Bird's Eye View Semantic Segmentation based on Improved Transformer for Automatic Annotation

Tianjiao Liang[1,2], Weiguo Pan[1,2*], Hong Bao[1,2*], Xinyue Fan[1,2], and Han Li[1,2]
[1] Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China
[2] College of Robotics, Beijing Union University, Beijing 100101, China
[e-mail : ldtweiguo@buu.edu.cn, baohong@buu.edu.cn]
[*]Corresponding author: Weiguo Pan, Hong Bao

Received April 12, 2023; revised May 29, 2023; accepted July 14, 2023;
published August 31, 2023

Abstract

High-definition (HD) maps can provide precise road information that enables an autonomous driving system to effectively navigate a vehicle. Recent research has focused on leveraging semantic segmentation to achieve automatic annotation of HD maps. However, the existing methods suffer from low recognition accuracy in automatic driving scenarios, leading to inefficient annotation processes. In this paper, we propose a novel semantic segmentation method for automatic HD map annotation. Our approach introduces a new encoder, known as the convolutional transformer hybrid encoder, to enhance the model's feature extraction capabilities. Additionally, we propose a multi-level fusion module that enables the model to aggregate different levels of detail and semantic information. Furthermore, we present a novel decoupled boundary joint decoder to improve the model's ability to handle the boundary between categories. To evaluate our method, we conducted experiments using the Bird's Eye View point cloud images dataset and Cityscapes dataset. Comparative analysis against state-of-the-art methods demonstrates that our model achieves the highest performance. Specifically, our model achieves an mIoU of 56.26%, surpassing the results of SegFormer with an mIoU of 1.47%. This innovative promises to significantly enhance the efficiency of HD map automatic annotation.

Keywords: Transformer; Semantic Segmentation; High-Definition maps; Automatic Annotation.

.

http://doi.org/10.3837/tiis.2023.08.002 **ISSN : 1976-7277**

1. Introduction

Autonomous driving systems enables vehicles to automatically navigate open roads, performing specific tasks such as robotaxi transport [1] and unmanned shipping container transport [2]. This area of research has gained significant popularity [3,4]. The acquisition of road information, particularly map information, is crucial for the smooth operation of autonomous driving system. These maps play a vital role in distinguishing between road areas and non-road areas, ensuring vehicles remain in safe zones and preventing potential accidents. Early studies relied on in-vehicle sensors like LiDAR and cameras to detect real-time road boundaries[5]. However, accurately capturing the diverse features of roads in real time poses significant challenges. Road boundaries are often narrow elongated, and irregular, making it difficult to define common features. Moreover, road boundaries are frequently obscured in real-world road scenes, greatly impacting the perception capabilities of autonomous driving systems. High-definition (HD) maps, a type of geographic information system, have become an essential component of autonomous driving systems. These maps encompass lane lines, road edges, zebra crossings, no-stopping areas, diversion areas, and other surface elements, offering precise geometric and semantic information about the static traffic environment. HD maps are manually labeled from bird's-eye-view (BEV) images, which include high-resolution aerial images, overhead images from pre-built point-cloud maps, and front view images captured by CMOS sensors positioned on the front of the vehicle. As autonomous driving and map labeling continue to advance, point cloud data collected by a LiDAR system equipped with a mobile map system (MMS) undergoes preprocessing to generate BEV point cloud images. These images possess clearer clearer map elements and lower absolute error compared to front view images, making them suitable for certain scenarios such as vehicle occlusion scenes and HD map annotation. In this paper, we propose an automatically annotate method for HD maps utilizing BEV point cloud images.

The challenge of HD map annotation primarily resides in element annotation. The task of marking road elements in urban areas for HD maps demands a substantial workforce. It entails labor-intensive efforts, resulting in low production efficiency and high production cost. However, thanks to the rapid advancements in deep learning over past years, automatic HD map annotation based on deep learning has shown remarkable progress, presenting the potential to enhance map production efficiency and elevate the automation rate. Automatic map annotation refers to the utilization of artificial intelligence techniques to automatically detect diverse elements in HD maps.

Currently, there is a dearth of research on the specific issue of automatically annotating the HD maps from BEV point cloud images. Existing studies predominantly concentrate on particular tasks such as road lane detection [6–8] or road grid detection [9–12]. These investigations can be classified into two primary categories: iterative graph growth methods [13] and segmentation-based methods. Although iterative graph growth methods ensures topology accuracy, they suffer from drawbacks such as low efficiency, limited parallelism, and drift, leading to inadequate precision. To overcome these challenges, we employ semantic segmentation for automated element labeling. Semantic segmentation is a fundamental computer vision task and holds significant important in the production of HD maps for autonomous driving. Differing from iterative graph growth methods, semantic segmentation operates at the pixel-level level, exhibiting characteristics such as high accuracy and efficient production.

Currently, the conventional approach to semantic segmentation is based on convolutional neural networks (CNNs) such as FCN [14], DeepLab [15–18], and U-Net [19]. These networks utilize a CNN to extract features from the input sample and subsequently restore the feature map size through upsampling. This enables pixel-level classification in an end-to-end manner.

However, a limitation of the CNN-based semantic segmentation method is relatively small and localized effective receptive field of feature map. Consequently the feature extraction ability is restricted, preventing a comprehensive consideration of long-distance pixels dependence wihtin the image.

Transformers [20] have gained widespread usage in natural language processing, primarily for their ability to efficiently acquire global information in a parallelized manner. Building upon the success of Transformer design in natural language processing, the Vision Transformer (ViT) [21] was introduced for image classification. Carion et al. [22] developed DETR for object detection, achieving the state-of-the-art performance on public datasets. Transformers have also been explored in semantic segmentation with methods like SETR [23] and SegFormer [24]. While these approaches have demonstrated impressive results, they do have certain limitations. Firstly, SETR, based on ViT, relies on single-scale feature maps for predictions instead of leveraging multi-scale information. Secondly, SegFormer adopts the main Transformer encoder design but utilizes a simple ALL-MLP decoder for feature decoding.

In this paper, we propose a novel semantic segmentation method called Mapformer for automatic HD map annotation, aiming to perceive map elements efficiently. Our proposed network takes a BEV point cloud image generated by the MMS as input and directly segment map features, including line and area features, from that image. The contributions of this work are summarized as follows:

(1)We propose a new encoder, the convolutional transformer hybrid encoder, to enhance the model’s feature extraction capability.

(2)We propose a multi-level fusion module that enables the model to aggregate diverse levels of detail and semantic information.

(3)We propose a novel decoupled boundary joint decoder that enhances the model's ability to handle category boundaries effectively.

2. Related Work

HD map annotation plays a crucial role in the rapidly advancing field of autonomous driving, with extensive research focusing on various tasks such as object detection [25,26], image classification [27], and semantic segmentation [28]. HD maps encompass essential static features of the road environment necessary for autonomous driving, including roads, buildings, traffic lights, and road markings. They also include semantic objects that may be occluded and therefore not directly detectable by sensors. In recent years, HD maps for autonomous driving have gained prominence due to their exceptional accuracy and rich geometric semantic information. Automatic HD map annotation relies on three primary data sources, which are outlined below.

Map annotation based on 2D aerial imagery involves extracting road marking from aerial images, allowing for the efficient extraction and storage of large-scale road markings in HD maps , thereby reducing detection time [29]. However, this approach is highly susceptible to data defects caused by factors such as lighting conditions, occlusion, and worn road markings. While traditional methods have demonstrated significant success in extracting road markings from images of sidewalks or concrete roads, mere extraction without correctly identifying the various types of road markings falls short in enabling vehicles to comprehend the rules of the road. Thanks to the rapid advancements in CNNs, CNN-based methods have emerged and gained widespread use in the detection and identification of road markings [30–32].

Map annotation based on 3D point cloud extraction involves utilizing LiDAR point cloud data, typically employing two annotation methods: the bottom-up method [33–35] and top-down method [36,37]. In the bottom-up approach, road markings are directly extracted from the original data by segmenting road markings and backgrounds, relying on detection and location. This annotation approach is efficient but sensitive to noise present in the raw data. On the other hand, the top-down method initially detects predefined geometric models and subsequently reconstructs road markings based on the detection results. While this method is less affected by noise in the original data, it is more time-consuming due to the extensive search space of the model.

Annotation methods based on 3D point cloud data primarily target stereoscopic objects in traffic scenes, including traffic lights, traffic signs, streetlights, trees, and poles. These methods are instrumental in object localization and motion planning.

Map annotation based on in-vehicle vision sensors: These methods [38,39] typically leverage image data captured by the vehicle's front view camera or a combination of cameras capturing the front, side, and rear views. Such methods enable the detection of the surrounding traffic conditions [40]. However, these approaches suffer from low recognition accuracy and are unable to address the issue of object occlusion. As a result, they are commonly employed for real-time in-car maps generation, bypassing the need for offline maps.

In comparison to aerial images, the front view image has a narrower field of view, resulting in longer detection and processing times for road marker extraction. However, due to real-time image acquisition, this approach offers greater flexibility in handing changes to road markings, including wear and occlusion.

Utilizing MLS (Mobile Laser Scanning)3D point clouds for feature extraction offers a high-performance approach to enhance HD maps with detailed road information. HD maps enriched with extracted 3D features provide depth information and up-to-date environmental data. This paper primarily focuses on ground information rather than pole-like targets, proposing an automatic HD map annotation method that converts MLS 3D point clouds into 2D georeferenced grayscale images, known as BEV point cloud maps. This conversion leverages the high accuracy and low error of 3D point clouds, significantly reducing computational complexity while preserving semantic information.

Semantic segmentation: Semantic segmentation involves pixel level image classification. FCN is a fundamental network architecture designed specifically for semantic segmentation. DeepLab v1 utilizes dilated convolutions to enlarge the receptive field while retaining resolution and edge information. Additionally, it employs conditional random fields [41] to refine boundary details. U-Net emphasizes feature fusion in an FCN, employing a shallow network structure to preserve low-level visual information. However, it may not perform optimally in semantic segmentation dataset with complex classification tasks. Seg-Net improves the upsampling performance by utilizing convolution with trainable decoder filters.

Dur to the inherent limitations of CNNs, the aforementioned network models struggle to capture global features of the input samples. Consequently, much of the existing research in this field is dedicated to enhancing feature extraction capabilities by expanding the receptive field of the feature map, refining edge information, and devising novel network architectures.

Thanks to its capability to integrate global information during the feature extraction stage, the Transformer emerged as the first to surpass RNNs [42] and LSTMs [43] in machine translation tasks. The ViT, initially introduced by Dosovitskiy et al., made its debut in image classification by treating images as a sequence of feature tokens and feeding them into Transformer modules. Network models based on the Transformer architecture have demonstrated state-of-the-art performance across various computer vision tasks, finding applications in fields such as autonomous driving, time series forecasting, and medical image processing. SETR utilizes the ViT as the encoder for feature extraction, leading to improve semantic segmentation performance. TransUNet [44], on the other hand, leverages the Transformer to enhance U-Net's ability for long-distance context modeling, resulting in promising outcomes in medical image segmentation. SegFormer has achieved state-of-the-art performance on numerous semantic segmentation datasets by enhancing the positional embedding coding, reducing time complexity, and preserving local continuity. However, it should be noted that SegFormer employs a simple ALL-MLP decoder for feature decoding, which may not capture fine-gained segmentation details effectively.

3 Proposed Method

The overall pipeline of our proposed method isdepicted in **Fig. 1**. The main contribution of the proposed method is the following three key components: (1) the convolutional transformer hybrid encoder, (2) the multi-level fusion module, and (3) decoupled boundary joint decoder.

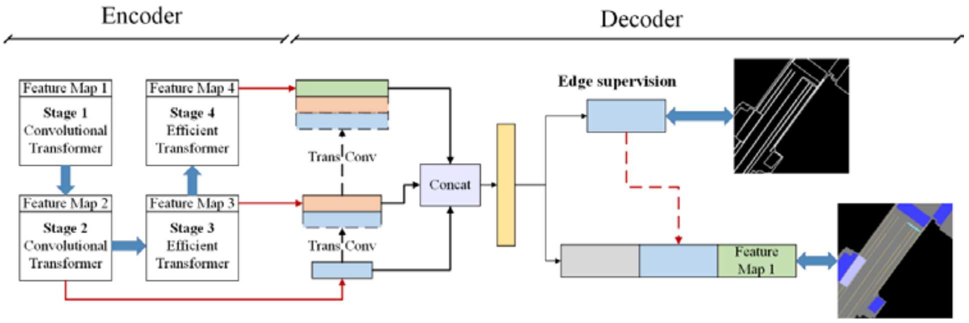
Encoder
Decoder
Feature Map 1
Stage 1 Convolutional Transformer
Feature Map 2
Stage 2 Convolutional Transformer
Feature Map 4
Stage 4 Efficient Transformer
Feature Map 3
Stage 3 Efficient Transformer
Trans Conv
Trans Conv
Concat
Edge supervision
Feature Map 1

Fig. 1. Overall architecture of the proposed method. The encoder is used to extract image features, and the decoder is responsible for segmenting the image. Note that the level of detail segmentation of the model on the image is one of the indicators for evaluating the performance of the model.

3.1. Convolutional transformer hybrid encoder

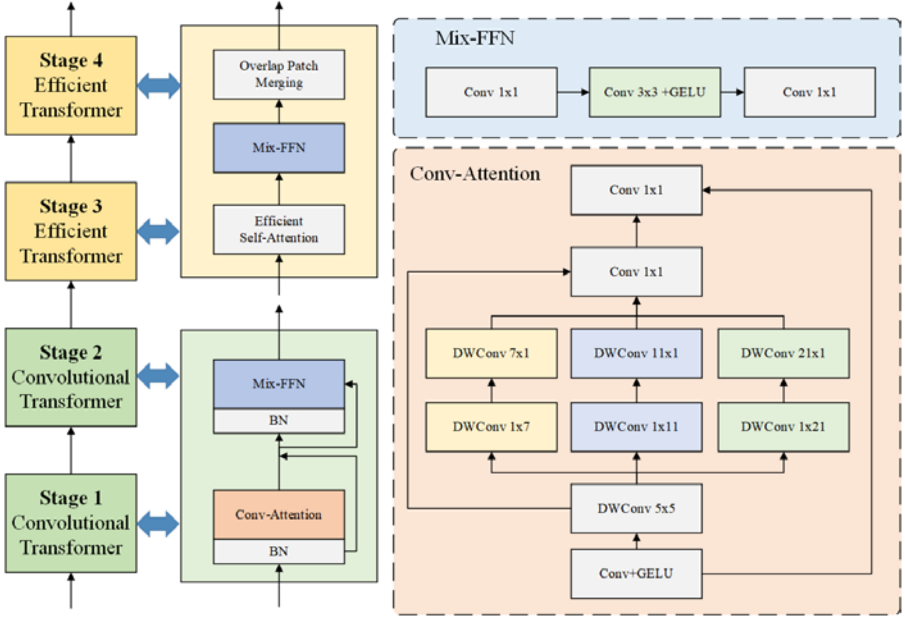

Fig. 2. Structure of the convolutional transformer hybrid encoder. The encoder consists of a convolutional transformer part to strengthen the model's ability to extract line features and an efficient transformer part to reduce computational complexity.

We propose an encoder that consists of convolutional transformer modules and efficient transformer modules, as illustrated in **Fig. 2**. The encoder comprises two modules: an efficient transformer and a convolutional transformer, each serving different purposes. The convolutional Transformer is employed in the initial two stages (stage 1 and stage 2) of the encoder. Unlike the traditional Transformer structure, the convolutional Transformer replaces self-attention with a multi-scale convolutional attention module. This module consists of multi-branch depth-wise strip convolutions and employs a 1 × 1 convolution for feature alignment. The effectiveness of convolutional modules in Transformer structures has been demonstrated by SegNeXt [45]. In our proposed method, certain strip-like objects, such as lane lines and curbs, are present in the HD maps. Hence, strip convolution can be employed to extract shallow feature maps that encompass rich in image detail, facilitating the extraction of strip-like features. The feature maps undergo strip convolution operations with varying kernel sizes (e.g., 7×1, 11×1, and 21×1), followed by a 1×1 convolution to fuse the features. The specific operation can be summarized as follows:

$$\alpha = Conv_{1\times 1}(\sum_{i} DWConv_{i}(F)) \tag{1}$$

$$\tilde{F} = \alpha \otimes F \tag{2}$$

Here, α is the attention weight, F is the feature map, $\tilde{F}$ is the result of α multiplied elementwise with F . $DWConv_i$ represents two DWConv operations with different convolution kernel sizes, for example, $DWConv_1$ denotes 1×7 depth-wise convolution and 7×1 depth-wise convolution.

The efficient transformer is utilized in the final two stages (stage 3 and stage 4) of the encoder to enhance the model's global attention. This module replaces the traditional self-attention module with an efficient self-attention mechanism, aiming to decrease computational complexity while maintaining accuracy. The efficient Transformer employs a scale parameter S to govern the input vector' size. Specifically,

$$\tilde{I} = Reshape(\hat{I})\ from\ HW \times C\ to\ \frac{HW}{S} \times C \bullet S \tag{3}$$

$$I = Linear(\tilde{I})\ from\ \frac{HW}{s} \times C \bullet S\ to\ \frac{HW}{S} \times C \tag{4}$$

Here, $\tilde{I}$ is the original input sample with dimensions $HW \times C$. Moreover, I is the input sample after dimensionality reduction. Its dimensions are $\frac{HW}{S} \times C$. Compared with the time complexity of the traditional self-attention module, the time complexity is reduced from $O(N^2)$ to $O(\frac{N^2}{S})$.

3.2. Multi-level Fusion

We have developed a multi-level fusion module to effectively capture high-level semantics and restore the original image details, which is integrated into the encoder. As depicted in **Fig. 2**, we conducted evaluations on three feature fusion methods. The CNN-based models employ a single structure to capture high-level semantics. Furthermore, SegFormer incorporates an MLP to fuse features from various feature layer outputs, aiming to design a compact decoder. However, this approach compromises the network's capability to extract fine details. To address this issue, we propose the multi-level fusion module, which maximizes the utilization of features.

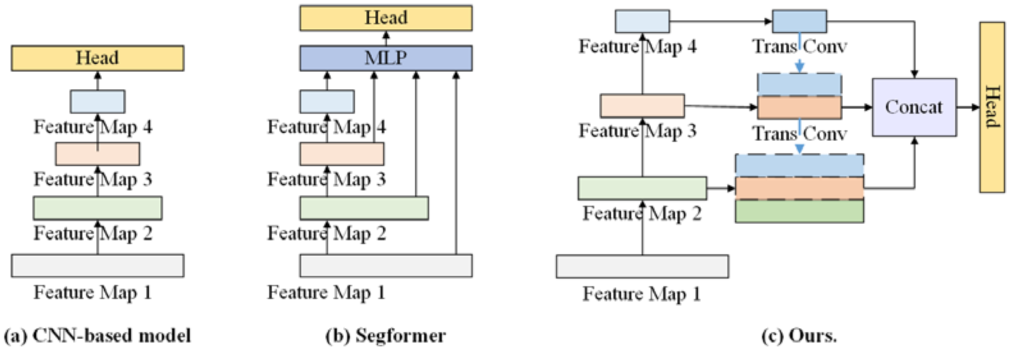

Fig. 3. Three decoder designs. (a) Structure of CNN-based model (b) Structure of Segformer (c) Structure of multi-level fusion Module.

Traditionally, prior studies have commonly employed bilinear interpolation for upsampling. In contrast, we enhance our model’s ability to restore intricate details by replacing the upsampling component with a transposed convolution.

We employ transposing convolution to upsample the feature map from stage 4 of the encoder. This upsampled feature map is then fused with the feature map from stage 3, resulting in a new feature map for stage 3. Similarly, we perform the same operation between stages 3 and 2. Next, the feature map data from these three levels (stage 2 to stage 4) are aligned, concatenated, and fused using an MLP layer. The resulting feature map contains a comprehensive representation of both low-level detail information and high-level semantic information.

3.3. Decoupled Boundary Joint Decoder

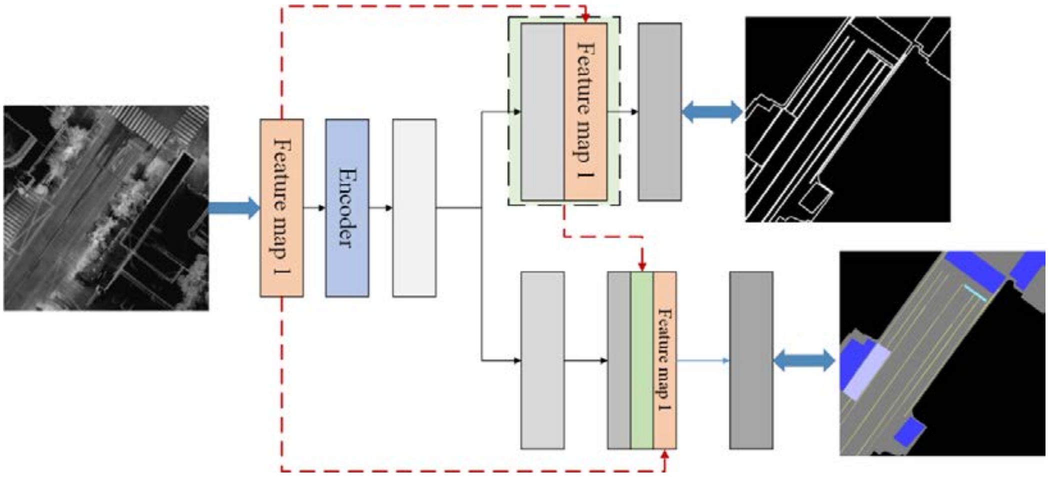

Fig. 4. Structure of the decoupled boundary joint decoder. The decoder adds a boundary supervision branch to improve the model's ability to restore category boundary details.

HD maps encompass a significant quantity of strip-like objects, and current segmentation networks prove inadequate in accurately segmenting fine image details. These networks often exhibit incorrect segmentation of slender objects like lane lines, leading to blurred classification boundary between distinct categories. The downsampling operation in FCNs contributes to imprecise predictions, while SegFormer’s lightweight decoder and bilinear interpolation upsampling hinder detail restoration. Consequently, the segmentation boundaries of the predictions tend to be blurry, thus compromising overall performance. To address these challenges, we introduce the decoupled boundary joint decoder, as illustrated in **Fig. 4**, which effectively resolves the aforementioned issues.

We decouple the high-frequency information from the image as boundary features and incorporate a dedicated branch in the decoder to oversee these boundary features. Subsequently, we generate a combined feature map consisting of the predicted feature map of high-frequency information and the feature map from the main branch. The main branch receives features from feature map 1 as well as the boundary supervision branch. The boundary features are carefully supervised using boundary masks, enabling the model to learn accurate boundary predictions, while the main branch is responsible for outputting comprehensive segmentation results.

We simultaneously supervise both the main and boundary branches. In the boundary module, we predict a boundary map y^l that encompasses all the outlines of the categories. The loss function defined as follows

$$L = \lambda_1 L_{main}(y^m, y^{gt}) + \lambda L_{edge}(y^l, y_e^{gt}) \quad (5)$$

where y^{gt} represents the groundtruth semantic labels and y_e^{gt} represents the groundtruth boundary masks labels. In addition, y^m and y^l denote the segmentation map predictions from the main branch and boundary branch, respectively. Finally, λ_1 and λ are hyperparameters controlling the weights.

3.4. Efficient Data Augmentation

We employed various traditional data augmentation methods, including random resizing, random horizontal flipping, and random cropping. Since BEV point cloud images are grayscale, we utilized CLAHE to equalize the histogram of the image, making the element information more distinct. The enhanced result is depicted in **Fig. 5**.

During the training process, each batch comprises two components, one consists of samples processed as described above, and the other consists of samples after AutoAugmentation.

Furthermore, upon analyzing the dataset and the training model, we identified certain categories that present challenges in learning due to there small areas (special lane lines) or blurred boundaries with other categories (such as double solid lines or one real and one virtual double dashed line). To address this, we curated a set of challenging cases and introduced a 10% resampling probability during training to include these data points in the training set.

The experiments demonstrate effectiveness of our data augmentation method without introducing additional time complexity.

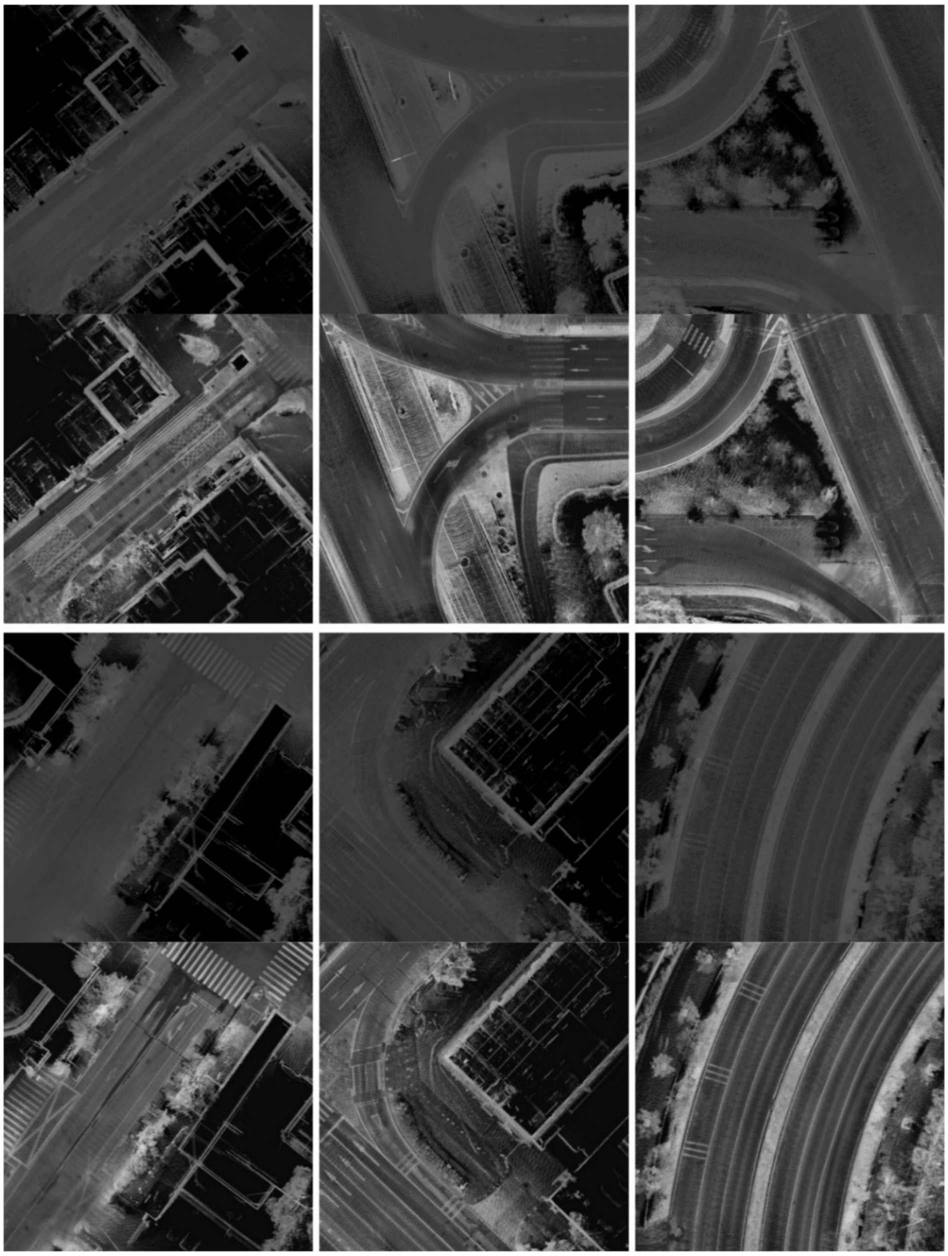

Fig. 5. Efficient data augmentation. The element information is clearer.

4. Expriments

4.1. Datasets

The BEV point cloud image dataset consists of 10,000 single-channel grayscale images with a resolution of 1536×1536. These images were derived from preprocessed point cloud data collected by a LiDAR system equipped with a MMS. The generation process of BEV point cloud images is illustrated in **Fig. 6**. The dataset encompasses a total of 16 categories, which are further classified into the following classes: invalid, pavement, non-road, single solid line, double solid line, single dashed line, double dashed line, fence double line, one real and one virtual, special lane line, horizontal marking, other line, green belt in the road, no-stopping area, zebra crossing, and diversion area. For test set, we randomly selected 800 images. Additionally, to assess the model's generalization, we conducted a comparative evaluation on the publicly available Cityscapes dataset [46].

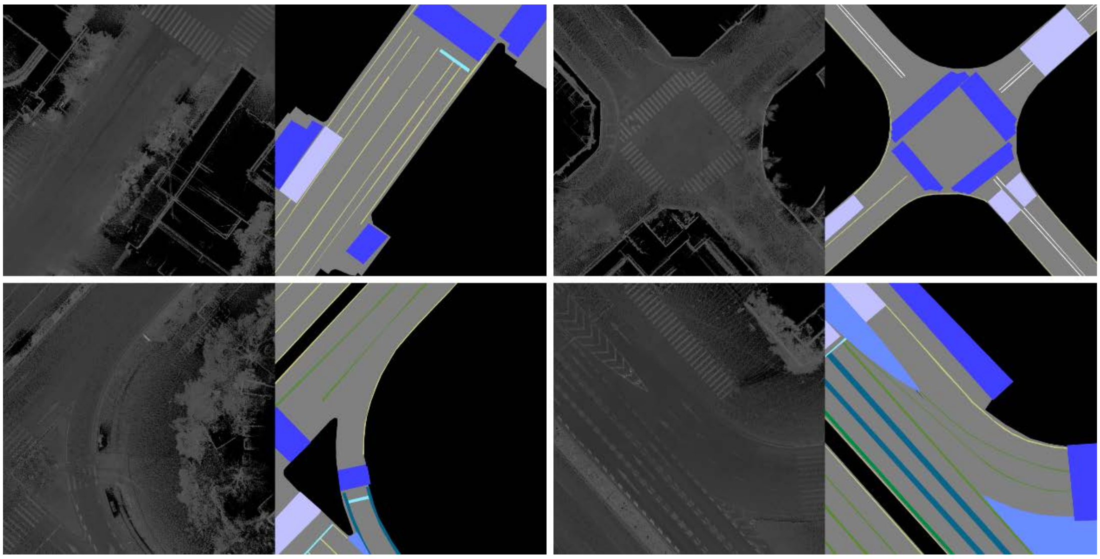

Fig. 6. Our BEV point cloud images dataset. (Left) Original images and (right) the groundtruth.

4.2. Implementation details

We utilized PyTorch and PaddleSeg [47] to implement and train our model using four Tesla T4s. The convolutional transformer hybrid encoder and multi-level fusion components were pre-trained on the ImageNet-1K dataset, while parameter initialization in the decoupled boundary joint decoder was performed using Xavier initialization. The AdamW [48] optimizer was employed, with an initial learning rate of 6×10−5 and a polynomial decay learning rate scheduler. The images are resized to 1024×1024 while preserving their aspect ratio. All models underwent training for 200,000 iterations.

4.3. Evaluation metrics

For quantitative evaluation, we measured the accuracy, Dice coefficient, mIoU, and kappa of the segmentation results and the groundtruth.

Accuracy is the overall classification accuracy.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

The Dice coefficient is used to measure the similarity between the groundtruth and predicted results.

$$Dice = \frac{2TP}{FN + 2TP + FP} \tag{7}$$

The mIoU calculates the coincidence ratio of the intersection of two sets and their union.

$$mIoU = \frac{TP}{FN + TP + FP} \tag{8}$$

The kappa coefficient judges the classification accuracy based on the confusion matrix, which is often used for consistency testing.

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{9}$$

Here, p_o is the number of correctly classified samples in the confusion matrix (values on the diagonal) divided by the number of all samples. In addition, p_e is the sum of the real value of each class in the confusion matrix and the predicted value of each class divided by the square of the total number of samples.

4.4. Ablation Studies

We first evaluated the effectiveness of the various proposed units as presented in **Table 1**. The convolutional transformer hybrid encoder (CTHE), multi-level fusion, decoupled boundary joint decoder (DBJD) and data augmentation modules were incrementally incorporated into SegFormer. This allowed us to evaluate the generalization capability of the proposed method and its detection performance in traffic scenes.

Table 1. Proposed modules evaluated in the ablation study.

Method	Kappa	Dice	mIoU
Segformer	0.8509	0.6691	0.5479
+ CTHE	0.8583	0.6748	0.5516
+ Multi-level Fusion	0.8627	0.6814	0.5547
+ DBJD	0.8642	0.6853	0,5583
+ Data Augmentation	0.8714	0.6902	0.5626

Furthermore, we conducted a comparetive analysis of various state-of-the-art encoders on the ImageNet validation dataset. As illustrated in **Table 2**, we compared the proposed CTHE with CNN-based and Transformer-based classification models, which have recently demonstrated outstanding performance. The results presented in **Table 2** clearly indicate that CTHE outperforms the other models, showcasing its superior performance.

Table 2. Comparison of the proposed method with state-of-the-art encoders on the ImageNet validation set.

Method	Params. (M)	Top-1 Acc. (%)
Vit-B/16	86	77.9
Swin-S[49]	50	83.0
MiT-B3	45	83.1
ConvNeXt-S[50]	45	83.1
CTHE	45	83.5

Table 3 demonstrates the utilization of transposed convolution, nearest neighbor, and bilinear upsampling methods in the multi-level fusion module. The experimental results substantiate that the use using transposed convolution yields superior results, suggesting its effectiveness in restoring the details of the original image.

Table 3. Upsampling ablation study.

Method	mIoU	Δ (%)
Ours. (trans-conv)	0.5626	-
nearest neighbor	0.5274	-3.52
bilinear	0.5583	-0.43

In the convolutional transformer hybrid encoder, we investigated the significance of the two modules: efficient transformer and convolutional transformer, as indicated in **Table 4**. By removing and interchanging these modules, we observed that both of them hold pivotal roles in feature extraction. Furthermore, the convolutional transformer excels in capturing information related to strip-like objects and category boundaries while extracting low-level feature maps.

Table 4. Ablation study results for the convolutional transformer hybrid encoder.

Method	mIoU	Line-class mIoU	mIoU Δ (%)
CTHE	0.5626	0.5202	-
w/o efficient transformer	0.5614	0.5201	-0.12
w/o convolutional transformer	0.5562	0.5125	-0.64
Swap two module	0.5597	0.5175	-0.29

In **Table 5**, we investigated the efficacy of boundary supervision. For both the CNN-based models (DDRNet and BiSeNet) and the Transformer-based model (SegFormer), incorporating distinct boundary supervision into the head resulted in enhanced performance.

Table 5. Ablation study results for boundary branch supervision.

Method	Acc	Kappa	Dice	mIoU
DDRNet[51]	0.8427	0.7708	0.6301	0.5052
+ Boundary sup.	0.8431	0.7865	0.6306	0.5196
BiSeNet[52]	0.8265	0.7467	0.5671	0.4482
+ Boundary sup.	0.8283	0.7535	0.5685	0.4738
Segformer	0.9330	0. 8509	0.6691	0.5479
+ Boundary sup.	0.9337	0. 8583	0.6743	0.5541

The feature maps of the main branch comprise three components: the feature derived from multi-level fusion, the boundary feature from the boundary supervision branch, and feature map 1. We illustrate the impact of the information contained in these three feature maps on the main branch. The boundary feature contributes edge information, while feature map 1 preserves detail information from the original image. As demonstrated in **Table 6**, omitting both components leads to a notable decrease in model performance, particularly in the segmentation of strip-like objects. The line mIoU decreases from 52.02% to 48.33% when both features are removed.

Table 6. Ablation study results for main branch supervision.

Method	mIoU	Line mIoU	Line mIoU Δ (%)
Original method	0.5626	0.5202	-
w/o Boundary feature	0.5411	0.5035	-1.67
w/o Feature map 1	0.5473	0.4982	-2.20
w/o Both	0.5129	0.4833	-3.69

4.5. Comparative results

4.5.1. The quantitative comparative results

We conducted a comparison of various methods and presented the mIoU scores for each category in **Table 7**. Our method attains state-of-the-art results in the majority of categories.

Table 7. Per-category performance comparison of different models.

	Category/Method		**DDRNet**	**STDC2**	**Segformer**	**Ours**
mIoU	Invalid class		0.794	0.7737	0.9213	**0.9219**
	Line	single_solid_line	0.3937	0.342	0.4261	**0.441**
		double_solid_line	0.1681	0.2065	0.4245	**0.4363**
		single_dashed_line	0.4247	0.3583	**0.5313**	0.5257
		double_dashed_line	0.0037	0.0050	0.1004	**0.1745**
		dashed_and_solid_line	0.1894	0.0903	0.2074	**0.3284**
		special_line	0.6776	0.6571	0.6041	**0.6174**
		stop_line	0.532	0.4073	**0.2649**	0.2564
	Surface	paved_road	0.8559	0.8457	0.8411	**0.8415**
		no_stop_zone	0.6349	0.6265	**0.786**	0.7776
		zebra_crossing	0.8403	0.8372	0.7906	**0.7945**
		diagonal_line	0.4879	0.5062	0.653	**0.6174**

Table 8 presents the results achieved on our BEV point cloud image dataset using both CNN-based and Transformer-based methods. Our model demonstrates competitive performance, with accuracy, kappa, Dice, and mIoU scores of 95.26%, 87.14%, 69.02%, and 56.26%, respectively.

Table 8. Comparison results for different models on the BEV point cloud image dataset.

Method	**Acc**	**Kappa**	**Dice**	**mIoU**
STDC2[53]	0.8295	0.7520	0.5871	0.4640
DDRNet	0.8427	0.7708	0.6301	0.5052
BiSeNet	0.8265	0.7467	0.5671	0.4482
DNLNet[54]	0.8193	0.7357	0.5760	0.4534
Segformer	0.9330	0.8509	0.6691	0.5479
Ours.	**0.9526**	**0.8714**	**0.6902**	**0.5626**

The proposed method was additionally assessed on the Cityscapes dataset. As shown in **Table 9**, our model outperforms other methods, delivering superior results.

Table 9. Comparison results for different models on the Cityscapes dataset.

Method	Backbone	mIoU (%)
STDC2	STDC2	77.60
FCN	HRNet_W18	78.97
DDRNet	ddrnet	79.85
Deeplabv3	ResNet50_OS8	79.90
	ResNet101_OS8	80.85
SETR	ViT-L	77.29
Segformer	MiT-B3	82.47
Ours.	CTHE	**82.93**

4.5.2. Qualitative analysis of different models

Model robustness and finer details are crucial aspects of semantic segmentation. The visualization of results from our model and SegFormer can be observed in **Fig. 7** and **Fig. 8**. Notably, SegFormer struggles in effectively handle single and double lines, exhibiting intermittent segmentation on single lines, and encountering challenges with double lines. In contrast, our model adeptly extracts features of strip-like objects and accurately segment them.

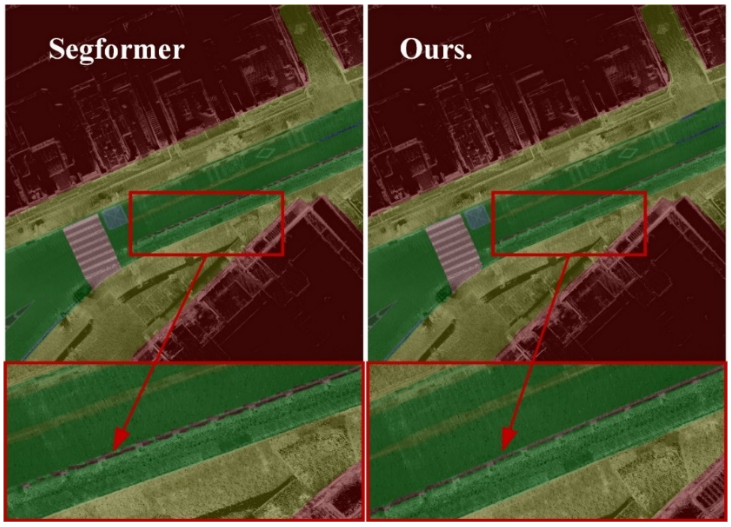

Fig. 7. Qualitative comparison of SegFormer and our model on single lines.

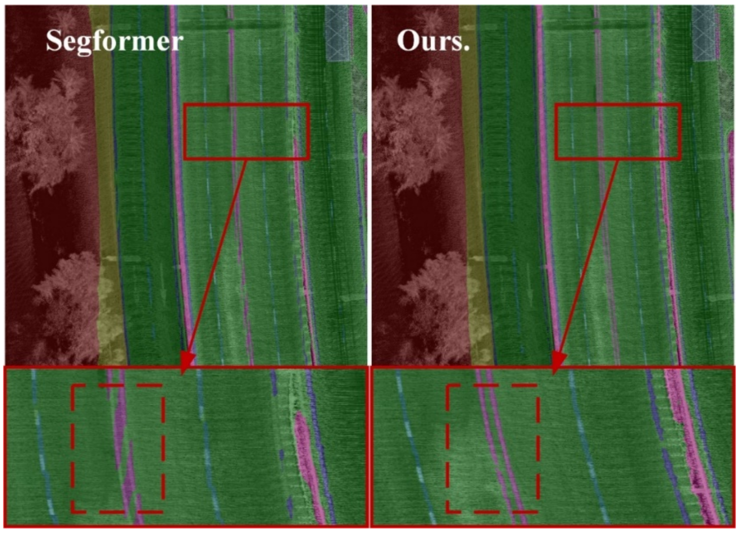

Fig. 8. Qualitative comparison of SegFormer and our model on double lines.

The results of our model and the CNN-based model DDRNet are visualized in **Fig. 9** and **Fig. 10**. Notably, our model exhibits excellent performance in accurately segmenting lane lines even in scenarios with vehicle occlusion. Additionally, when it comes to large-scale surface segmentations, our model surpasses the CNN-based mode by effectively distinguishing various types of regions, thanks to its ability to reduce detail and extract features.

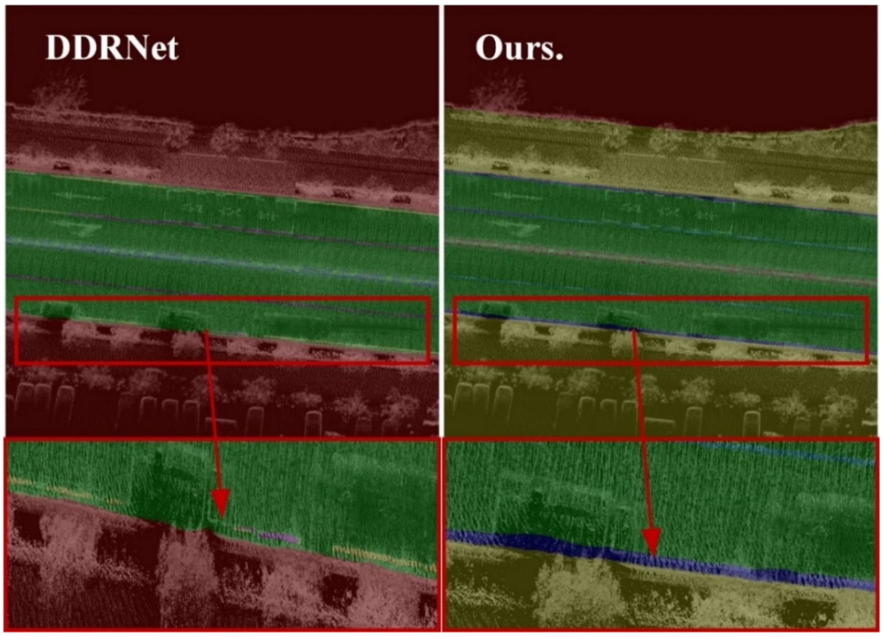

Fig. 9. Qualitative comparison of DDRNet and our model in the scene of vehicle occlusion.

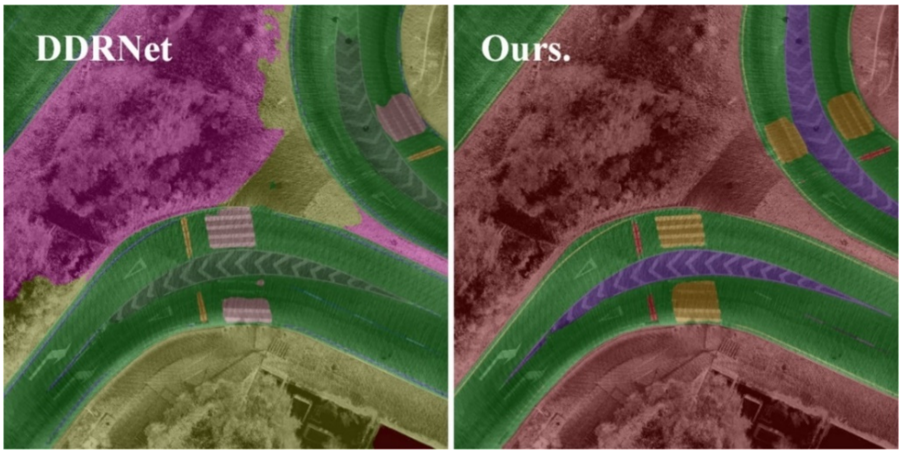

Fig. 10. Qualitative comparison of DDRNet and our model in surface segmentation task.

5. Conclusion

HD maps play a vital role in autonomous driving system, but their annotation currently relies on inefficient manual methods. In this paper, we proposed a novel semantic segmentation method called Mapformer for automatic HD map annotation. Our approach involves a carefully designed an encoder-decoder architecture that enhances feature extraction for strip-like objects and category boundaries, while also improving the model's ability to restore details. By achieving an mIoU of 56.26%, Mapformer outperforms SegFormer, setting a new state-of-the-art performance benchmark. With its potential to automate HD map annotation, Mapformer holds promise for significantly improving the annotation automation rate. However, our model still has some limitations, such as insufficient segmentation ability for hard samples and tail categories. Going forward, our future research will focus on addressing these limitations by exploring techniques for hard sample mining and addressing long-tail distributions challenges.

Acknowledgements

This work supported part by Beijing Natural Science Foundation (4232026); National Natural Science Foundation of China (grant nos. 62272049, 61932012, 61871039, 61906017, 62171042, 62102033 and 62006020) and Key project of science and technology of Beijing Education Commission (KZ202211417048); The Project of Construction and Support for high-level Innovative Teams of Beijing Municipal Institutions (Grant No. BPHR20220121) academic research projects of Beijing Union University (ZK10202202)

References

[1] J.Y. Yoon, J. Jeong, W. Sung, “Design and Implementation of HD Mapping, Vehicle Control, and V2I Communication for Robo-Taxi Services," *Sensors*, vol.22, no.18, pp 1-23, Sep. 2022. Article (CrossRef Link)

[2] Daduna, J.R, "Automated and Autonomous Driving in Freight Transport - Opportunities and Limitations," in *Proc. of International Conference on Computational Logistics*, pp. 457–475, 28-30 Sep. 2020. Article (CrossRef Link)

[3] N. Ma, Y. Gao, J.H. Li, D.Y. Li, “Interactive Cognition in Self-Driving,” *SCIENTIA SINICA Informationis*, vol.48, no.8, pp, 1083–1096, 2018. Article (CrossRef Link)

[4] N. Ma, D.Y. Li, W. He, et al., “Future Vehicles: Interactive Wheeled Robots,” *Sci. China Inf. Sci.* vol.64, pp, 1-3, Apr. 2021. Article (CrossRef Link)

[5] X. Lu, Y. Ai, B. Tian, “Real-Time Mine Road Boundary Detection and Tracking for Autonomous Truck,” *Sensors*, vol.20, no.04, pp, 1-19, Feb. 2020. Article (CrossRef Link)

[6] Y.B. Can, A. Liniger, D.P. Paudel, L.V. Gool, “Structured Bird’s-Eye-View Traffic Scene Understanding from Onboard Images,” in *Proc. of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, pp. 15641–15650, Oct. 2021. Article (CrossRef Link)

[7] N. Homayounfar, J. Liang, W.-C. Ma, J. Fan, X. Wu, and Urtasun, R., "DAGMapper: Learning to Map by Discovering Lane Topology,” in *Proc. of the IEEE/CVF International Conference on Computer Vision*, Seoul, Korea, pp. 2911–2920, Oct. 2019. Article (CrossRef Link)

[8] N. Homayounfar, W.-C. Ma, S.K. Lakshmikanth, R. Urtasun, “Hierarchical Recurrent Attention Networks for Structured Online Maps,” in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, pp. 3417–3426, June 2018. Article (CrossRef Link)

[9] A.V. Etten, “City-Scale Road Extraction from Satellite Imagery v2: Road Speeds and Travel Times,” in *Proc. of the IEEE Winter Conference on Applications of Computer Vision*; Snowmass Village, CO, USA, pp. 1775–1784, Mar. 2020. Article (CrossRef Link)

[10] S. He, F. Bastani, S. Jagwani, M. Alizadeh, et al., “Sat2Graph: Road Graph Extraction Through Graph-Tensor Encoding,” in *Proc. of European Conference on Computer Vision*, Glasgow, United Kingdom, pp. 51–67, Aug. 2020. Article (CrossRef Link)

[11] Y.Q. Tan, S.H. Gao, X.Y. Li, M.M. Cheng, B. Ren, “VecRoad: Point-Based Iterative Graph Exploration for Road Graphs Extraction,” in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 8907–8915, Jun. 2020. Article (CrossRef Link)

[12] F. Bastani, S. He, S. Abbar, M. Alizadeh, et al., “RoadTracer: Automatic Extraction of Road Networks from Aerial Images,” in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, pp. 4720–4728, Jun. 2018. Article (CrossRef Link)

[13] Z. Xu, Y. Sun, M. Liu, “ICurb: Imitation Learning-Based Detection of Road Curbs Using Aerial Images for Autonomous Driving,” *IEEE Robot. Autom. Lett.*, vol. 06, no. 02, pp, 1097–1104, Apr. 2021. Article (CrossRef Link)

[14] E. Shelhamer, J. Long, T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” *IEEE Trans. Pattern Anal.*, vol.39, no.04, pp, 640–651, Apr. 2017. Article (CrossRef Link)
L.C. Chen, G. Papandreou, I. Kokkinos, et al., “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs,” *arXiv paper*, pp, 1-14, 2014. Article (CrossRef Link)

[15] L.C. Chen, G. Papandreou, I. Kokkinos, et al., “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” *IEEE Trans. Pattern Anal.*, vol.40, no. 04, pp, 834–848, Apr. 2018. Article (CrossRef Link)

[16] L.C. Chen, G. Papandreou, F. Schroff, H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation,” *arXiv paper*, pp. 1-14, Dec. 2017. Article (CrossRef Link)

[17] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation,” in *Proc. of European Conference on Computer Vision In Computer Vision*, Munich, Germany, pp. 833–851, Sep. 2018. Article (CrossRef Link)

[18] O. Ronneberger, P. Fischer, T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Proc. of International Conference on Medical Image Computing and Computer-Assisted*, Munich, Germany, pp. 234–241, Oct. 2015. Article (CrossRef Link)

[19] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention Is All You Need,” in *Proc. of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, pp. 6000–6010, Dec. 2017. Article (CrossRef Link)

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *arXiv paper*, pp. 1-22, Jun. 2021. Article (CrossRef Link)

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Proc. of European Conference on Computer Vision*, Glasgow, United Kingdom, pp. 213–229, Aug. 2020. Article (CrossRef Link)

[22] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, et al., "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 6877–6886, Jun. 2021. Article (CrossRef Link)

[23] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *arXiv paper*, pp. 1-18, Oct. 2021. Article (CrossRef Link)

[24] T. Liang, H. Bao, W.G. Pan, F. Pan, "Traffic Sign Detection via Improved Sparse R-CNN for Autonomous Vehicles," *J. Adv. Transportation*, vol. 2022, pp. 1–16, Mar. 2022. Article (CrossRef Link)

[25] T. Liang, H. Bao, W.G. Pan, X. Fan, H. Li, "DetectFormer: Category-Assisted Transformer for Traffic Scene Object Detection," *Sensors*, vol. 22, no. 13, pp. 1-17, Jun. 2022. Article (CrossRef Link)

[26] H. Fujiyoshi, T. Hirakawa, T. Yamashita, "Deep Learning-Based Image Recognition for Autonomous Driving," *IATSS Research*, vol. 43, no. 04, pp. 244–252, Dec. 2019. Article (CrossRef Link)

[27] D. Feng, C. Haase-Schutz, L. Rosenbaum, et al., "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges," *IEEE Trans. Intell. Transport. Syst*, vol. 22, no. 03, pp.1341–1360, Mar. 2021. Article (CrossRef Link)

[28] S.M. Azimi, P. Fischer, M. Korner, P. Reinartz, "Aerial LaneNet: Lane-Marking Semantic Segmentation in Aerial Imagery Using Wavelet-Enhanced Cost-Sensitive Symmetric Fully Convolutional Neural Networks," *IEEE Trans. Geosci. Remote Sensing*, vol.57, no. 05, pp. 2920–2938, May 2019. Article (CrossRef Link)

[29] J. Wang, K. Sun, T. Cheng, B. Jiang, et al., "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Trans. Pattern Anal.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021. Article (CrossRef Link)

[30] C. Wen, X. Sun, J. Li, C. Wang, Y. Guo, A. Habib, "A Deep Learning Framework for Road Marking Extraction, Classification and Completion from Mobile Laser Scanning Point Clouds," *ISPRS J PHOTOGRAMM*, vol. 147, pp. 178–192, Jan. 2019. Article (CrossRef Link)

[31] Y. Yu, Y. Yao, H. Guan, et al., "A Self-Attention Capsule Feature Pyramid Network for Water Body Extraction from Remote Sensing Imagery," *Int J Remote Sens*, vol.42, no. 05, pp.1801–1822, Oct. 2020. Article (CrossRef Link)

[32] C. Ye, J. Li, H. Jiang, H. Zhao, L. Ma, M. Chapman, "Semi-Automated Generation of Road Transition Lines Using Mobile Laser Scanning Data," *IEEE Trans. Intell. Transport. Syst.*, vol.21, no.05, pp.1877–1890, May 2020. Article (CrossRef Link)

[33] L. Ma, Y. Li, J. Li, Z. Zhong, M.A. Chapman, "Generation of Horizontally Curved Driving Lines in HD Maps Using Mobile Laser Scanning Point Clouds," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol.12, no.05, pp. 1572–1586, May 2019. Article (CrossRef Link)

[34] L. Ma, Y. Li, J. Li, et al., "Capsule-Based Networks for Road Marking Extraction and Classification From Mobile LiDAR Point Clouds," *IEEE Trans. Intell. Transport. Syst*, vol.22, no. 04, pp.1981–1995, Apr. 2021. Article (CrossRef Link)

[35] X. Mi, B. Yang, Z. Dong, C. Liu, Z. Zong, Z. Yuan, "A Two-Stage Approach for Road Marking Extraction and Modeling Using MLS Point Clouds," *ISPRS J PHOTOGRAMM*, vol. 180, pp.255–268, Oct. 2021. Article (CrossRef Link)

[36] D. Prochazka, J. Prochazkova, J. Landa, "Automatic Lane Marking Extraction from Point Cloud into Polygon Map Layer," *EUR J REMOTE SENS*, vol. 52, pp. 26–39, Oct. 2018. Article (CrossRef Link)

[37] P. Lu, S.Xu, H. Peng, "Graph-Embedded Lane Detection," *IEEE Trans. on Image Process.*, vol. 30, pp. 2977–2988, Feb. 2021. Article (CrossRef Link)

[38] Y. Zhang, Z. Lu, D. Ma, J.H. Xue, Q. Liao, “Ripple-GAN: Lane Line Detection With Ripple Lane Line Detection Network and Wasserstein GAN,” *IEEE Trans. Intell. Transport. Syst*., vol. 22, no.03, pp. 1532–1542, Mar. 2021. Article (CrossRef Link)

[39] J. Zhang, T. Deng, F. Yan, W. Liu, “Lane Detection Model Based on Spatio-Temporal Network With Double Convolutional Gated Recurrent Units,” *IEEE Trans. Intell. Transport. Syst*. vol. 23, no. 07, pp. 6666–6678, Jul. 2022. Article (CrossRef Link)

[40] J. Lafferty, A. McCallum, F. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” in *Proc. of the Eighteenth International Conference on Machine Learning*, Williamstown, MA, USA, pp 282–289, Jun. 2001. Article (CrossRef Link)

[41] W. Zaremba, I. Sutskever, O. Vinyals, “Recurrent Neural Network Regularization,” *arXiv paper*, pp. 1-8, Feb. 2015. Article (CrossRef Link)

[42] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer Berlin Heidelberg: Berlin, Heidelberg, 2012.

[43] J. Chen, Y. Lu, Q. Yu, X. Luo, et al., “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation,” *arXiv paper*, pp.1-13, Feb. 2021. Article (CrossRef Link)

[44] M.H. Guo, C.Z. Lu, Q. Hou, Z. Liu, M.M. Cheng, S.M. Hu, “SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation,” *arXiv paper*, pp.1-15, Feb.2022. Article (CrossRef Link)

[45] M. Cordts, M. Omran, S. Ramos, et al., “The Cityscapes Dataset for Semantic Urban Scene Understanding,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 3213–3223, Jun. 2016.Article (CrossRef Link)

[46] Y. Liu, L. Chu, G. Chen, Z. Wu, Z. Chen, B. Lai, Y. Hao, “PaddleSeg: A High-Efficient Development Toolkit for Image Segmentation,” *arXiv paper*, pp. 1-11, Jan. 2021. Article (CrossRef Link)

[47] I. Loshchilov, F. Hutter, “Decoupled Weight Decay Regularization,” *arXiv paper*, pp. 1-19, Jan. 2019. Article (CrossRef Link)

[48] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” in *Proc. of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, pp. 9992–10002, Oct. 2021. Article (CrossRef Link)

[49] Z. Liu, H. Mao, C.Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, “A ConvNet for the 2020s,” *arXiv paper*, pp. 1-15, Mar. 2023. Article (CrossRef Link)

[50] H. Pan, Y. Hong, W. Sun, Y. Jia, “Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Traffic Scenes,” *IEEE Trans. Intell. Transport. Syst*., vol. 24, no. 03, pp. 3448-3460, Mar. 2023. Article (CrossRef Link)

[51] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, “BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation,” in *Proc. of European Conference on Computer Vision*, Munich, Germany, pp. 334–349, Oct. 2018.Article (CrossRef Link)

[52] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, X. Wei, “Rethinking BiSeNet For Real-Time Semantic Segmentation,” in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 9711–9720, June 2021. Article (CrossRef Link)

[53] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, H. Hu, “Disentangled Non-Local Neural Networks,” in *Proc. of European Conference on Computer Vision*, Glasgow, United Kingdom, pp. 191–207, Nov. 2020. Article (CrossRef Link)

TIANJIAO LIANG was born in Shandong province in 1998. Received the B.S. degree from the University of Jinan, in 2020, China. He is currently pursuing the M.S. degree in software engineering with the Beijing Union University. His research interests include computer vision, deep learning, and object detection.

WEIGUO PAN was born in Handan, Hebei Province, China. He received the B.S. degree North China University of Water Resources and Electric Power in 2009, and the M.S. degree in Beijing Union University, in 2012 and the Ph.D. degree in University of Chinese Academy of Sciences, in 2015. His research interest includes machine learning, object detection and intelligent driving.

HONG BAO was born in Beijing in 1958. He received the B.S. degrees in computer science from the Beijing Union University in 1983 and the Ph.D. degree in computer science from Bejing Jiaotong University in 2012. His research interests include intelligent driving, cognitive computing, networks and distributed systems.

XINYUE FAN was born in Heilongjiang province in 1998. Received the B.S. degree from Beijing Institute of Graphic communication, in 2020, China. She is currently pursuing the M.S. degree in software engineering with the Beijing Union University. Her research interests include computer vision, deep learning, and long-tail instance segmentation.

Li Han was born in 1998 in Hebei Province, China. In 2020, he received a bachelor's degree in engineering from Hebei Normal University of Science and Technology. He is currently pursuing the M.S. degree in software engineering with the Beijing Union University. His research interests include computer vision, deep learning and optical character recognition.