–13737.
21. Liu, B.; Chen, W.; Wang, Z.; Pouriyeh, S.; Han, M. RAdam-DA-NLSTM: A Nested LSTM-Based Time Series Prediction Method for Human–Computer Intelligent Systems. *Electronics* **2023**, *12*, 3084. [CrossRef]
22. Solares-Canal, A.; Alonso, L.; Picos, J.; Armesto, J. Automatic tree detection and attribute characterization using portable terrestrial lidar. *Trees* **2023**, *37*, 963–979. [CrossRef]
23. García, A.; Valbuena, G.D.; García-Tuero, A.; Fernández-González, A.; Viesca, J.L.; Battez, A.H. Compatibility of Automatic Transmission Fluids with Structural Polymers Used in Electrified Transmissions. *Appl. Sci.* **2022**, *12*, 3608. [CrossRef]
24. Yu, J.; Gao, H.; Chen, Y.; Zhou, D.; Liu, J.; Ju, Z. Deep object detector with attentional spatiotemporal LSTM for space human–robot interaction. *IEEE Trans. Hum. Mach. Syst.* **2022**, *52*, 784–793. [CrossRef]
25. Sivaraman, S.; Trivedi, M.M. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 1773–1795. [CrossRef]
26. Thompson, S.; Sheat, D. Exploiting telecommunications to deliver real time transport information. In Proceedings of the 9th International Conference on Road Transport Information and Control, 1998, London, UK, 21–23 April 1998; pp. 59–61.
27. Yu, J.; Zheng, W.; Chen, Y.; Zhang, Y.; Huang, R. Surrounding-aware representation prediction in Birds-Eye-View using transformers. *Front. Neurosci.* **2023**, *17*, 1219363. [CrossRef]

28. Tang, T.; Li, J.; Huang, H.; Yang, X. A car-following model with real-time road conditions and numerical tests. *Measurement* **2014**, *48*, 63–76. [CrossRef]
29. Rosas, S.R.; Kane, M. Quality and rigor of the concept mapping methodology: A pooled study analysis. *Eval. Program Plan.* **2012**, *35*, 236–245. [CrossRef]
30. Dutta, P.; Sistu, G.; Yogamani, S.; Galván, E.; McDonald, J. A hierarchical transformer network for monocular birds-eye-view segmentation. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–7.
31. Ma, J.; Liu, Y.; Han, M.; Hu, C.; Ju, Z. Propagation Structure Fusion for Rumor Detection Based on Node-Level Contrastive Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *17*, 3319661. [CrossRef] [PubMed]
32. Liu, T.; Yang, J.; Li, B.; Xiao, C.; Sun, Y.; Wang, Y.; An, W. Nonconvex tensor low-rank approximation for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5614718. [CrossRef]
33. Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; Hauptmann, A.G. Infrared patch-image model for small target detection in a single image. *IEEE Trans. Image Process.* **2013**, *22*, 4996–5009. [CrossRef]
34. Sebastien, R.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203.
35. Wang, K.; Du, S.; Liu, C.; Cao, Z. Interior attention-aware network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5002013. [CrossRef]
36. Atcheson, B.; Ihrke, I.; Heidrich, W.; Tevs, A.; Bradley, D.; Magnor, M.; Seidel, H.P. Time-resolved 3d capture of non-stationary gas flows. *ACM Trans. Graph. (TOG)* **2008**, *27*, 1–9. [CrossRef]
37. Liu, S.; Chen, P.; Woźniak, M. Image enhancement-based detection with small infrared targets. *Remote Sens.* **2022**, *14*, 3232. [CrossRef]
38. Hu, C.; Liu, Z.; Li, R.; Hu, P.; Xiang, T.; Han, M. Smart Contract Assisted Privacy-Preserving Data Aggregation and Management Scheme for Smart Grid. *IEEE Trans. Dependable Secur. Comput.* 2023; *early access*. [CrossRef]
39. Du, J.; Lu, H.; Zhang, L.; Hu, M.; Chen, S.; Deng, Y.; Shen, X.; Zhang, Y. A spatial-temporal feature-based detection framework for infrared dim small target. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 3000412. [CrossRef]
40. Zheng, Y.; Pan, S.; Lee, V.; Zheng, Y.; Yu, P.S. Rethinking and scaling up graph contrastive learning: An extremely efficient approach with group discrimination. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10809–10820.
41. Yu, C.; Xiao, B.; Gao, C.; Yuan, L.; Zhang, L.; Sang, N.; Wang, J. Lite-hrnet: A lightweight high-resolution network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, Nashville, TN, USA, 20–25 June 2021; pp. 10440–10450.
42. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
43. Gosala, N.; Valada, A. Bird's-eye-view panoptic segmentation using monocular frontal view images. *IEEE Robot. Autom. Lett.* **2022**, *7*, 1968–1975. [CrossRef]
44. Dwivedi, I.; Malla, S.; Chen, Y.-T.; British, B.D. Bird's eye view segmentation using lifted 2D semantic features. In Proceedings of the Machine Vision Conference (BMVC), Online, 22–25 November 2021; pp. 6985–6994.
45. Ma, Y.; Wang, T.; Bai, X.; Yang, H.; Hou, Y.; Wang, Y.; Qiao, Y.; Yang, R.; Manocha, D.; Zhu, X. Vision-centric bev perception: A survey. *arXiv* **2022**, arXiv:2208.02797.
46. Saha, A.; Mendez, O.; Russell, C.; Bowden, R. "The Pedestrian next to the Lamppost" Adaptive Object Graphs for Better Instantaneous Mapping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 19528–19537.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.