Boundary-Aware Semantic Bird-Eye-View Map Generation Based on Conditional Diffusion Models

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

***Abstract*—Semantic bird-eye-view (BEV) map is an efficient data representation for environment perception in autonomous driving. In real driving scenarios, the collected sensory data usually exhibit class imbalance. For example, road layouts are often the majority classes and road objects are the minority. Such imbalanced data could lead to inferior performance in BEV map generation, particularly for minority objects due to insufficient learning samples. This work attempts to mitigate this issue from the perspective of network and loss function design. To this end, a diffusion-guided semantic BEV map generation network with a boundary-aware loss is proposed. The network learns the underlying distribution of the data, including the relationship between majority and minority classes. The boundary-aware loss increases weighting for minority classes during training, making the network focus on these classes. Experimental results on a public dataset demonstrate our superiority over the state-of-the-art methods, and our effectiveness in addressing the class imbalance issue.**

***Index Terms*—Semantic BEV map, class imbalance, semantic scene understanding, autonomous driving.**

I. Introduction

EFFICIENT data representation is essential for self-driving vehicles to perceive surrounding environments. The semantic bird-eye-view (BEV) map stands out as a popular choice in this area owing to its efficiency. It has attracted significant attention in autonomous driving research communities. Compared with the common front-view data representation, the notable Characteristics of BEV are evident in the following aspects: 1) its top-down perspective facilitates the seamless integration of information from heterogeneous sensors, such as cameras and LiDAR; 2) it functions as an intermediate data, effectively bridging the disparity between the real world and simulation environments through its semantic representation; 3) it eliminates distortions caused by perspective projection, making it an inherently suitable option for downstream tasks that rely on grid-like data representation, such as trajectory prediction [1], [2] and autonomous navigation [3], [4], [5].

Semantic BEV map generation involves both semantic labeling and view transformation from front-view to top-down view. For semantic labeling, correctly labeling road objects, such as vehicles, pedestrians, and barriers, is important to the safety of autonomous driving [6], [7]. However, in the real collected sensory data from self-driving vehicles, such as visual images, road objects usually have small sizes, while road layouts, such as drivable areas, and walkways, usually occupy a larger portion. Such a case leads to class imbalance, which is a common issue in semantic scene understanding [8]. Fig. 1 demonstrates the pixel distribution between different semantic classes in the nuScenes dataset [9], a well-known public dataset for autonomous driving. The figure visualizes the imbalance between road layouts and road objects. This data imbalance issue could lead to inferior segmentation accuracy for the minority objects due to insufficient learning samples. Unfortunately, virtually all the existing semantic BEV map generation methods focus on network design to improve the overall segmentation performance across all classes, overlooking this class imbalance issue.

The Transformer model has been widely used in BEV perception [10], [11], [12]. These methods use BEV queries to look up the BEV feature from the front-view information via the cross-attention mechanism. The BEV queries are randomly initialized as blank templates in the same format as the required output (e.g., the detection boxes in the object detection task or the BEV map in the semantic BEV map generation task), and the target representation is gradually learned from the source domain during training. However, those methods also do not take the class imbalance problem into consideration.

In the fields outside autonomous driving, there are some attempts to solve this issue. These attempts can be generally categorized into data-level methods and cost-sensitive methods. The former uses various data augmentation approaches [13], [14], [15] to create new minority data. Besides data augmentation, oversampling [16], [17] is also a commonly used data-level method to enrich datasets. The latter [18], [19] focuses on developing new loss functions, guiding the network to learn from the minority classes.

In line with the latter category of methods, we try to solve the class imbalance issue by introducing a boundary-aware loss, which was initially introduced in our conference paper [20]. The boundary-aware loss emphasizes minority classes

Received 19 September 2024; revised 16 December 2024; accepted 16 April 2025. Date of publication 24 April 2025; date of current version 6 October 2025. This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010116 and in part by the City University of Hong Kong under Grant 9610675. This article was recommended by Associate Editor X. Chang. *(Corresponding author: Yuxiang Sun.)*

Shuang Gao is with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, and also with Harbin Institute of Technology, Harbin 150001, China (e-mail: gaoshuang.sarah@outlook.com).

Qiang Wang is with Harbin Institute of Technology, Harbin 150001, China (e-mail: wangqiang@hit.edu.cn).

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

Digital Object Identifier 10.1109/TCSVT.2025.3564002

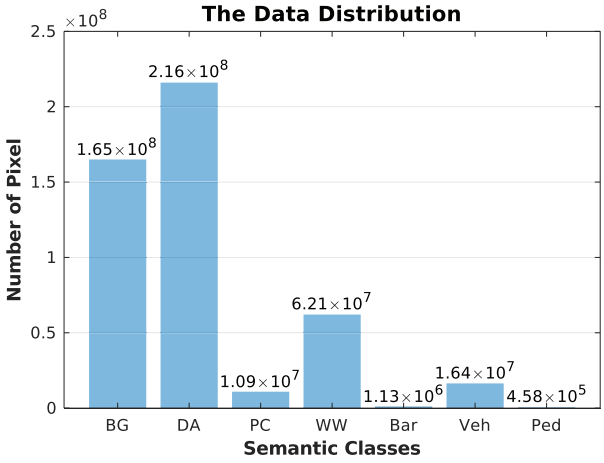

Fig. 1. The statistics of the pixel number for each semantic category in the nuScenes dataset. For our semantic BEV map generation task, we create the semantic BEV ground truth for 7 classes, including the background (BG), drivable area (DA), pedestrian crossing (PC), walkway (WW), barrier (Bar), vehicle (Veh), and pedestrian (Ped).

during training, thereby improving the network sensitivity to minority objects. However, experimental results indicate that relying solely on the loss function to improve minority class detection can compromise the accuracy of majority class predictions. To mitigate this issue, we further refined the network's design to achieve a more balanced focus across classes.

Recently, the diffusion model [21] has attracted great attention due to its ability to model the underlying data distribution. The denoising process of the diffusion model can capture the distribution of both majority and minority classes, benefiting the detection of minority objects while having less impact on the majority classes. In this work, we use the diffusion model to initialize the BEV queries in the cross-attention module of the Transformer, rather than use blank templates. In such a way, the BEV queries learned by the diffusion model could be seen as the prior information that encodes the global scene understanding, making the queries stick to physical principles. For example, pedestrians should appear on top of roads rather than on top of a car. The road layouts should be subject to certain changing rules instead of sudden changes in shape. Thus, we propose a diffusion-guided semantic BEV map generation network with boundary-aware loss. Our motivation is to improve the segmentation performance for the minority classes. This paper is an extension of our conference paper [20]. Our code is open-sourced.[1] The contributions are summarized as follows:

1) We design a novel diffusion-guided semantic BEV map generation network.
2) We design the boundary-aware loss to mitigate the impact of the class imbalance problem.
3) We compare our network with the state of the arts to demonstrate our superiority.

The remainder of this paper is structured as follows: Section II reviews the related work. Section III delves into the details of our network. Section IV showcases our experiment results. The last section summarizes our work and suggests potential directions for future research.

[1]https://github.com/lab-sun/Boundary-aware-BEV

II. Related Work

A. Semantic Segmentation Networks

Semantic segmentation aims to classify each pixel in an image into individual classes [22], [23], [24]. U-Net [25] introduced the skip connection to the encoder-decoder structure, enabling the information interaction between up- and down-sampling. U-Net++ [26], and U-Net3+ [27] are the U-Net-based methods, attempting to create new information fusion manners for the skip connections. The DeepLab family [28], [29], [30] employed the atrous convolution to obtain the different receptive field ranges by setting the various dilation rates. Nowadays, the Transformer model [31] has become prevalent due to its remarkable ability to extract attention. Zheng et al. [32] proposed a SETR network, extending the ViT [33] into semantic segmentation. Chen et al. [34] designed a hybrid model for semantic segmentation, combining convolution operations and self-attention mechanisms. Leveraging extensive training on large-scale datasets, the SAM [35] attained a high-level generalization in segmentation tasks across diverse domains. MASA [36] utilized the powerful segmentation capabilities of SAM to effectively accomplish cross-frame segmentation of identical objects.

B. Semantic BEV Map Generation

The semantic BEV map finds extensive applications in a variety of practical scenarios [37], [38], establishing its significance within the perception domain. Generating semantic BEV maps differs from semantic segmentation as it conducts both view transformation and semantic prediction. Lu et al. [39] first introduced semantic BEV map generation into driving scenarios via a convolutional variational encoder-decoder structure. Monolayout [40] utilized the adversarial learning framework to respectively predict the static and dynamic semantics. LSS [41] designed a depth prediction mechanism to lift the 2D feature to the 3D space during the view transformation. Also based on the idea of depth prediction, BEVDepth [42] used a network to predict depth with the additional depth supervision. Yang et al. [43] used a cross-attention module in the semantic BEV map generation to look up the BEV feature from the front-view image. BEVFormer [12] introduced the Transformer framework to perform the BEV perception. S2G2 [44] employed the semi-supervised framework to alleviate dependence on labeled data. Additionally, there exists research [45] aimed at forecasting the future driving environment using semantic BEV maps.

C. Class Imbalance Learning

Long-tailed distribution is a common issue in deep learning. This could hinder the learning capacity for minority classes. Many methods have been proposed to address this issue. Some methods [14], [15], [16] explored the data-level solution by enriching data diversity or oversampling the minority classes

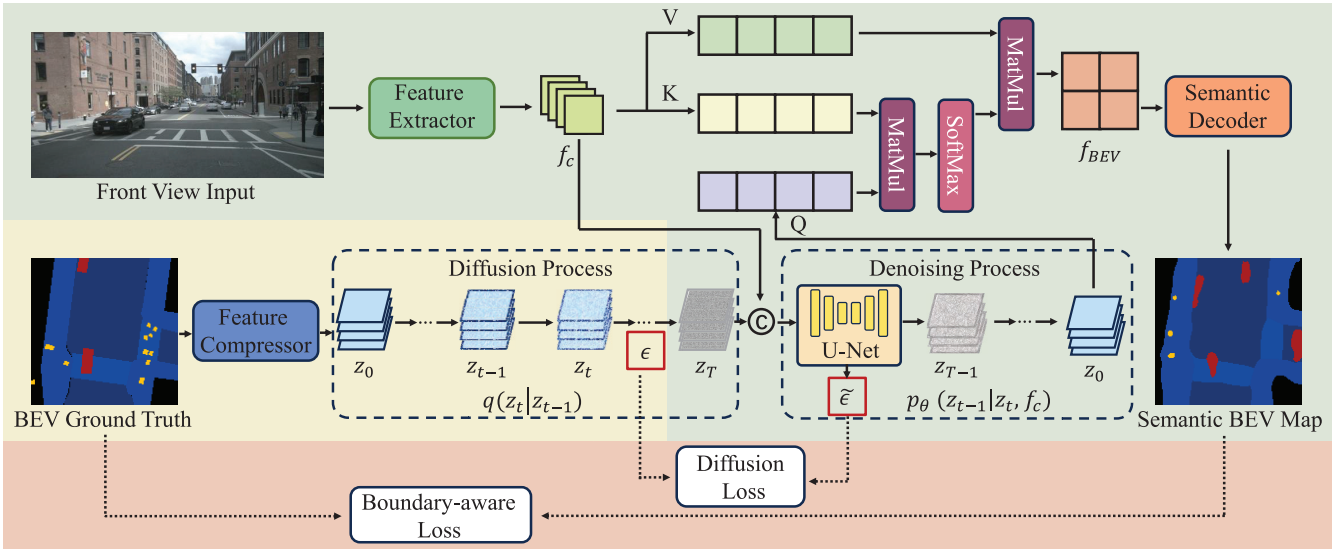

Fig. 2. The architecture of our proposed network. The front-view RGB image is taken as a condition to guide the diffusion module after passing through the feature extractor. The compressed BEV ground truth is corrupted with stochastic noise iteratively to get a pure Gaussian noise. The denoising process recovers the latent BEV feature map structure step-by-step with a U-Net structure, during which the underlying pattern of the data distribution can be learned. Then, we use this recovered BEV feature map to initialize the BEV queries in the cross-attention mechanism instead of the blank query templates to get the BEV feature map, f_{BEV}. The semantic BEV map is generated after feeding f_{BEV} to the semantic decoder. The boundary-aware loss and diffusion loss are employed to train this network. Note that the procedures shown in the yellow background are implemented only in the training phase, while those highlighted in green represent the inference steps. The figure is best viewed in color.

to re-balance the data. Ebenuwa et al. [46] tackled this issue by introducing a new attribute selection technique, variance ranking, to address class imbalance problems. The cost-sensitive approaches, like [18] and [19], designed proper loss functions to guide the network training by assigning larger weights to minority classes, thereby addressing the class imbalance issue.

D. Diffusion Model

Recently, the diffusion model has been well-studied as a powerful generative model, which learns to simulate and generate the underlying data distribution of the training set by iteratively applying a denoising process. Diffusion is a well-known concept in non-equilibrium statistical physics, which was first introduced into the deep-learning field by Sohl-Dickstein et al. [47]. Ho et al. [21] proposed Denoising Diffusion Probabilistic Models (DDPM), laying the foundation for image generation. During the diffusion phase, the original image is corrupted with random noise step-by-step until it becomes a white Gaussian noise. Then, a neural network is trained to predict the noise added in each step to recover the structure of the input image. According to whether a prompt is provided to guide the generation process, the diffusion model can be divided into conditional and unconditional ones. The unconditional diffusion [48], [49], [50] randomly samples from the learned data distribution as the generation, while the conditional ones [51], [52], [53] take the text or images as the control of the generation to obtain an image of the specified content. Since the diffusion and denoising processes are implemented in the original image size, it takes a long time to sample an image by simulating a Markov chain for many steps. To accelerate sampling, the Latent Diffusion Model [54] was proposed by compressing the original image into the smaller latent space with an auto-encoder.

III. The Proposed Network

A. The Overall Architecture

To reduce the impact caused by the class-imbalanced training data on the network performance, we propose a diffusion-guided semantic BEV map generation network with a boundary-aware loss function. The overall architecture of the proposed network is illustrated in Fig. 2. During the training phase, the network takes as input the front-view images and the corresponding BEV ground truth. The two types of inputs are fed into a feature extractor and a feature compressor, respectively. The former extracts the environment information from the natural driving surroundings, while the latter maps the high-dimensional BEV labels into the latent space to accelerate the diffusion sampling. The diffusion process adds the stochastic noise to the latent BEV feature, z_0, step by step until a pure Gaussian noise is obtained. Note that the diffusion process is only applied during training but not during inference. Subsequently, U-Net is employed to remove the added noise iteratively under the guidance of the front-view feature map, f_c. This denoising process can recover the original BEV features from the pure Gaussian noise. This recovered BEV feature includes the prior information on the various classes' distribution patterns learned from the diffusion process. Thus, we take the recovered BEV feature map as the BEV query to look up the related information from the front-view feature maps in a cross-attention module, which initializes the query with the prior underlying distribution of the data. After the semantic decoder, the BEV feature, f_{BEV}, is decoded into the semantic BEV map. In contrast to

certain diffusion-based segmentation methods, the proposed method leverages the inherent strengths of diffusion models, particularly their capacity to effectively capture the underlying patterns of data distributions, which specifically addresses the class imbalance problem.

The whole network is end-to-end trainable by punishing boundary-aware and diffusion losses. The boundary-aware loss focuses on the detection performance of the minority classes. The diffusion loss reduces the disparity between the noise added during the diffusion process and the noise predicted during the denoising process.

B. The Feature Extractor for the Diffusion Condition

Our network aims to generate a corresponding semantic BEV map from a given front-view RGB image, leveraging the powerful generative capabilities of the diffusion model. Since the diffusion model could generate an arbitrary random sample that conforms to the distribution pattern of the training dataset, our task requires guidance to control the generation process to get a semantic BEV map that aligns with the front-view image. Therefore, we use the feature map extracted from the front-view RGB image as the condition for the diffusion model. In this context, it is expected that both global and local environment information included in the front-view image is accurately represented, thereby providing high-quality conditions for the diffusion model.

We employ a pre-trained CNN model, DeepLab V3+ [30], as our feature extractor. The ASPP module in DeepLab V3+ enables multi-scale feature extraction by using various dilation rates, capturing the information at various scales. The front-view RGB image $\mathbf{I}_{FV} \in \mathbb{R}^{3\times H\times W}$ is provided to the feature extractor $\mathcal{G}$ to get the diffusion condition, f_c:

$$f_c = \mathcal{G}(\mathbf{I}_{FV}). \tag{1}$$

C. The Diffusion-Guided BEV Queries

1) The Feature Compressor: We integrate the diffusion model with the cross-attention mechanism to make the BEV queries encode the prior information of the dataset. In this way, the distribution patterns of the majority and minority classes could be captured by the BEV queries. The structure of the data distribution is disrupted through the iterative forward diffusion process, during which random noise is added to the original image. Then, a reverse denoising network is trained to restore the data structure from the Gaussian noise. The reverse step is performed at the original image scale and repeated for T timesteps to predict the noise added at each step of the forward process, leading to a low sampling rate. However, the BEV label is composed of several semantic areas without complicated details, allowing a compact representation to hold all the information. To speed up the sampling, we first shrink the BEV label to a small size with the feature compressor and implement the diffusion model on the latent space instead of pixel space.

A variational auto-encoder model [55] is used to trim off redundant pixel-level information. We pre-train this model by first encoding the BEV label $\mathbf{y}_{BEV}$ into the latent space and then decoding the latent into the original space. In our proposed network, only the encoder, $\mathcal{H}$, is preserved to provide the diffusion process with a reliable compact representation, z_0:

$$z_0 = \mathcal{H}(\mathbf{y}_{BEV}). \tag{2}$$

2) The Conditional Diffusion: A diffusion model comprises two key processes: the forward diffusion process and the reverse denoising process, both of which can be conceptualized as Markov chains. The forward process begins with clear input data and progressively corrupts the data by adding random noise over T timesteps, where each step depends on the noisy sample from the preceding step. This gradual addition of noise transforms the original complex data distribution into a simplified isotropic distribution. Conversely, the reverse process seeks to reconstruct the original data distribution by iteratively removing the noise. Guided by a parameterized model, it learns to reverse the forward steps, withdrawing the added noise at each stage until the original structure is restored. Intuitively, the forward process can be understood as a controlled mechanism that incrementally disperses the data into a simpler representation, while the reverse process functions as a stepwise recovery system, restoring the intricate details of the initial pattern.

In our case, the BEV latent tensor follows a certain distribution, $z_0 \sim q(z)$. In the forward diffusion process, a small amount of Gaussian noise incrementally added to the latent sample in each timestep is determined by a predefined variance schedule, $\beta_t \in (0, 1)$, yielding a series of noisy samples z_0, z_1, $\cdots$, z_T:

$$q(z_{1:T}|z_0) = \prod_{t=1}^{T} q(z_t|z_{t-1}), \tag{3}$$

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t;\ \sqrt{1-\beta_t}z_{t-1}, \beta_t \mathrm{I}). \tag{4}$$

The diffusion process q gradually converts the complex distribution (the data, z_0) to the sample Gaussian distribution (the pure Gaussian noise, z_T). Since this process is subject to the Markov chain and variance schedule β_t is known, the noisy sample can be calculated at arbitrary step t:

$$q(z_t|z_0) = \mathcal{N}(z_t;\ \sqrt{\bar{\alpha}_t}z_0, (1-\bar{\alpha}_t)\mathrm{I}), \tag{5}$$

$$z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{6}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Eq. 6 is the re-parameterized form to get z_t, where $\epsilon \sim \mathcal{N}(0, \mathrm{I})$. At T step, $z_T \sim \mathcal{N}(0, \mathrm{I})$, becomes the isotropic Gaussian distribution.

The denoising process reverses the previously mentioned operation by training a neural network p_θ to approximate the conditional probabilities $q(z_{t-1}|z_t, f_c)$, conditioning on the front-view feature, f_c:

$$p_\theta(z_{0:T}|f_c) = p(z_T) \prod_{t=1}^{T} p_\theta(z_{t-1}|z_t, f_c), \tag{7}$$

$$p_\theta(z_{t-1}|z_t, f_c) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t, f_c), \Sigma_\theta(z_t, t, f_c)). \tag{8}$$

We follow [21], and set $\Sigma_\theta(z_t, t, f_c) = \sigma_t^2 \mathrm{I}$ as an untrained constants, where $\sigma_t^2 = \beta_t$. μ_θ is learned by the parameterized model θ. Because the back-propagation cannot be performed

on the random sampling process, the re-parameterization operation is introduced to sample z_{t-1} from $p_\theta(z_{t-1}|z_t, f_c)$:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t, f_c)\right) + \sigma_t n, \tag{9}$$

where $n \sim \mathcal{N}(0, \mathrm{I})$, and $\epsilon_\theta(z_t, t, f_c)$ represents the U-Net we used in the denoising process, which predicts the noise added at the t step, given the noisy sample z_t, the time embedding t, and the front-view condition f_c as input. To supervise this conditional diffusion model, we minimize the distance between the noise added in the forward diffusion process and the predicted noise from the U-Net in the reverse denoising process at each time step:

$$L_{diff} = E_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, f_c)\|^2\right]. \tag{10}$$

The forward diffusion and reverse denoising processes are displayed in Fig. 2. Note that the diffusion process is not implemented in the inference phase, instead, we use a random noise sampled from the Gaussian distribution as input during inference.

3) The Cross-Attention Module: The transformation from the front view to the bird-eye view can be interpreted as a perspective translation. To facilitate this process, the cross-attention mechanism is employed, enabling the model to selectively focus on specific regions of the front-view data that are relevant to the BEV perspective through a query-based operation. Consequently, the use of an appropriately initialized query tensor can significantly enhance the effectiveness of the view transformation. The diffusion model reveals underlying patterns across various semantic classes through its noise corruption and denoising steps. Within the cross-attention module, samples derived from the learned data distribution serve as initializations for the BEV queries, facilitating the model's ability to capture relationships among semantic classes, particularly between majority and minority classes.

As displayed in Fig. 2, the key and value vectors are calculated from the front-view feature map f_c, and the BEV queries are initialized with the BEV feature z_0 sampled from the learned data distribution:

$$\begin{aligned} Q &= \mathbf{W_Q} z_0, \\ K &= \mathbf{W_K} f_c, \\ V &= \mathbf{W_V} f_c, \end{aligned} \tag{11}$$

where $\mathbf{W_Q}$, $\mathbf{W_K}$, $\mathbf{W_V}$ are the trainable weight matrices. The cross-attention module establishes the matching relationship between the BEV and the front-view features through the query operation, which is measured by the attention score:

$$f_{BEV} = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{12}$$

where d is the dimension of the BEV queries.

After the cross-attention module, the BEV feature map f_{BEV} encodes the information from both the front-view and BEV. Then, the semantic decoder up-samples the BEV feature to generate the semantic BEV map.

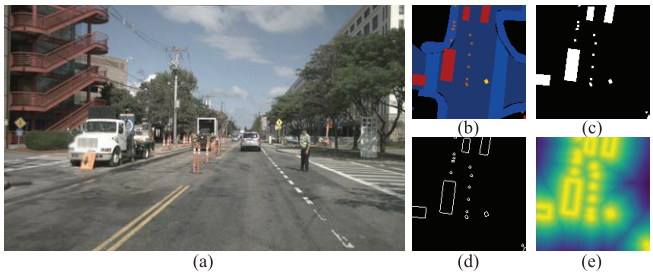

Fig. 3. The visualization of the boundary-aware loss. (*a*) is the input front-view image. (*b*) is the corresponding ground truth. (*c*) is the binary mask for the minority objects that appear in the image. (*d*) is the extracted edges of the minority objects. (*e*) is the boundary-aware score. Note that the brighter color represents the higher score. The figure is best viewed in color.

D. The Boundary-Aware Loss

Due to the data scarcity of minority classes (e.g., pedestrians), semantic BEV map generation networks may not perform well in detecting these objects. So, besides the network design, a boundary-aware loss is introduced in this work to address this issue. This loss is expected to guide the network to learning process and improve its ability to accurately detect minority classes during training phase.

Fig. 1 shows the statistics on the number of pixels of each class in our used dataset, which includes 7 semantic classes. We group the background (BG), drivable area (DA), pedestrian crossing (PC), and walkway (WW) into the majority classes, and the barriers (Bar), vehicle (Veh), and pedestrian (Ped) into the minority classes. To stress the minority classes, it is intuitive to increase the weight of these objects in loss calculation during training. Given that the loss function measures the difference between the predicted output of the proposed network and the ground truth, assigning higher loss weights to minority classes can guide the network toward better performance in detecting these classes through the iterative backpropagation and weight adjustments.

To implement this loss function, we begin by processing the semantic BEV label. The one-hot encoded label is transformed into a single-channel image where each pixel is assigned its corresponding class ID. Then, the objects belonging to the minority classes are gathered onto a binary mask $M \in \mathbb{R}^{h\times w}$, distinguishing from the background and other classes:

$$M = \begin{cases} \mathbf{V}(u,v) = 0, & \mathbf{V}(u,v) < T \\ \mathbf{V}(u,v) = 255, & \mathbf{V}(u,v) \geq T, \end{cases} \tag{13}$$

where $\mathbf{V}(u, v)$ is the pixel value of the single-channel label. T is the threshold to pick up the minority classes. Fig. 4 (c) displays the binary mask M. Subsequently, the object edges are extracted using the Canny edge detection [56] (shown in Fig. 4 (d)). The last step is to calculate the Euclidean distance from each pixel to the nearest edge as the boundary-aware score S. Note that the smaller the pixel distance, the higher its score. The visualization of this score is displayed in Fig. 4 (e), where the brighter parts indicate the higher boundary-aware scores.

This boundary-aware score is applied to the Mean Squared Error (MSE) loss calculation. We assign the score to each pixel

Fig. 4. The bar charts for the impacts of the boundary intensity factor (γ) on the mIoU and mAP. The performance of the majority, minority, and overall classes are reported separately. The x-axis represents various γ values, while the upper y-axis denotes the mIoU scale and the lower y-axis represents the mAP scale. Certain areas of the y-axis are magnified to enhance the visibility of differences. The figure is best viewed in color.

to enhance the network's attention to the minority objects:

$$L_{bound} = \sum_{u=1}^{H}\sum_{v=1}^{W}(S^{u,v})^{\gamma}(\mathbf{y}_{BEV}^{u,v} - \mathbf{p}_{BEV}^{u,v})^2, \tag{14}$$

where $\mathbf{y}_{BEV}^{u,v}$ and $\mathbf{p}_{BEV}^{u,v}$ stand for the pixel (u, v) in the semantic BEV label and the predicted map, respectively. $S^{u,v}$ is the boundary-aware score for the pixel (u, v) and γ is a hyperparameter to control the boundary intensity. During training, the Focal loss [57] is employed to supervise the segmentation. The total loss of the proposed network is calculated as follows:

$$L = L_{seg} + \delta L_{bound} + \lambda L_{diff}, \tag{15}$$

where δ and λ are the weight parameters to balance different loss functions. They are both set to 1 in our experiments.

IV. Experimental Results and Discussions

A. The Dataset and Training Details

We use the public dataset, nuScenes [9], to train and test the performance of the proposed network. The dataset is collected from different regions of Singapore and Boston, with a total of 850 sequences of annotated scenes. The annotation data includes 3D object bounding boxes, HD maps, scene semantic information, and camera matrices. We project the 3D bounding boxes onto the HD map to create the semantic BEV labels for training. Note that due to the limitations of the flat plane assumption, some semantic BEV labels may not be generated correctly. After excluding scenes containing incorrect labels, we randomly divide the entire dataset into 548 training sets, 150 validation sets, and 148 testing sets. The size of the input image is 256×512 and the semantic BEV labels contain 150×150 semantic grids, each with a resolution of $0.2\times0.2\ m^2$.

The proposed network is implemented on NVIDIA RTX 3090 with 24GB GPU memory. The learning rate for the diffusion module and the rest of the network is initialized to 1×10^{-4} and 1×10^{-3}, respectively. The warmup strategy is applied to the adjustment of the learning rates. The weight decay is set to 1×10^{-1}. We train our proposed network for 200 epochs. The total time steps for the diffusion sampling process is 1000. The hyperparameter γ controls the intensity of the boundary-aware score during the loss calculation, which is set to 1.5. The details of the hyperparameter selection are discussed in the ablation study.

TABLE I

The Results (%) of the Ablation Study on the Diffusion Timesteps. The Timesteps Control the Noise Level in Each Forward and Inverse Step. We Assess the Network's Performance Across the 100, 500, 1000, and 1500 Timesteps, Reporting Outcomes for Both Majority and Minority Classes. The Best Results Are Highlighted in Bold Font

timesteps	Majority		Minority		mIoU	mAP
	mIoU	mAP	mIoU	mAP		
100	49.77	67.58	11.42	27.29	33.33	50.31
500	50.00	64.82	11.64	25.63	33.56	49.77
1000	**50.50**	64.98	**11.89**	**27.47**	**33.95**	**50.52**
1500	50.26	**64.99**	11.17	26.38	33.51	50.25

B. Ablation Study

We perform ablation studies to validate the effectiveness of the proposed network and to optimize parameter selection. For evaluation metrics, we use the mean Intersection over Union (mIoU) and mean Average Precision (mAP).

1) Ablation on the Diffusion Process: The diffusion module incrementally introduces the random noise to the original data over T time steps during the forward process, until the data is transformed to pure Gaussian noise. Then, the reverse process eliminates the noise, step by step, to reconstruct the original data. The time step T determines the amount of noise added in the forward steps and removed in the reverse steps, thereby impacting the overall diffusion process. Thus, we conduct an ablation study to evaluate the network performance under varying diffusion time steps. The time step is respectively set to 100, 500, 1000, and 1500.

Since this work aims to mitigate the impacts of class imbalance, we focus on the network performance across both minority and majority classes. The results shown in Tab. I, report the performance of these classes in terms of mIoU and mAP. The data in the table indicates that both excessively

TABLE II

THE RESULTS (%) OF THE ABLATION STUDY ON THE SIZE OF THE LATENT FEATURE. TO ACCELERATE THE SAMPLING RATE, THE DIFFUSION PROCESS IS CONDUCTED IN THE LATENT SPACE. WE ASSESS THE NETWORK'S PERFORMANCE TO DETERMINE THE OPTIMAL LATENT FEATURE SIZE AMONG THE RESOLUTION OF 64×64, 32×32, AND 16×16. THE OUTCOMES ARE MEASURED USING MIOU AND MAP FOR BOTH MAJORITY AND MINORITY CLASSES, WITH THE BEST RESULTS HIGHLIGHTED IN BOLD FONT

Resolution	Majority		Minority		mIoU	mAP
	mIoU	mAP	mIoU	mAP		
64×64	49.27	**66.86**	11.14	23.80	32.93	48.40
32×32	**50.50**	64.98	**11.89**	**27.47**	**33.95**	**50.52**
16×16	39.72	60.49	4.62	10.82	24.68	39.20

TABLE III

THE RESULTS (%) OF THE ABLATION STUDY ON THE VARIOUS CONDITION FUSION METHODS. WE ASSESS THE PERFORMANCE OF THREE COMMONLY USED FUSION METHODS IN THIS EXPERIMENT. 'ADD', 'ATT', AND 'CAT' IN THE TABLE ARE ABBREVIATIONS FOR ELEMENT-WISE ADDITION, ATTENTION-BASED, AND FEATURE CONCATENATION METHODS, RESPECTIVELY. THE OUTCOMES ARE MEASURED USING MIOU AND MAP FOR BOTH MAJORITY AND MINORITY CLASSES, WITH THE BEST RESULTS HIGHLIGHTED IN BOLD FONT

Fusion Methods	Majority		Minority		mIoU	mAP
	mIoU	mAP	mIoU	mAP		
add	44.86	59.73	9.91	27.71	29.88	46.01
att	49.89	**68.75**	11.63	**28.60**	33.49	**51.55**
cat	**50.50**	67.80	**11.89**	27.47	**33.95**	50.52

small and large time steps degrade the generation performance of the proposed network. The optimal performance is achieved when the diffusion process is divided into $1,000$ steps. One possible explanation for this is that the small number of time steps leads to less gradual noise addition, causing turbulence during the distribution transition (from the original data distribution to a Gaussian one or vice versa). Conversely, using larger time steps requires more thorough training and increases the duration of the reverse process. To trade off the network accuracy and time efficiency, we opt for $1,000$ as the time step.

2) Ablation on the Latent Feature: To increase the sampling speed in the reverse denoising process, we implement the diffusion model in the latent space rather than the pixel space. Using a feature compressor, the semantic BEV label is reduced from the resolution of 150×150 to a smaller size. It is crucial to determine an appropriate latent feature size that retains the primary components of the semantic BEV labels. In this ablation study, we compare the performance of the diffusion process with different latent feature sizes. The various latent feature sizes can be obtained by adjusting the number of downsampling layers in the feature compressor. Specifically, the diffusion process is performed on latent features with the resolutions of 64×64, 32×32, and 16×16.

The ablation results are displayed in Tab. II, indicating that the network performs best when the semantic BEV labels are compressed to the resolution of 32×32. From the table, we can infer that the smaller latent features do not provide sufficient semantic information due to the high compression ratio, resulting in inferior performance. However, it is also observed that the outcomes with a bigger latent feature (64×64) are not as good as those with the resolution of 32×32. This result can probably be attributed to the feature compressor's ability to extract the primary components, effectively summarizing the necessary semantic BEV information as a pre-processing step for the subsequent diffusion module.

3) Ablation on the Condition Fusion: In this work, we aim to generate the semantic BEV map from the given front-view input. However, the diffusion process, which directly samples random data from the learned data distribution, fails to establish the corresponding relationship between the BEV and the front view. To address this, the front-view image is employed as the condition to guide the diffusion learning and sampling. In our network, the front-view feature is fused with the latent semantic BEV feature before the forward diffusion process. The commonly used fusing methods include feature addition, concatenation, and attention mechanisms. These fusion methods could influence the effectiveness of guidance from the front-view image. Therefore, we explore the optimal fusion strategy to provide the most effective guidance information to the diffusion process.

TABLE IV

THE RESULTS (%) OF THE ABLATION STUDY ON THE LOSS FUNCTION USED IN THE TRAINING. WE ASSESS THE PERFORMANCE OF THE COMMONLY USED LOSS FUNCTION FOR THE CLASS IMBALANCED PROBLEM IN THIS EXPERIMENT. 'CE', 'FOCAL', 'DICE', AND 'BA' IN THE TABLE ARE ABBREVIATIONS FOR CROSS-ENTROPY LOSS, FOCAL LOSS, DICE LOSS, AND BOUNDARY-AWARE LOSS, RESPECTIVELY. THE OUTCOMES ARE MEASURED USING MIOU AND MAP FOR BOTH MAJORITY AND MINORITY CLASSES, WITH THE BEST RESULTS HIGHLIGHTED IN BOLD FONT

Loss Function	Majority		Minority		mIoU	mAP
	mIoU	mAP	mIoU	mAP		
CE	48.95	65.57	10.55	16.35	32.49	44.48
Focal	50.41	67.43	10.88	26.84	33.47	50.03
Dice	49.27	66.86	11.14	23.80	32.93	48.40
BA	**50.50**	**67.80**	**11.89**	**27.47**	**33.95**	**50.52**

The results can be found in Tab. III, according to which, the attention-based fusion method performs well in terms of mAP, while the direct concatenation counterpart achieves the highest mIoU. The inferior performance of the element-wise additional may be attributed to the corruption of both the latent BEV feature and the front-view condition. To avoid introducing additional learning parameters during the training phase, we utilize the direct concatenation method to fuse the conditional and latent features.

TABLE V

THE COMPARATIVE RESULTS (%) COMPARED WITH THE STATE-OF-THE-ART METHODS. THIS STUDY AIMS TO IMPROVE THE SEGMENTATION PERFORMANCE OF THE MINORITY CLASSES. PARTICULAR ATTENTION SHOULD BE GIVEN TO MINORITY OBJECTS, INCLUDING BARRIERS, VEHICLES, AND PEDESTRIANS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT

Methods	Background		Drivable Area		Ped. Crossing		WalkWay		Barrier		Vehicle		Pedestrain		mIoU	mAP
	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP		
CVT	63.00	73.91	70.88	84.32	23.19	44.70	39.73	55.43	8.27	33.84	18.02	39.77	0.00	0.00	31.87	47.42
VED	56.32	71.99	67.36	74.40	27.31	59.74	33.57	60.22	0.00	0.00	3.60	**53.80**	0.00	0.00	26.88	45.74
MonoLayout	53.84	64.43	62.58	78.15	2.82	16.97	26.99	43.88	0.82	1.58	8.60	23.56	0.00	0.00	22.13	32.65
LSS	56.65	73.30	67.88	78.36	33.17	51.29	35.30	52.29	14.88	43.90	12.67	35.66	0.27	1.08	31.54	47.98
VPN	**66.58**	76.43	73.21	**84.38**	26.78	60.49	**44.32**	**63.56**	3.33	28.25	16.07	39.49	0.00	0.00	32.18	49.78
BEVFormer	50.73	66.72	63.45	74.05	17.11	43.39	26.50	45.23	5.50	21.70	14.17	37.56	0.02	0.20	25.35	41.26
BEVDepth	63.85	**77.87**	**75.53**	81.48	**36.71**	**63.68**	23.90	47.38	0.00	0.00	0.41	16.74	0.00	0.00	28.63	47.86
MatrixVT	54.07	61.37	63.71	80.20	24.67	57.41	27.67	48.33	6.76	33.12	6.13	33.67	0.00	0.00	26.14	44.87
Ours-pre	58.44	69.85	67.75	81.96	36.64	60.19	35.97	54.23	16.12	**54.57**	18.21	27.47	**0.76**	**3.21**	32.84	50.21
Ours	60.63	76.27	71.99	82.92	27.79	55.55	40.34	56.47	**16.24**	44.38	**20.19**	36.21	0.49	1.83	**33.95**	**50.52**

4) Ablation on the Boundary Intensity: The boundary-aware loss increases the weight assigned to minority objects during loss calculation, thereby addressing the class imbalance issue. As outlined in Eq. 14, a boundary intensity factor, denoted as γ, is utilized to modulate the extent to which the edge information of minority objects influences the training process of the network. An increase in γ enables the network to focus more on minority objects when computing the distance between the predicted output and the ground truth. To determine an optimal value for γ, we evaluate the network accuracy for γ values within the set of $[0, 0.5, 0.75, 1.0, 1.5, 3.0]$.

The results are presented in Fig. 4, where the performance of the majority and minority classes is reported separately. The x-axis represents different γ values. The upper y-axis denotes the scale of mIoU. The lower y-axis indicates mAP. Note that certain areas of the y-axis have been magnified to better view the differences. The bar charts illustrate that changes in the γ predominantly affect the performance of the minority classes. The fluctuations in most classes are not significant. During network training, we set γ to 1.5, as this value leads to the highest mIoU and a competitive mAP for both minority classes and the overall performance.

5) Ablation on the Loss Function: One approach to addressing the challenge of class imbalance involves the development of specialized loss functions. Widely used loss functions like Focal Loss and Dice Loss have proven effective in alleviating this problem. Accordingly, we compared the network utilizing these loss functions with the proposed boundary-aware loss.

The experimental results are listed in Tab. IV. The Cross-Entropy loss is introduced as the baseline for comparison. The results demonstrate that the network incorporating the proposed boundary-aware loss outperforms those employing other loss functions. This outcome highlights the effectiveness of the boundary-aware loss in balancing the network's attention between majority and minority classes.

C. Comparative Experiment

A comparative experiment is conducted between the proposed network and the existing BEV methods, including semantic BEV map generation networks and BEV detection networks. The semantic BEV map generation networks include Variational Encoder-Decoder Networks (VED) [39], MonoLayout [40], Lift split shot network (LSS) [41], View Parsing Network (VPN) [58] and Cross-view Transformation (CVT) [43]. The BEV detection networks include BEVFormer [12], BEVdepth [42], and MatrixVT [59]. To achieve comparison, the detection heads of the BEV detection networks are replaced with the semantic heads, while the original structures of the semantic BEV map generation networks are kept unchanged. We further assess the approach introduced in our conference paper [20], which utilizes an LSS-based model integrated with boundary-aware loss, to illustrate the proposed network's capability in enhancing segmentation accuracy for both majority and minority classes in a balanced fashion. This baseline method is referred to as Ours-pre. This research focuses on enhancing the accuracy of minority object detection. Accordingly, specific emphasis should be placed on analyzing the segmentation outcomes for minority classes, including barriers, vehicles, and pedestrians, within the framework of this comparative evaluation.

1) The Quantitative Results: The comparative results of the aforementioned networks are presented in Tab. V. For a more detailed comparison, results are reported for each semantic class. The first four classes represent the majority classes, while the remaining classes belong to the minority classes. Overall, our proposed network achieves the highest performance in both mIoU and mAP across all classes. VPN [58] and BEVDepth [42] perform better in generating semantic BEV maps for the majority classes, but our network performance is closely competitive. The slightly reduced detection performance for majority classes may be attributed to the enhanced detection of minority classes. In comparison, existing methods often misclassify minority classes as majority classes, resulting in relatively higher performance metrics for majority classes. Conversely, the proposed network leverages a generative model to capture the underlying data distribution, which generally necessitates larger datasets and extended training durations to achieve convergence compared to the discriminative models employed by the comparison

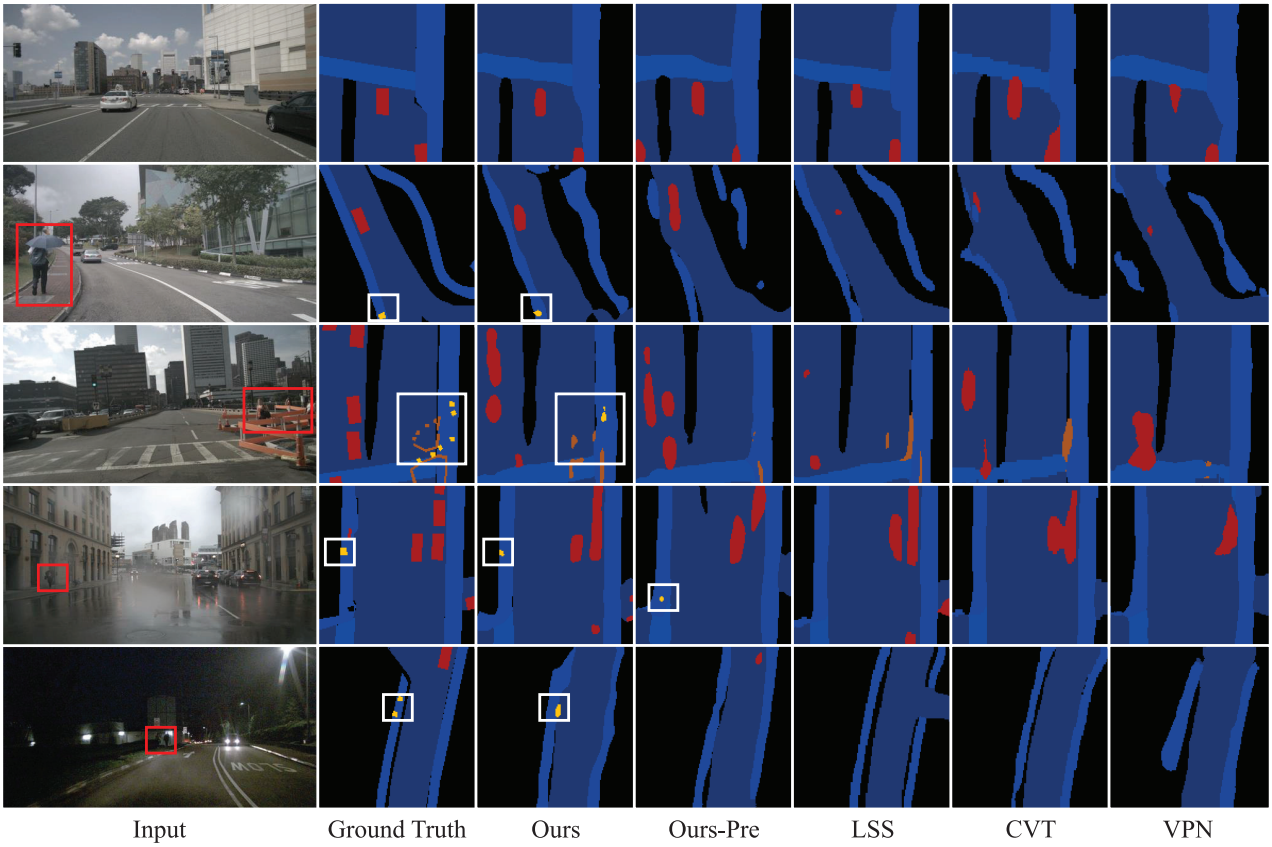

Fig. 5. Sample qualitative demonstrations for the semantic BEV map generation networks. Each row presents the outcomes from different networks tested with identical inputs. The minority objects are marked with bounding boxes to attract attention. Due to space constraints, only the outputs from the top-performing networks are displayed. The results demonstrate the superiority of our network. The figure is best viewed in color.

methods. Since all networks were trained under identical experimental conditions, the generative model's performance may not be as fully optimized as the other methods. However, our network exhibits a notable improvement in segmenting minority classes. In comparison to our earlier work, denoted as Ours-pre, the current model enhances the segmentation performance for minority classes while preserving competitive accuracy for majority classes.

2) Qualitative Demonstrations: Some sampled qualitative demonstrations are shown in Fig. 5. Generally, compared to the other networks, our network generates the most accurate and clear semantic BEV maps, especially for minority objects like pedestrians. The third and bottom rows illustrate driving scenarios in rainy weather or under low-light conditions, where the proposed network can also successfully detect pedestrians. However, in cases where objects are occluded, as depicted in the second row with pedestrians positioned behind barriers, our network fails to detect them. Also at night, the headlights of the vehicle impede the correct detection.

V. Conclusion and Future Work

We proposed here a diffusion-guided semantic BEV map generation network with a boundary-aware loss, which aims to address the class imbalance issue. Within our network, a conditional diffusion module is utilized in the latent space to learn the underlying data distribution for both majority and minority classes. The BEV feature sampled from the learned distribution then initializes the BEV query in the cross-attention module, providing prior knowledge about data distribution. Additionally, a boundary-aware loss is designed to balance the network's attention toward the minority classes during training. This loss function increases the weight of the minority classes in the loss calculation to guide the network focus on the segmentation of those objects. Extensive experimental results confirm the effectiveness of our designed network and its superiority over the state of the arts, particularly in segmenting minority classes.

In this work, we observed that complex driving environments, such as rainy and low-lighting conditions, could degrade the network performance. Exploring methods to adapt the network's capabilities from standard scenarios to these specialized conditions represents a promising research direction. Insights from domain shift [60] and knowledge distillation [61] could provide valuable guidance in addressing this issue. Furthermore, future work will investigate the integration of contrastive learning [62], [63], [64], [65] to enhance the effectiveness of addressing class imbalance challenges. Building on advancements in inference acceleration methods [66], further research on optimizing the inference efficiency of diffusion-based techniques is also a critical area to explore.

REFERENCES

[1] D. Li, Q. Zhang, S. Lu, Y. Pan, and D. Zhao, "Conditional goal-oriented trajectory prediction for interacting vehicles," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 18758–18770, Dec. 2024.

[2] X. Zhang, H. Yu, Y. Qin, X. Zhou, and S. Chan, "Video-based multi-camera vehicle tracking via appearance-parsing spatio-temporal trajectory matching network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 10077–10091, Oct. 2024.

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

[4] H. Thanh Le, S. L. Phung, and A. Bouzerdoum, "Bayesian Gabor network with uncertainty estimation for pedestrian lane detection in assistive navigation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5331–5345, Aug. 2022.

[5] X. Shao, L. Zhang, T. Zhang, Y. Shen, and Y. Zhou, "MOFISSLAM: A multi-object semantic SLAM system with front-view, inertial, and surround-view sensors for indoor parking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4788–4803, Jul. 2022.

[6] Z. Ma, Z. Zheng, J. Wei, Y. Yang, and H. T. Shen, "Instance-dictionary learning for open-world object detection in autonomous driving scenarios," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3395–3408, May 2024.

[7] C. Tao, J. Cao, C. Wang, Z. Zhang, and Z. Gao, "Pseudo-mono for monocular 3D object detection in autonomous driving," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3962–3975, Aug. 2023.

[8] Y. Pan, F. Xie, and H. Zhao, "Understanding the challenges when 3D semantic segmentation faces class imbalanced and OOD data," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 7, pp. 6955–6970, Jul. 2023.

[9] H. Caesar et al., "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[10] K. Chitta, A. Prakash, and A. Geiger, "NEAT: Neural attention fields for end-to-end autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15773–15783.

[11] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, vol. 164, 2022, pp. 180–191.

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

[13] M. Saini and S. Susan, "Deep transfer with minority data augmentation for imbalanced breast cancer dataset," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106759.

[14] Y. Shi et al., "Improving imbalanced learning by pre-finetuning with data augmentation," in *Proc. 4th Int. Workshop Learn. Imbalanced Domains, Theory Appl.*, 2022, pp. 68–82.

[15] M. Temraz and M. T. Keane, "Solving the class imbalance problem using a counterfactual method for data augmentation," *Mach. Learn. Appl.*, vol. 9, Sep. 2022, Art. no. 100375.

[16] J. Li, Q. Zhu, Q. Wu, and Z. Fan, "A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors," *Inf. Sci.*, vol. 565, pp. 438–455, Jul. 2021.

[17] Y. Xie, M. Qiu, H. Zhang, L. Peng, and Z. Chen, "Gaussian distribution based oversampling for imbalanced data classification," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, pp. 667–679, Feb. 2022.

[18] K. Cao, C. Wei, A. Gaidon, N. Aréchiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Jan. 2019.

[19] S. R. Bulo, G. Neuhold, and P. Kontschieder, "Loss max-pooling for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7082–7091.

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

[21] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NIPS*, vol. 33, 2020, pp. 6840–6851.

[22] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, Jul. 2019.

[23] Z. Feng, Y. Guo, D. Navarro-Alarcon, Y. Lyu, and Y. Sun, "InconSeg: Residual-guided fusion with inconsistent multi-modal data for negative and positive road obstacles segmentation," *IEEE Robot. Autom. Lett.*, vol. 8, no. 8, pp. 4871–4878, Aug. 2023.

[24] Z. Feng, Y. Guo, and Y. Sun, "CEKD: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images," *IEEE Robot. Autom. Lett.*, vol. 8, no. 4, pp. 2205–2212, Apr. 2023.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[26] Z. Zhou et al., "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[27] H. Huang et al., "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

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

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

[31] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2022.

[32] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

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

[34] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[35] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[36] S. Li et al., "Matching anything by segmenting anything," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 18963–18973.

[37] D. Yuan et al., "An attention mechanism based AVOD network for 3D vehicle detection," *IEEE Trans. Intell. Vehicles*, early access, Oct. 12, 2023, doi: 10.1109/TIV.2023.3323960.

[38] D. Yuan, X. Chang, Z. Li, and Z. He, "Learning adaptive spatial–temporal context-aware correlation filters for UAV tracking," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 18, no. 3, pp. 1–18, 2022.

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

[40] K. Mani, S. Daga, S. Garg, N. S. Shankar, K. M. Jatavallabhula, and K. M. Krishna, "MonoLayout: Amodal scene layout from a single image," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Feb. 2020, pp. 1689–1697.

[41] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. 16th Eur. Conf. Comput. Vis.-ECCV*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 194–210.

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

[43] W. Yang et al., "Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15536–15545.

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

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

[46] S. H. Ebenuwa, M. S. Sharif, M. Alazab, and A. Al-Nemrat, "Variance ranking attributes selection techniques for binary classification problem in imbalance data," *IEEE Access*, vol. 7, pp. 24649–24666, 2019.

[47] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[48] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, M. Meila and T. Zhang, Eds., Jul. 2021, pp. 8162–8171.
[49] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.
[50] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. NIPS*, vol. 34, 2021, pp. 8780–8794.
[51] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.
[52] X. Du, Y. Sun, X. Zhu, and Y. Li, "Dream the impossible: Outlier imagination with diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, Jan. 2023.
[53] F. Fan et al., "Hierarchical masked 3D diffusion model for video outpainting," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7890–7900.
[54] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.
[55] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[56] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[57] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[58] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4867–4873, Jul. 2020.
[59] H. Zhou, Z. Ge, Z. Li, and X. Zhang, "MatrixVT: Efficient multi-camera to BEV transformation for 3D perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 8548–8557.
[60] M. Li et al., "A benchmark for cycling close pass detection from video streams," 2023, *arXiv:2304.11868*.
[61] Z. Li et al., "When object detection meets knowledge distillation: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10555–10579, Aug. 2023.
[62] M. Li, P. Huang, X. Chang, J. Hu, Y. Yang, and A. Hauptmann, "Video pivoting unsupervised multi-modal machine translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3918–3932, Mar. 2022.
[63] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest X-ray report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3334–3343.
[64] M. Li et al., "Contrastive learning with counterfactual explanations for radiology report generation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2024, pp. 162–180.
[65] R. Liu, M. Li, S. Zhao, L. Chen, X. Chang, and L. Yao, "In-context learning for zero-shot medical report generation," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 8721–8730.
[66] C. Li, G. Wang, B. Wang, X. Liang, Z. Li, and X. Chang, "DS-Net++: Dynamic weight slicing for efficient inference in CNNs and vision transformers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4430–4446, Apr. 2023.

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

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

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

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

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

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

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