S & M 3733

Image Signal Communication and Sensing for Traffic Key Representation Prediction

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

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China
[2]Innovation Center for Smart Medical Technologies & Devices, Binjiang Institute of Zhejiang University, Hangzhou 310053, China

(Received January 23, 2024; accepted March 12, 2024)

Keywords: autonomous vehicles, sensors, perception transmission technology, real-time traffic monitoring

Autonomous vehicles are a pivotal technology in today's intelligent transportation systems, improving traffic conditions by enhancing road safety, reducing congestion, conserving energy, decreasing emissions, and boosting travel efficiency and comfort. The key to achieving these objectives is to enable autonomous vehicles to perceive their surroundings in real time, with accuracy and stability, and to make appropriate decisions and controls on the basis of this perception. This perception relies on onboard sensors such as cameras, radars, lidars, and ultrasonic sensors. Radar sensors, one of the most frequently used perception devices in autonomous vehicles, are known for their strong anti-interference ability, effectiveness under occlusion, accurate range measurement, and wide coverage. However, radar sensors are limited by low resolution, high cost, large data volume, and difficulty in identifying object types and colors. Therefore, compensating for the inadequacies of radar sensing technology and enhancing the perceptual abilities of autonomous vehicles are ongoing challenges. In this paper, we introduce a novel image signal-based perception and transmission technology, aimed at overcoming the limitations of radar sensing and improving the perception efficiency and quality of autonomous vehicles. This technology employs image signals as the information medium, transmitting road information from the perception end to the processing end via wireless communication, thus enabling road condition perception. Compared with radar sensing technology, this image-based perception transmission technology is more cost-effective, as it only requires standard cameras and eliminates the need for expensive radar equipment. It captures a richer array of information, including object shape, size, position, direction, color, and texture, rather than just distance and velocity. The processing is more flexible, utilizing existing image processing and machine learning technologies for image signal compression, encoding, decoding, recognition, and analysis, thereby enhancing the accuracy and real-time capabilities of perception.

[*]Corresponding author: e-mail: mhan@zju.edu.cn
https://doi.org/10.18494/SAM4998

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

1. Introduction

In the context of intelligent transportation systems, the accurate prediction of traffic elements is crucial for enhancing road safety and traffic efficiency. In this study, we aim to explore the latest advancements in this field. Currently, the mainstream method for traffic element prediction is object detection, which plays a vital role in several aspects. Firstly, object detection technology can effectively identify and locate key elements such as vehicles and pedestrians in complex road environments, providing timely environmental information(1) to drivers. Secondly, this method is essential for implementing safety features such as collision warnings and automatic emergency braking systems. Lastly, object detection is critical in real-time traffic monitoring and management, helping to reduce traffic congestion and accidents. However, with the development of intelligent transportation systems, semantic segmentation has shown more forward-looking and practical applications than traditional object detection.(2) Unlike object detection, which merely draws bounding boxes around objects, semantic segmentation recognizes and categorizes each pixel in an image, offering more detailed and comprehensive environmental information. Firstly, semantic segmentation accurately delineates the shape of roads, sidewalks, and other key areas, providing essential information for path planning in autonomous vehicles. Secondly, it is more effective in dealing with occluded and overlapping targets, offering more accurate scene understanding. Lastly, semantic segmentation is particularly useful in complex urban environments, adapting better to varying road conditions and different weather situations.

Although semantic segmentation has broad prospects in intelligent transportation systems, achieving accurate semantic prediction is more challenging than achieving accurate object detection.(3) Firstly, semantic segmentation requires classifying each pixel in an image, which is computationally more complex and demands greater computational power. Secondly, given the highly dynamic nature of road environments, accurately predicting future semantic(4) information requires models to have higher environmental adaptability and time sensitivity. In the background of intelligent transportation systems, research based on converting images into bird's-eye view (BEV) maps,(5) particularly in conjunction with sensors that recognize road signs, is especially cutting-edge and important. Such research utilizes monocular or binocular cameras(6) to capture road scenes and transforms them into BEVs through projection transformations. This conversion provides richer spatial information for autonomous driving systems, including the position, direction, and speed of vehicles, as well as the geometric shape and topology of roads, which is crucial for understanding and responding to road signs. Current research on image-to-BEV map conversion mainly falls into two categories: geometry-based and deep-learning-based methods.(7) The geometry-based method requires accurate knowledge of camera intrinsic and extrinsic parameters and the equation of the road plane, using principles of perspective projection for transformation. This method is very effective in dealing with known environments and standard road signs. However, its limitations become apparent when faced with complex road conditions and diverse road signs.

On the other hand, the deep-learning-based method learns the transformation function directly from images through neural networks without explicitly using camera parameters or the road plane. This method exhibits higher flexibility and robustness in handling complex road scenes and various road signs. Although it requires a large amount of training data and a large number of computational resources, this method can effectively identify and parse various road signs, especially in nonstandardized road environments.

In conclusion, by combining traditional sensors with image-based BEV map conversion technologies, intelligent transportation systems can more comprehensively understand and react to various situations on the road, including rapidly and accurately identifying road signs. This not only enhances the safety and efficiency of autonomous vehicles but also brings revolutionary progress to the entire intelligent transportation system. In this paper, we propose a novel method for image generation: (1) The generation of BEV maps from images is presented as a set of one-dimensional sequence-to-sequence transformations. We introduce a novel task of cross-view semantic segmentation, which aids in flexibly perceiving the surrounding environment for robots and extracting target and category information from original images. (2) We propose a framework with a view parsing network that effectively learns and aggregates features from multi-angle and modal first-view observations, transforming them into a semantic map from the BEV perspective. (3) We introduce a transformation module that converts the output of the object detection branch into the input of the semantic prediction branch, facilitating information fusion and complementarity between the two branches. An adaptive loss function is utilized to adjust loss weights, thereby improving the accuracy and robustness of semantic prediction.

2. Related Work

2.1 Research on radar information transmission

Radar information transmission technology is significant in studying intelligent transportation systems and traffic element detection, particularly in integrating deep learning. In the realm of intelligent transportation system research, the application of radar technology has become a crucial component. Radar systems operate by emitting radio waves[8] and receiving the waves reflected, detecting the position and speed of objects. This technology is widely used in automotive, aviation, aerospace, and maritime fields. Radar systems capture information such as the distance, speed, angle, and size of objects and transmit this data via radio waves, which are then captured and interpreted by radar receivers. In traffic element detection, radar technology plays a vital role. It is used to detect and track surrounding vehicles' positions and speeds in autonomous and assisted driving systems.[9] It also identifies pedestrians and other obstacles, thereby enhancing road safety. Particularly in complex road environments, radar provides key environmental perception capabilities,[10] especially under limited visibility conditions.

In recent years, combining radar data with other sensor data[11] (such as cameras and LiDAR) and using deep learning techniques for data fusion have become hot research directions. More accurate detection and comprehensive environmental perception are achieved by using deep learning algorithms (such as convolutional neural networks) to extract features from radar signals and perform object classification and recognition. Furthermore, deep learning models can process radar data in real time, providing immediate decision support for autonomous vehicles.(12)

Despite significant progress in traffic element detection with radar technology, challenges remain, such as handling radar signal noise, optimizing multisensor data fusion,(13) and maintaining stability and accuracy under different weather and lighting conditions. Future research may focus on improving radar system resolution, optimizing deep learning models for more effective radar data processing, and developing advanced data fusion and processing techniques to further enhance the performance and reliability of intelligent transportation systems. In the research fields of intelligent transportation systems and autonomous vehicles, radar information transmission technology is increasingly attracting attention. This technology, utilizing radar signals for wireless communication, not only enhances the safety and efficiency of autonomous vehicles but also conserves valuable spectrum resources. Radar information transmission is closely related to the deep learning detection of traffic elements. As essential perception tools in autonomous driving, they complement and integrate, enhancing perception accuracy and robustness.

Radar information transmission technology utilizes the multidimensional features of radar waveforms, such as frequency, phase, polarization, and coding, to convey information, achieving dual functions of radar and communication. This transmission can occur between autonomous vehicles or vehicles and infrastructure, exchanging information such as position, speed, direction, and intention, thereby enhancing cooperative driving capabilities.(14) Additionally, radar information transmission can be an auxiliary perception method, fused with data from sensors such as visual cameras and LiDAR, to improve detection and recognition capabilities for dynamic and static objects. The deep learning detection of traffic elements refers to using deep neural networks to detect and recognize traffic elements, such as vehicles, pedestrians, lane lines, and traffic signs, from the data collected by onboard cameras or LiDAR sensors. This core perception task in autonomous driving directly impacts decision-making and planning. To tackle issues such as occlusion, variations in lighting, small objects, and scenarios involving multiple scales and classes, appropriate network structures, loss functions, and optimization methods(15) must be designed to enhance detection accuracy and speed. The relationship between radar information transmission and the deep learning detection of traffic elements is primarily manifested in the following aspects: (1) Radar information transmission provides additional information for deep learning detection, such as target distance, speed, and direction, improving detection accuracy and robustness. (2) Radar information transmission and deep learning detection can be jointly optimized, for instance, using deep learning to design radar waveforms, or transmitting parameters or output results of deep learning models via radar information transmission, thereby enhancing system performance and efficiency. (3) Radar information transmission and deep learning detection can achieve data fusion, such as using multimodal data to train deep neural networks or fusing features at different levels, thus improving the system's generalization ability and adaptability. These research directions showcase the cutting-edge nature of radar information transmission technology and indicate potential pathways for enhancing the perception capabilities of autonomous driving systems.

2.2 Research on image signal transmission

In the context of intelligent transportation systems and road information prediction research, the integrated application of data image signal transmission(16) and perception is emerging as a significant focus(17) in image processing. Combining these technologies enhances the efficiency and quality of image data transmission and optimizes the use of spectrum resources. In intelligent transportation systems, this integrated application is particularly crucial for improving road monitoring images' perceptual capability and application value.

Specifically, data image signal transmission technology utilizes the multidimensional features of image signals, such as grayscale, color, texture, and shape, for information encoding and decoding. This technology can facilitate information exchange between image-capturing devices or with infrastructure, effectively improving image transmission efficiency and quality. In intelligent transportation systems, images depicting road conditions, such as traffic flow and accidents, can be transmitted and processed with greater speed and precision, leading to more timely and precise responses. Simultaneously, data perception technology, which involves collecting image data through sensors or cameras, is the front-end part of image processing. It includes steps such as image acquisition, preprocessing, and enhancement and works with data image signal transmission technology. In intelligent transportation systems, data perception provides high-quality image data, such as clear, stable, and complete road scene images, and enhances the effectiveness(18) and performance of data image signal transmission.(19) Moreover, data image signal transmission technology provides effective image encoding and decoding methods for data perception, such as image compression, encryption, and verification, enhancing the security and reliability of image data processing. In intelligent transportation systems, this bidirectional optimization not only improves the accuracy of road information prediction but also enhances the system's overall performance.

In summary, combining data image signal transmission and data perception provides a powerful tool for intelligent transportation systems, enabling more effective processing and analysis of road information, thereby bringing significant benefits to autonomous driving and traffic management.(20) This integrated application's development reflects technological innovation and showcases the future direction of intelligent transportation systems.

2.3 Semantic prediction of traffic elements

The semantic prediction of traffic elements uses deep learning models to segment and recognize road elements such as roads, vehicles, pedestrians, and traffic signs from the data collected by onboard cameras or LiDAR sensors. This is a core perceptual task in autonomous driving and directly affects the effectiveness of decision-making and planning. Addressing the challenges of complex scenes in traffic element semantic prediction,(21) such as occlusion, changes in illumination, small objects, and multiscale and multiclass scenarios, necessitates the design of appropriate network structures, loss functions, and optimization methods to enhance the accuracy and speed of segmentation.

Bird's eye view (BEV) is a perspective that projects objects from three-dimensional space onto a two-dimensional plane, providing a more intuitive and comprehensive view of the scene, and is beneficial for the semantic prediction of traffic elements. Research on BEV primarily includes two aspects: the generation and segmentation of BEV. BEV generation involves using sensor data, such as images, point clouds, and radar, to construct a scene representation from the BEV perspective. There are generally two methods: geometry-based(22) and learning-based.(23) Geometry-based methods transform the coordinates of objects in three-dimensional space into planar coordinates in the BEV perspective based on the intrinsic and extrinsic parameters of the sensors. This approach is simple and fast but requires precise sensor calibration and is susceptible to noise and occlusion. Learning-based methods use deep neural networks to learn scene representations from the BEV perspective from sensor data. This approach can overcome sensor calibration errors and allows for multimodal data fusion. However, it requires extensive annotated data for model training and has a high computational demand. BEV segmentation refers to using deep learning models to segment and recognize the types and locations of traffic elements from the scene representation in the BEV perspective. There are typically two methods: detection-based and segmentation-based. Detection-based methods first use object detection models, such as YOLO and SSD, to detect bounding boxes of traffic elements in the BEV scene, followed by classification models such as ResNet and VGG to classify objects within each box. This method can accurately locate traffic elements but does not reflect the shape and details of objects nor distinguish overlapping and occlusion between objects. Segmentation-based methods use semantic segmentation models, such as FCN, U-Net, and PSPNet, to generate pixel-level semantic maps from the BEV scene directly. This method preserves the shape and details of objects and distinguishes between overlapping and occlusion but requires higher resolution and computational resources.

2.4 Attention and transformer-driven image translation

Attention and Transformer are two neural network models based on the self-attention mechanism, which have significantly advanced Natural Language Processing (NLP) and computer vision (CV) in recent years. Attention is a technique for enhancing the representational capability of neural networks and capturing long-range dependences. It enables the model to automatically focus on essential parts of the input sequence while ignoring irrelevant parts. Transformer, a model entirely based on attention, functions as an encoder-decoder and effectively processes sequential data, such as text and images, without the need for recurrent neural networks (RNNs) or convolutional neural networks (CNNs). Originally developed to address machine translation issues in NLP, Attention and Transformer were first proposed in the 2017 paper "Attention Is All You Need" and have since achieved remarkable results in various NLP tasks,(24) including text classification, text generation, question answering, and sentiment analysis. The advantages of Attention and Transformer are primarily reflected in their ability to utilize the global information of input sequences without being limited by local information, thereby enhancing the model's semantic understanding and generalization capabilities; they can process input sequences in parallel rather than sequentially, which increases computational efficiency and speed; and they can flexibly combine different attention mechanisms, such as self-attention, cross-attention, and multihead attention, to enhance the model's expressive power and diversity. Inspired by the success of Attention and Transformer in NLP,(25) researchers began exploring their application in the CV field to address tasks such as image classification, object detection, and image segmentation. Given the significant differences between image and text data, such as dimensions, structure, and distribution, directly transferring Attention and Transformer to the CV field requires modifications and innovations. In the CV domain, the development of Attention and Transformer has mainly manifested in the following aspects:

1. To adapt to image data's high dimensionality and resolution, some methods propose dividing images into multiple small patches, each treated as a token to be input into Transformer, thereby reducing computational complexity and memory consumption.
2. To cater to image data's local structure and global semantics, some methods have introduced hybrid models combining CNNs and Transformer. These models use CNNs to extract low-level features of images and Transformer to capture high-level features, enhancing the model's representational ability and robustness.
3. To address the diversity and multimodality of image data, some methods utilize different attention mechanisms to process various data types, such as images, text, and videos, thereby improving the model's generalization ability and multitasking capability.

Attention and Transformer are two self-attention-based neural network models that have significantly propelled the development of NLP and CV. Initially introduced to solve machine translation problems in NLP, they have demonstrated significant effectiveness in various NLP tasks. In CV, the evolution of Attention and Transformer is mainly evident in their ability to divide images into multiple small patches, integrate CNNs and Transformer in hybrid models, and use different attention mechanisms to handle various data types.

3. Methods

3.1 Encoder and decoder

In a three-dimensional environment, we utilize multiple sensors to gather first-view observations from different angles and modalities, obtaining comprehensive information. These first-view observations are encoded by different encoders corresponding to various modalities. These CNN-based encoders extract multiple spatial feature maps for the first-view inputs. Then, all these feature maps are input into a view transformation module, which transforms these view feature maps from the first-view space to the BEV feature space and fuses them to produce a final feature map containing sufficient spatial information. Finally, we use a convolutional decoder to predict the semantic map of the BEV. The overall framework and flow chart are shown in Fig. 1.

Despite the huge success of the encoder-decoder structure in classic semantic segmentation, the experiments show that it performs poorly in cross-view semantic segmentation tasks. We speculate that this is because, in standard semantic segmentation structures, the receptive field of the output spatial feature maps roughly aligns with the input spatial feature maps. However, in

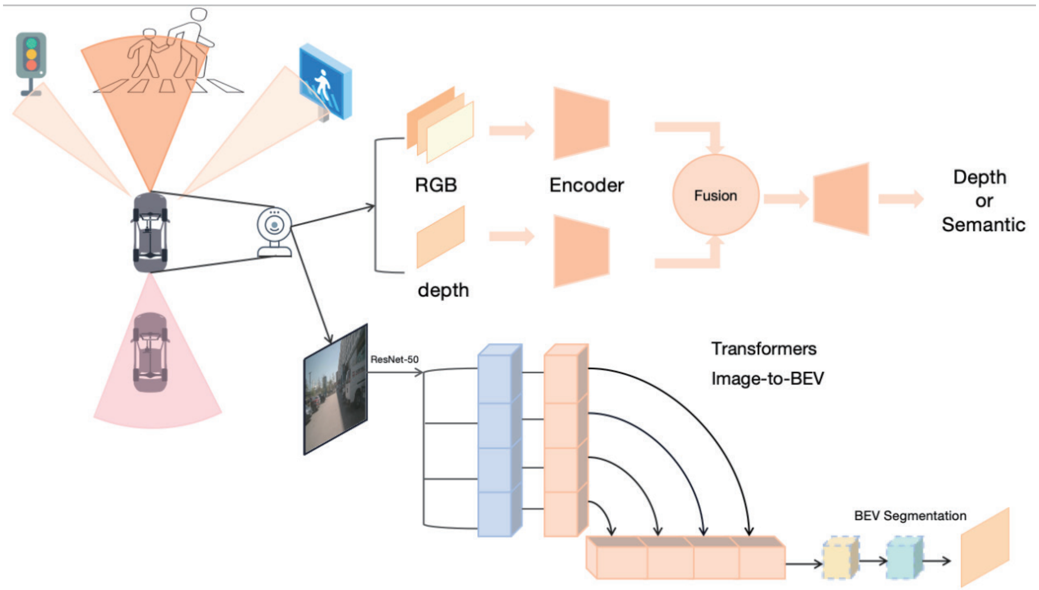

Fig. 1. (Color online) The compiler framework of cross-view semantic segmentation is applied to the prediction of road information in real-life intelligent traffic scenarios.

cross-view semantic segmentation, each pixel on the BEV should consider all the first-view feature maps of the inputs, not just the local receptive field areas. After considering the limitations of the current semantic segmentation structures, we designed a view transformation module to learn the dependences between all spatial positions of the first-view and BEV feature maps. This module does not alter the shape of the input feature maps, allowing it to be integrated into any existing encoder-decoder-type network architecture for classic semantic segmentation. It consists of a view relation module and a view fusion module.

The lower half of Fig. 1 illustrates the entire process. First-view feature maps are initially flattened while keeping the channel dimension unchanged. We then employ a view relation module P to learn the relationship between any two pixel positions of the flattened first-view and BEV feature maps, where $i, j \in [0, HW]$ is the index of the first view feature map $f \in R^{HW \times C}$ along the flattening dimension of the top view feature map, whereas P^i simulates the relationship between the i th pixel on the top view feature map and each pixel on the first view feature map. In the view relation module P, we simply use a multilayer perceptron (MLP). Subsequently, the BEV feature map is reshaped back into $H \times W \times C$. Each first-view input has its own view relation module, producing a BEV feature map $t_i \in R^{H \times W \times C}$ based on its observation. To aggregate information from all observation inputs, we use a view fusion module to merge these BEV feature maps t_i.

In terms of predicting the trajectories of target vehicles (TVs), we represent their future behavior as states they are expected to pass through, defined as

$$X_{TVs} = \left\{ X_t^i, X_{t+1}^i, \ldots, X_{t+m}^i \right\}_{i=1}^{N}, \tag{1}$$

where A denotes the state of vehicle i at time t (for example, position), N is the number of target vehicles, and m is the length of the prediction window. The goal is to compute the conditional distribution $P(X_{TVs}|Y_{EV})$, where Y_{EV} is the observation available for environmental vehicles (EVs). This distribution is a joint distribution of sequences of states for multiple interdependent vehicles, which can be complex to handle. To reduce the computational demand of estimating $P(X_{TVs}|Y_{EV})$, we assume that the future behaviors of the vehicles are independent. Therefore, the behavior of each target vehicle can be predicted separately, with a manageable computational requirement. At each step, select a $P(X_{TVs}|Y_{EV})$ vehicle as the target vehicle and calculate its $P(X_{TVs}|Y_{EV})$, where in Eq. (2), T is the selected target vehicle.

$$X_{TV} = \left\{X_t^T, X_{t+1}^T, \ldots, X_{t+m}^T\right\} \tag{2}$$

Sensors must be considered in the context of intelligent traffic and the detection and recognition of road traffic signals. Sensors are an essential component of intelligent traffic systems; they perceive, collect, transmit, and process traffic information, providing data support for traffic management and control. Depending on different detection principles and application scenarios, sensors can be categorized into various types, such as inductive loop detectors, image sensors, microwave radar sensors, and geomagnetic sensors. In the method, we primarily use image sensors to obtain observations of environmental vehicles because image sensors offer the following advantages:

1. Image sensors can provide rich visual information, including the position, shape, color, and type of vehicle, which helps improve the accuracy and robustness of the conditional distribution.
2. Image sensors can cover an extensive detection range, observing multiple lanes and directions of vehicles simultaneously, aiding in handling the interdependences among vehicles.
3. Image sensors can provide observations with different perspectives and modalities based on the behavior and maneuvers of vehicles, enhancing the multimodality of vehicle behaviors.

3.2 Image feature translation

The method utilizes cross-view semantic segmentation to generate semantic maps of polar coordinate BEVs from multiple camera perspectives. The model implicitly learns the mapping from a single camera view to a polar coordinate BEV, employing a camera calibration-based cross-view attention mechanism. Each camera uses position embeddings dependent on its intrinsic and extrinsic parameters. These embeddings allow a transformer to learn mappings between different views without explicit geometric modeling. The model comprises a convolutional image encoder and a cross-view transformer layer to infer the polar coordinate BEF semantic segmentation. The model is simple, easily parallelizable, and capable of real-time operation. It achieves state-of-the-art performance on the NuScenes dataset.

In the task of cross-view semantic segmentation, we consider the alignment between the semantic segmentation column of each camera view and its corresponding polar coordinate BEV ground truth as a "hard" alignment, where each pixel on the polar coordinate ray corresponds to a semantic category in the camera view column. To achieve such a hard alignment, the only uncertainty to resolve is the depth of each pixel. However, while aligning in this manner, we need to assign features that contribute to solving both semantic and depth aspects. Therefore, a hard alignment is disadvantageous. Instead, we prefer a soft alignment, where each pixel on the polar coordinate ray is assigned a combination of elements from the camera view column, i.e., a context vector. Specifically, when generating each radial element $S_i^{\varphi(BEV)}$, we want to give it a context c_i based on the convex combination of elements in the camera view column S^I and the radial position of element $S_i^{\varphi(BEV)}$ along the polar ray. This need for context allocation inspired us to use soft attention between the camera view column and its polar rays. Formally, let $h \in R^{H\times C}$ represent the encoded "memory" of a camera view column of height H, and let $y \in R^{r\times C}$ represent a position query that encodes the relative position of a polar ray along a length r. We generate a context c based on the input sequence h and query y by the alignment a between the elements in the input sequence and their radial positions. First, in Eq. (3), the input sequence h and the position query y are projected by the matrices $W_M \in R^{C\times D}$ and $W_N \in R^{C\times D}$ to the corresponding representations Q and K:

$$\begin{aligned} M\left(y_i\right) &= y_i W_M, \\ N\left(h_i\right) &= h_i W_N. \end{aligned} \tag{3}$$

We refer to Q and K as "query" and "key", respectively. After projection, we use the scaled dot product to produce an unnormalized alignment score between each memory-query combination in Eq. (4):

$$e_{i,j} = \frac{\left\{M\left(y_i\right), N\left(h_j\right)\right\}}{\sqrt{D}}. \tag{4}$$

Then, in Eq. (5), we use a softmax to normalize the energy scalar to a probability distribution on memory:

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

Finally, the context vector in Eq. (6) is computed as a weighted sum of K:

$$c_i = \sum_{j=1}^{H} \partial_i N\left(h_j\right). \tag{6}$$

The context is generated in such r_i way that each radial slot r_i independently collects relevant information from the camera view column and represents the initial allocation of components from the image to their aerial view position. Such an initial allocation is similar to raising a pixel based on its depth. However, it is elevated to a depth of distribution and thus should be able to overcome the common problems of sparsity and elongated object cross sections. This means that the image context available for each radial slot is independent of its distance to the camera. Finally, to generate an aerial view feature $S_i^{\varphi(BEV)}$ at radial position r_i, in Eq. (7), we operate globally on the context $c = \{c_i, ..., c_r\}$ assigned to all radial positions:

$$S_i^{\varphi(BEV)} = g\left(c\right)X \, , \tag{7}$$

where $g(.)$ is a nonlinear function that deduces along the entire polar ray. We describe its role below.

In this approach, the method addresses the everyday challenges of mapping camera views to their corresponding BEV positions. Using a transformer with soft attention mechanisms facilitates learning spatial relationships between the camera and the BEV. This model architecture enables efficient and accurate semantic segmentation, necessary for applications such as autonomous navigation and advanced driver-assistance systems. The global operation at the end consolidates the context information from all radial positions, ensuring a coherent and comprehensive representation of the BEV.

3.3 Loss design

To provide a training signal for the predicted occupancy grid, while addressing the uncertainties in both semantics and location, we employ a multiscale loss function similar to Eq. (8). For each scale u and category k, the average loss is as described in Eq. (8):

$$L_u = 1 - \frac{1}{|k|} \sum_{k=1}^{K} \frac{2\sum_i^N \widehat{y_i^k}\, y_i^k}{\sum_i^K \widehat{y_i^k} + y_i^k + \in} \, , \tag{8}$$

where y_i^k is the ground truth binary variable grid cell and $\widehat{y_i^k}$ is the network's predicted sigmoid output, a constant used to prevent division by zero.

The loss function is used for image segmentation tasks. Its fundamental idea is to compute the overlap between the predicted and actual results, optimizing the model by minimizing the difference between them. Compared with the cross-entropy loss function, the Dice loss can better handle class imbalance issues, as it considers the weight of each pixel in the calculation rather than simply using the number of pixels as the weight. In the context of intelligent traffic and the detection and recognition of road traffic signals, using the Dice loss function can improve the performance of polar coordinate BEV semantic segmentation, as it can overcome common problems of sparsity and elongated object cross sections. We utilize sensors to acquire depth information in the scene and the direction and speed of vehicles, thereby achieving the mapping from a single camera view to a polar coordinate BEV. The sensors include stereo vision systems, microwave radars, and geomagnetic sensors. These sensors assist us in detecting and tracking vehicles and collecting information on traffic flow, occupancy, and vehicle types.

In summary, the approach leverages the multiscale loss function and a combination of various sensors to enhance the semantic segmentation performance from camera views to polar coordinate BEVs. This method improves the accuracy of vehicle detection and tracking and facilitates the collection of comprehensive traffic data, which is essential for intelligent traffic management and advanced driver-assistance systems. The use of the Dice loss function, in particular, addresses the challenges posed by sparse and elongated objects common in overhead views of traffic scenes.

4. Experiments and Discussion

4.1 Experimental settings and datasets

In this study, two primary public datasets, Cityscapes and NuScenes, were utilized. Cityscapes is designed for urban street scenes, providing high-quality, pixel-level, and instance-level semantic annotations from 50 European cities. It encompasses various road and urban environments, including static road information (such as traffic signs, signals, and pedestrian paths), roads, buildings, vehicles, and pedestrians. This dataset aims to evaluate and promote research and development in algorithms and models related to semantic image segmentation, object detection, and scene understanding in urban contexts, particularly in autonomous driving. Cityscapes also offers extensions such as a 3D vehicle detection benchmark and a 3D bounding box extension. It is an ideal dataset for research and development in urban scene understanding, providing a wealth of diverse scenes for areas such as autonomous driving, thereby aiding in improving visual perception capabilities within such systems. NuScenes comprises street scenes from Boston and Singapore, covering various traffic participants, complex traffic conditions, and diverse road conditions. Presented in a time-series format, this dataset supports understanding and predicting dynamic scenes. It also offers extensive annotated information, including labels for 23 object categories such as vehicles, pedestrians, and bicycles. Additional features such as map layers and raw sensor data are also provided. In developing intelligent traffic systems, the precise perception of vehicles, pedestrians, and roads is crucial for enhancing the performance and safety of autonomous driving systems. The core of this research is using six annular RGB cameras mounted on the vehicle, distributed around the vehicle to provide a 360-degree panoramic view. These cameras capture the road environment in real time, generating a complete annular view that aids in the comprehensive perception of the vehicle's surroundings.

To analyze and utilize these image data effectively, metrics such as the "Intersection over Union" (IOU) are commonly employed to assess the performance of BEV map generation algorithms. IOU is an indicator measuring the overlap between maps generated by algorithms and the real-world environment. It is particularly suited for measuring the correspondence of objects (such as vehicles and pedestrians) in BEV maps to their actual counterparts. The higher the IOU value, the closer the algorithm-generated map is to the actual environment, indicating higher accuracy. Therefore, in this study, we aim to develop a high-quality BEV map generation algorithm that can utilize the annular RGB cameras on vehicles to capture road conditions in real time and employ evaluation metrics such as IOU to verify the accuracy and performance of the generated BEV maps. This approach can significantly enhance autonomous driving systems' environmental perception and path-planning capabilities, playing a crucial role in creating a safer and more efficient intelligent traffic system.

Aside from IOU, we also employ *Precision–Recall* (*P–R*) curves to assess the model performance further. *P–R* curves depict the overall performance trend of a model by calculating precision and recall at different thresholds. The area under the curve, known as the AP value, represents the accuracy of model recognition. Overall, improvements in mAP (mean Average Precision across all categories) reflect an overall enhancement in model performance in object detection. In this study, we aim to provide more accurate and comprehensive technical support for road information prediction in intelligent traffic through these comprehensive evaluation methods.

4.2 Comparison results and discussion

In this study, we evaluated the performance of various semantic segmentation models in urban scenes, focusing on three key metrics: drivable area accuracy (drivable), vehicle detection accuracy (road markings), and pedestrian detection accuracy (ped), as illustrated in Table 1.

To evaluate the performance of the approach, we conducted experiments in various urban scenarios, including highways, urban areas, and suburbs. We used metrics such as accuracy, recall, and F1 score to measure the predictive performance of the method on different categories of targets. The method was compared with several benchmark methods, including VED1, PON, STA-ST, and TIIM-ST. The method outperforms all other methods on all metrics, indicating that the method has high accuracy and robustness.

Table 1
The sensor achieves an accurate perception of urban vehicles, pedestrians, and roads from front and side perspectives. The road under intelligent transportation relies on the vehicle's six ring RGB cameras for perception. The main RGB view of the road is obtained in real time, and the algorithm of BEV map generation is studied and evaluated as Index IOU.

Model	Drivable	Road markings	Ped
VED	58.7	7.4	7.7
PON	60	8.5	8.4
STA-ST	70.7	8.3	8.6
TIIM-ST	75.1	9.4	10
Our	81.2	13.2	13.3

Model	Drivable	Road markings	Ped
VED	58.2	7.3	7.3
PON	60.4	8.4	8.8
STA-ST	70.1	8.6	9
TIIM-ST	75.9	9.2	10.4
Our	81	13.9	13.7

By comparing four existing models, VED, PON, STA-ST, and TIIM-ST, and the proposed new model, we found that the model outperformed the existing model on all metrics. Specifically, as shown in Table 2, the model achieves a positive accuracy of 33% and a lateral accuracy of 32.7% on roadblock information, 18.1 and 18.2% higher than those of the existing best model IM-ST, respectively. In automobile detection, the front reaches 80.6% and the side reaches 80.9%, which is 5.9% higher than the front of TIIM-ST. Positive results on traffic signal signs were 13.2 and 13.9, 3.5 and 4.7% higher than those of TIIM-ST, respectively. These results reveal that the model shows remarkable robustness and generalization ability when dealing with complex and diverse urban scenes, and the corresponding ROI curve is shown in Fig. 2.

The model employs deep learning techniques and utilizes a large amount of data to train the model, enabling it to automatically learn the features of road information without the need for manually designed feature extractors. It uses a multiscale CNN to capture road information at different scales, enhancing the model's generalization ability and detailed representation. The model employs an attention mechanism that can automatically focus on essential areas and ignore irrelevant areas, thereby improving the model's efficiency and accuracy.

In summary, the model can effectively predict the road conditions in front of and on the side of intelligent vehicles in the context of intelligent transportation and road information prediction. It outperforms existing models on eight indicators, highlighting its high accuracy and robustness. These advantages are mainly attributed to the model's combination of advanced methods such as deep learning, multiscale CNNs, attention mechanisms, and multitask learning.

4.3 Qualitative results and discussion

In this paper, a deep-learning-based method that utilizes sensors in the front and side of autonomous vehicles is proposed. The multi-object prediction shown in Fig. 3 is a task to predict the location and category of multiple objects in the scene. It performs real-time predictions of

Table 2
IOU (%) of sensor for multitarget prediction (the table above is for the sensor side shot data, and the table below is for the sensor front shot data).

Model	Drivable	Crossing	Walkway	Carpark	Barricade	Traffic signal	Con.Veh	Mean
VED	58.6	26.9	29.8	12.7	10.6	7.4	4.8	21.5
PON	59.8	28.3	31.2	18.3	8.4	8.6	12	23.8
STA-ST	71	30.6	32.7	33	14	8.3	12.3	28.8
TIIM-ST	74.7	37.1	35.4	31.5	14.5	9.7	14.3	31.1
Our	80.6	41.4	41.2	41.8	32.7	13.2	25.9	39.5

Model	Drivable	Crossing	Walkway	Carpark	Barricade	Traffic signal	Con.Veh	Mean
VED	58.5	26.7	30	12.4	11	7.3	5.3	21.6
PON	59.4	28.2	31.3	18.5	8.8	8.4	11.6	23.7
STA-ST	71.2	30.8	33.2	33.5	13.7	8.6	12.8	29.1
TIIM-ST	75	36.9	35.8	31.8	14.9	9.2	14	31.1
Our	80.9	41.7	41.5	42	33	13.9	23.6	39.5

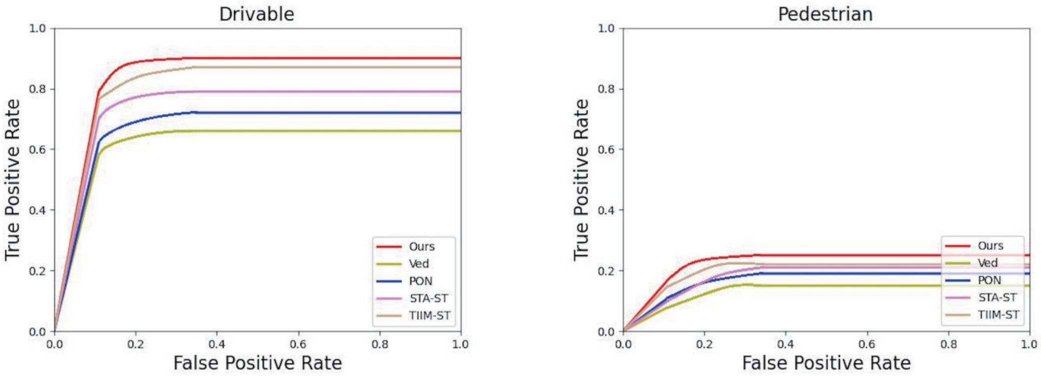

Fig. 2. (Color online) BEV map generation algorithm and index IOU (the image on the left shows the sensor front detection data, and the image on the right shows the sensor side detection data.).

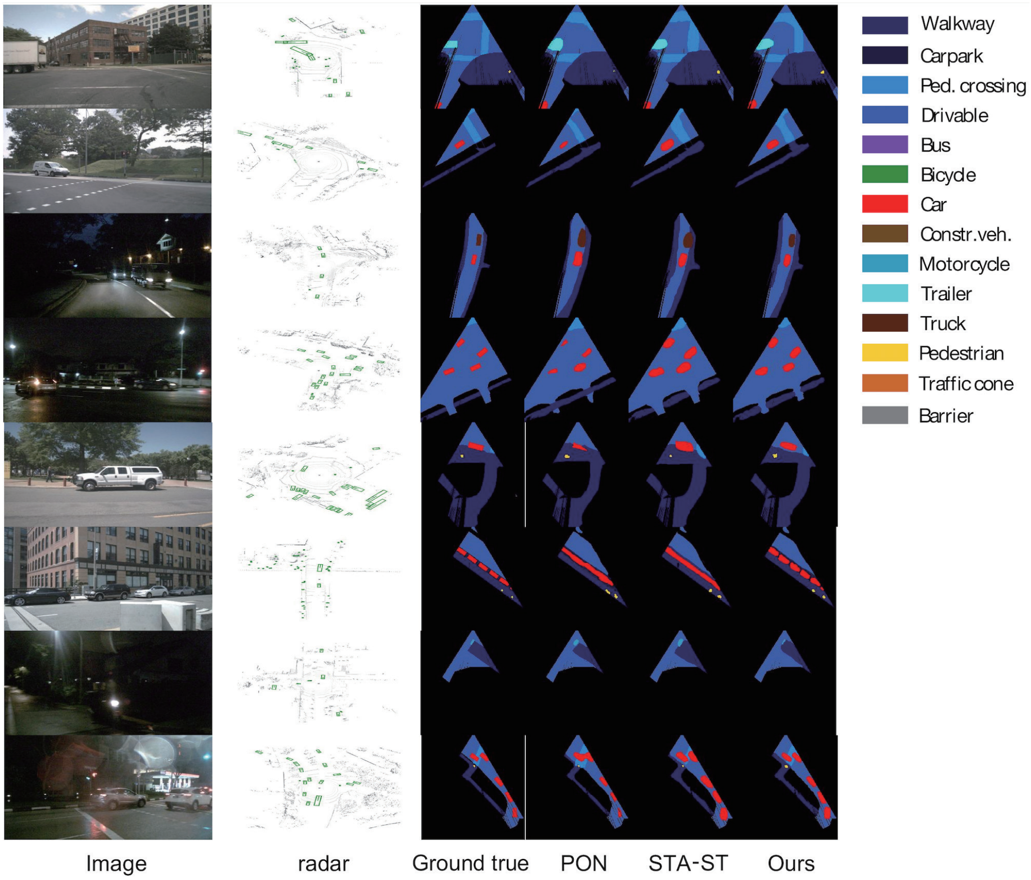

Fig. 3. (Color online) The sensors self-compile qualitative results on the NuScenes and Cityscapes validation sets at any time of day and night, front and side.

information on the road, such as pedestrians, parking lot signs, drivable cars, buses, bicycles, zebra crossings, traffic signs, and roadblocks, and presents the prediction results in a visualized

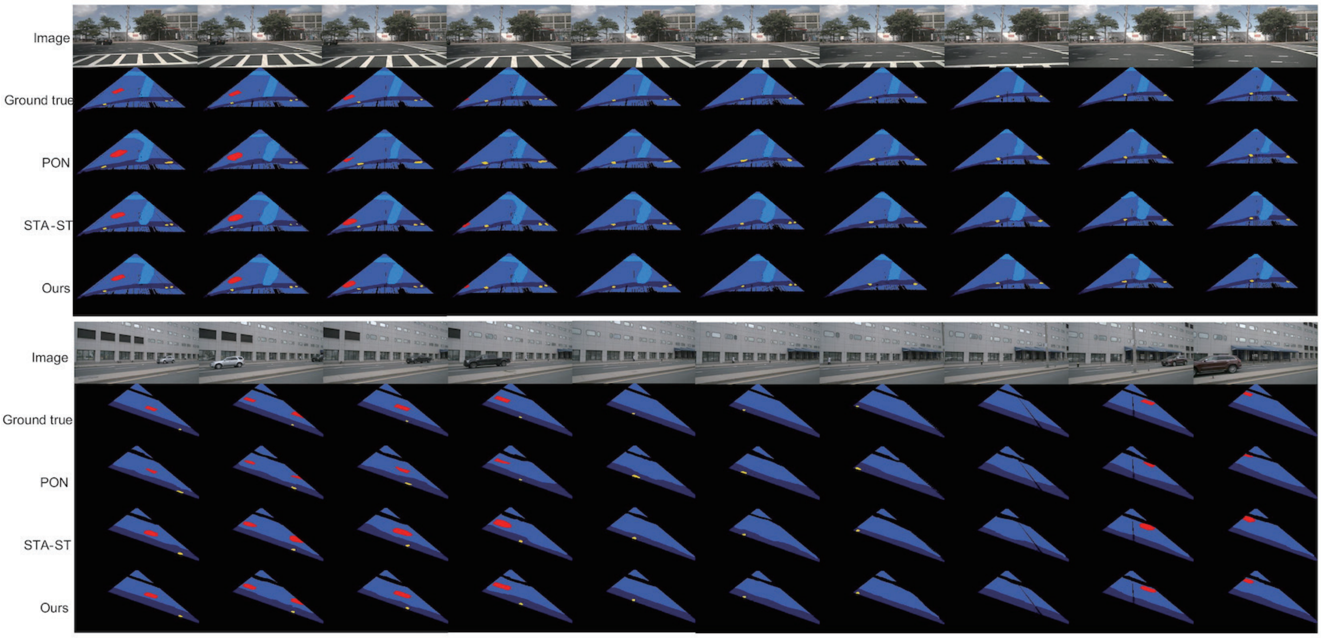

Fig. 4. (Color online) Visualization comparison of different models predicting automotive elements on 10 consecutive keyframes in NuScenes.

manner on the vehicle's display screen. The method can help autonomous vehicles perceive their surroundings better, improving safety and efficiency.

In developing intelligent transportation systems, the accuracy of dynamic temporal prediction is crucial. In this study, two large datasets are utilized, NuScenes and Cityscapes, to validate and demonstrate the performance of the prediction model in natural traffic environments.

Through these two datasets, high-precision predictions of the motion trajectories of targets such as vehicles, pedestrians, and bicycles can be made, which is of considerable significance for enhancing the safety and efficiency of the transportation system.The model can precisely capture and predict the motion trajectories of targets, including the position, speed, shape, size, and direction of the targets. In NuScenes and Cityscapes, the automotive elements are predicted in 10 consecutive keyframes, demonstrating the effectiveness and limitations of different models in handling complex traffic scenarios. The method integrates data from multiple perspectives to more accurately capture the dynamics of cars, including their trajectories and behavior patterns. In addition, the model considers historical data and future trends, generating smoother and more coherent motion trajectories when predicting the dynamics of automotive elements.

Including Cityscapes further enhances the generalization ability of the model. By validating the model in different urban environments, its high accuracy and robustness in various traffic scenarios can be ensured. This cross-dataset analysis and application are crucial for understanding complex traffic conditions and making accurate driving decisions. These accurate predictions improve the safety of autonomous vehicles and help optimize traffic flow, reduce congestion, and improve the efficiency of the entire transportation system. The research provides strong technical support for efficient and safe traffic management in intelligent transportation systems based on visual perception, especially in predicting complex traffic environments based on BEVs. Through the model, intelligent transportation systems can better adapt and respond to various road conditions, providing more accurate and real-time data support for autonomous driving and traffic management. The visual comparison of different models of predictive automotive elements in NuScenes considering 10 consecutive keyframes of timing information is shown in Fig. 4.

5. Conclusion

In this paper, we introduce a novel approach that combines object detection and semantic prediction (BEV) to produce high-quality intelligent traffic maps using the data collected from onboard cameras. Leveraging advanced sensor technology, including high-resolution cameras and multisensor fusion, detailed road scene information was captured, and precise data inputs were provided. Additionally, an efficient data transmission mechanism facilitated the real-time transmission of collected images and information to the central processing unit, ensuring the accuracy of the map generation process.

Addressing challenges inherent in the semantic prediction (BEV) of traffic elements, such as long-distance and multiscale targets, as well as the precision of critical road sign shapes, the method demonstrates effectiveness. By effectively leveraging sensor perception capabilities, various traffic elements—vehicles, pedestrians, traffic signs, and roadblocks—were accurately identified and distinguished, and the obtained information was transmitted to the analysis algorithm. The innovative approach was validated through extensive experiments on Cityscapes and NuScenes, showcasing its superiority in generating BEV representations and handling relationships between multiple targets and categories in complex scenarios. Integrating advanced sensor data processing and transmission technologies, the method not only offers a novel solution for robotic vision tasks but also provides valuable insights for interdisciplinary fields reliant on visual information acquisition and understanding.

In summary, this research presents a significant contribution—an effective method for generating accurate and information-rich intelligent traffic maps from the data captured by onboard cameras. Such advancements hold promise for furthering research and applications in intelligent transportation and autonomous driving realms.

Acknowledgments

The authors would like to acknowledge the support of the National Natural Science Foundation of China under Grant No. 62376246.

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.

References

1 J. Yu, Y. Xu, and H. Chen: IEEE Trans. on Neural Networks and Learning Systems (2022).
2 H. Qiu, X. Liu, S. Rallapalli, A. J. Bency, K. Chan, R. Urgaonkar, B. S. Manjunath, and R. K. K. Govindan: Proc. 2018 IEEE/ACM 3rd Int. Conf. Internet-of-Things Design and Implementation (IoTDI) (2018) 48–59.
3 R. Zhang and S. Cao: IEEE Access **7** (2019) 137065.
4 J. Yu, H. Gao, Y. Chen, D. Zhou, J. Liu, and Z. Ju: IEEE Trans. Hum. Mach. Syst. **52** (2022) 784.
5 Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew: Int. J. Multimed. Inf. Retr. **7** (2018) 87.
6 M. J. Daily, J. G. Harris, and K. Reiser: Proc. Image Understanding Workshop.
7 J. Ma, Y. Liu, M. Han, C. Hu, and Z. Ju: IEEE Trans. Neural Networks Learn. Syst. (2023) 1. https://doi.org/10.1109/TNNLS.2023.3319661
8 M. M. Rahman, Y. Tan, J. Xue, and K. Luy: IEEE Trans. Image Process **29** (2019) 2947.
9 J. Yu, W. Zheng, Y. Chen, Y. Zhang, and R. Huang: Front. Neurosci. **17** (2023) 1219363.
10 S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis: IEEE Trans. Intell. Transp. Syst. **23** (2020) 33.
11 Y. Fan, Z. Feng, A. Mannan, T. U. Khan, C. Shen, and S. Saeed: Remote Sens. **10** (2018) 1845.
12 B. Liu, W. Chen, Z. Wang, S. Pouriyeh, and M. Han: Electronics **12** (2023) 3084.
13 Z. Wang, J. Huang, K. Miao, X. Lv, Y. Chen, B. Su, L. Liu, and M. Han: Comput. Netw. **236** (2023) 110021.
14 J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi: Proc. Computer Vision—ECCV 2020: 16th European Conf. (2020) 720.
15 J. Yu, H. Gao, D. Zhou, J. Liu, Q. Gao, and Z. Ju: IEEE Trans. Cybern. **52** (2021) 13738
16 B. Liu, W. Chen, Z. Wang, and S. Pouriyeh, and M. Han: Electronics **12** (2023) 3084.
17 A. García, G. D. Valbuena, A. García-Tuero, A. Fernández-González, J. L. Viesca, and A. H. Battez: Appl. Sci. 12 (2022) 3608.
18 S. Thompson and D. Sheat: Proc. 9th Int. Conf. Road Transport Information and Control (1998) 59.
19 T. Liu, J. Yang, B. Li, C. Xiao, Y. Sun, Y. Wang, and W. An: IEEE Trans. Geosci. Remote Sens. **60** (2021) 5614718.
20 Z. Wang, J. Huang, K Miao, X. Lv, Y. Chen, B. Su, L. Liu, and M. Han: Comput. Netw. **236** (2023) 110021.
21 R. Sebastien and F. Jurie: J. Vis. Commun. Image Represent. **34** (2016) 187.
22 K, Wang, S. Du, C. Liu, and Z. Cao: IEEE Trans. Geosci. Remote Sens. **60** (2022) 5002013.
23 C. Hu, Z. Liu, R. Li, P. Hu, T. Xiang, and M. Han: IEEE Trans. Dependable Secur. Comput. **21** (2023) 2145. https://doi.org/10.1109/TDSC.2023.3300749
24 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin: Proc. 31st Int. Conf. Neural Information Processing Systems (NIPS, 2017) 129–130. https://arxiv.org/abs/1706.03762
25 H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (2020) 11621–11631.

About the Authors

Zhiguo Ma received his B.S. degree from North China Electric Power University, China, in 2006 and his M.S. degree from Zhejiang University, China in 2013. Currently, he is a Ph.D. candidate of Zhejiang University. His research interests are in data intelligence, machine learning, and intelligent systems. (11921159@zju.edu.cn)

Yutong Zhang received her B.S. degree in measurement and control technology and instruments from Shenyang Ligong University, China, in 2022. She is now working on her M.S. project in intelligent systems at Shenyang Ligong University, China, and collaborative training at Binjiang Institute of Zhejiang University, China. Her research interests include human motion analysis, healthcare robotics, and human–computer interaction. (yutongzhang@stu.sylu.edu.cn)

Meng Han received his Ph.D. degree in computer science and MBA degree from Georgia State University and Georgia Institute of Technology in 2021 and 2017, respectively. He is currently the Director of the Intelligent Fusion Research Center (IFRC). His research interests include data-driven intelligence, data security and privacy, and AI governance. (mhan@zju.edu.cn)