2024 IEEE Intelligent Vehicles Symposium (IV)
June 2-5, 2024. Jeju Island, Korea

Forecasting Semantic Bird-Eye-View Maps for Autonomous Driving

Shuang Gao, Qiang Wang, David Navarro-Alarcon, and Yuxiang Sun

2024 IEEE Intelligent Vehicle Symposium (IV) | 979-8-3503-4881-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/IV55156.2024.10588762

Abstract— **Correctly understanding surrounding environments is a fundamental capability for autonomous driving. Semantic forecasting of bird-eye-view (BEV) maps can provide semantic perception information in advance, which is important for environment understanding. Currently, the research works on combining semantic forecasting and semantic BEV map generation is limited. Most existing work focuses on individual tasks only. In this work, we attempt to forecast semantic BEV maps in an end-to-end framework for future front-view (FV) images. To this end, we predict depth distributions and context features for FV input images and then forecast depth-context features for the future. The depth-context features are finally converted to the future semantic BEV maps. We conduct ablation studies and create baselines for evaluation and comparison. The results demonstrate that our network achieves superior performance.**

I. Introduction

Semantic forecasting aims to segment future frames pixel-wisely from previous observations. It is important for semantic environment understanding, which is a fundamental capability of autonomous vehicles [1]–[9]. Semantic forecasting could facilitate the intelligent decision-making process [10], [11] by predicting the possible position of the other road agents and the road layout, enabling self-driving cars to avoid obstacles. The semantic bird-eye-view (BEV) map is an ideal format for such task because the BEV map is more flexible in representing the dynamically changing environment. The relative distance between the self-driving car and other agents can be explicitly illustrated. Compared with the front-view images, the BEV map could eliminate the foreshortening due to the perspective projection. Besides the advantages of representation, the BEV map provides a uniform coordinate to fuse the observation information from different modality inputs. This is in line with the development of autonomous driving, where an increasing number and variety of sensors are equipped for self-driving cars.

The conventional semantic segmentation tasks predict the semantic class for each pixel according to the observation. However, semantic forecasting is required to predict the semantic distribution for the unobserved frame according to the previous frames.

Early semantics-to-semantics (S2S) methods [12], [13] predict future semantic information with semantic segmentation from the past as the input to the network. Those S2S networks separate the semantic segmentation and forecasting into two tasks. Recently, feature-to-feature (F2F) forecasting has drawn attention in the forecasting research field. The methods [14], [15] adopt such approach to extract the feature from the origin RBG images and recover the feature maps to the semantic map directly. The semantic segmentation pays attention to the task under the front view. In contrast, the key point of the semantic BEV map generation is to predict the cross-view semantic position for the objects observed by the front-faced cameras. Recent works [16]–[19] have achieved satisfactory results with deep neural networks.

Most existing works focus on front-view semantic forecasting for the future frames or semantic BEV map generation for the current frame separately. Few attempts solve those two problems within a whole framework. In this work, we aim to forecast the short-term future semantic segmentation in the form of the BEV map. The most related work is proposed by Hoyer *et al.* [20] in 2019. However, they follow the S2S pipeline, conducting the semantic forecasting in two steps. They generate the semantic segmentation using the off-the-shelf method, DeepLabV3 [21], then transform the semantic information into the bird-eye view in the second step. Such two-step manipulation suffers error accumulation, resulting in inferior performance.

Different from the previous method, we propose an end-to-end semantic forecasting network to predict the semantic BEV map for future frames. We adopt the F2F framework, extracting the front-view feature from the previously observed images and then predicting the depth distribution with LSS [22]. We propose a dual-forecasting module for semantic forecasting, in which the context and depth of the unobserved frames can be predicted together. To the best of our knowledge, this is the first network that forecasts the semantic BEV map in an end-to-end manner. The contributions of this work are summarized as follows:

1) We propose an end-to-end framework to forecast semantic BEV map with F2F strategy for future frames.
2) We design a depth-context forecasting module to predict and fuse the future depth and context features.
3) We create a group of baseline methods based on the existing semantic BEV map generation networks and compare the performance of our network with those baselines.

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010116, in part by Hong Kong Research Grants Council under Grant 15222523, and in part by City University of Hong Kong under Grant 9610675. *(Corresponding author: Shuang Gao.)*

Shuang Gao is with Harbin Institute of Technology, Harbin, China, and also with The Hong Kong Polytechnic University, Kowloon, Hong Kong (email: gaoshuang.sarah@outlook.com).

Qiang Wang is with Harbin Institute of Technology, Harbin, China (email: wangqiang@hit.edu.cn).

David Navarro-Alarcon is with The Hong Kong Polytechnic University, Kowloon, Hong Kong (email: dnavar@polyu.edu.hk).

Yuxiang Sun is with City University of Hong Kong, Kowloon, Hong Kong (e-mail: yx.sun@cityu.edu.hk, sun.yuxiang@outlook.com).

II. Related Works

A. Semantic Segmentation Networks

Deep learning greatly improves the performance of the semantic segmentation task. Chen *et al.* [23] proposed the atrous convolution in DeepLabV3+. This operation further enlarges the convolutional receptive field and is beneficial for getting the global feature. The emergence of skip-connection enables the feature fusion between the downsampling and upsampling paths. The mainstream semantic segmentation methods [24], [25] gradually adopt the skip-connection structure and get competitive results.

Besides studying new semantic segmentation network architectures, Hu *et al.* [26] designed a fast attention module, achieving real-time semantic segmentation. Milletari *et al.* [27] developed a new loss function based on the Dice coefficient to improve the semantic segmentation performance. Zhang *et al.* [28] introduced ObjectAug, a new data augmentation method, into the semantic segmentation field.

B. Semantic Forecasting

Some early works [12], [13] forecast future semantics by mapping the past semantic segmentation to their future counterparts. Those S2S methods have been proven to be less efficient than the later proposed F2F approach [29]. In this work, the authors directly extracted the feature from the observed RGB images rather than using the past semantics. Chiu *et al.* [30] conducted a F2F network within the teacher-student framework to generate the supervision signal for semantic forecasting training. Hu *et al.* [14] proposed the Apanet, which explores the pyramid feature of the various network levels. Lin *et al.* [15] developed a Predictive Feature Autoencoder and established the connection between the segmentation features and the predictor, improving the future segmentation results.

C. Semantic BEV map generation

The semantic BEV map generation has recently received considerable attention in the autonomous driving community. Lu *et al.* [16] first attempted to predict the semantic BEV map using the convolutional variational encoder-decoder network. Yang *et al.* [19] performed the view transformation with a cross-view transformer. Besides using the agnostic convolution neural networks to conduct the view transformation and assign semantics, Philion *et al.* [22] proposed LSS, which first introduces the depth prediction into the semantic BEV map generation task. Later, Li *et al.* [31] further improved the depth perception ability by supervising the depth prediction network with the depth ground truth. In [32], the same author constructed a stereo-based method to achieve more reliable depth estimation.

Although the aforementioned methods perform well in respective research fields, the research in semantic forecasting for the future BEV map is insufficient. In this work, we aim to create a semantic forecasting network that could generate the future semantic BEV map given the previously observed RGB images. [20] is related to our work. We both forecast the future semantic BEV map, but our network is based on the F2F scheme rather than the S2S, eliminating the error from the intermediate step.

III. The Proposed Method

A. The Overall Architecture

Fig. 1 shows the overall architecture of our proposed network. The proposed network mainly consists of a feature extractor, a depth-context forecasting module, and a frustum grid generator. In the first part of our network, the EfficientNet [33] is adopted as the backbone network. We employ the backbone network to extract the front-view feature from the observed RGB images rather than directly using the semantic maps as the S2S methods. Then, the depth and context feature for the future frame is predicted based on the past frames' feature maps within the depth-context forecasting module. At the same time, the RGB images from the past frames and the camera matrix are fed into the frustum grid generator to get the frustum-shaped point cloud. Each point in this point cloud corresponds to a pixel of the given image at various depths. After getting the point cloud, we assign the obtained depth-context feature to each point and project those points into the BEV plane. Through the semantic head at the end of the proposed network, a semantic BEV map for the future frame can be generated.

B. BEV Feature Map Generation

In this work, the images from the front-faced camera and the extrinsic and intrinsic matrix are taken as input to the whole network. Let $I_t \in \mathbb{R}^{3\times H\times W}$ denote the input front-view images from the past t frames. Those images are fed into a pre-trained CNN model, EifficientNet, to get the individual feature, $\mathcal{F}_{front}$. The dimension of $\mathcal{F}_{front}$ is $B\times t\times C\times h\times w$, where B, t, C, h, w stand for the batch size, numbers of the past input, channel size, the height and width of the extracted feature maps. Given those feature maps, our semantic forecasting network can predict the semantic BEV maps for the future frame, F_{t+m}, where m denotes the timestamp of the future.

After getting the past t frames' feature maps, we transfer perspective from the front view to BEV. To this end, the feature maps in BEV space are generated by first lifting the 2D front-view images, I_t, into the 3D point cloud, P_t. Because the input images are from a monocular camera, the depth estimation for a single image seems like an ill-posed problem without any other input. Taking the camera extrinsic matrix, and intrinsic matrix as the input, each pixel in the image can be projected into the world coordinate, but the individual depth is not sure, which is formulated as:

$$z_c \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = K \cdot \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x_w \\ y_w \\ z_w \\ 1 \end{pmatrix} \tag{1}$$

where $(u, v, 1)$ is the coordinate of an image pixel p, represented in the form of homogeneous coordinates. K denoted the camera intrinsic matrix. R and T are the rotation and translation matrix, describing the camera's motion pose.

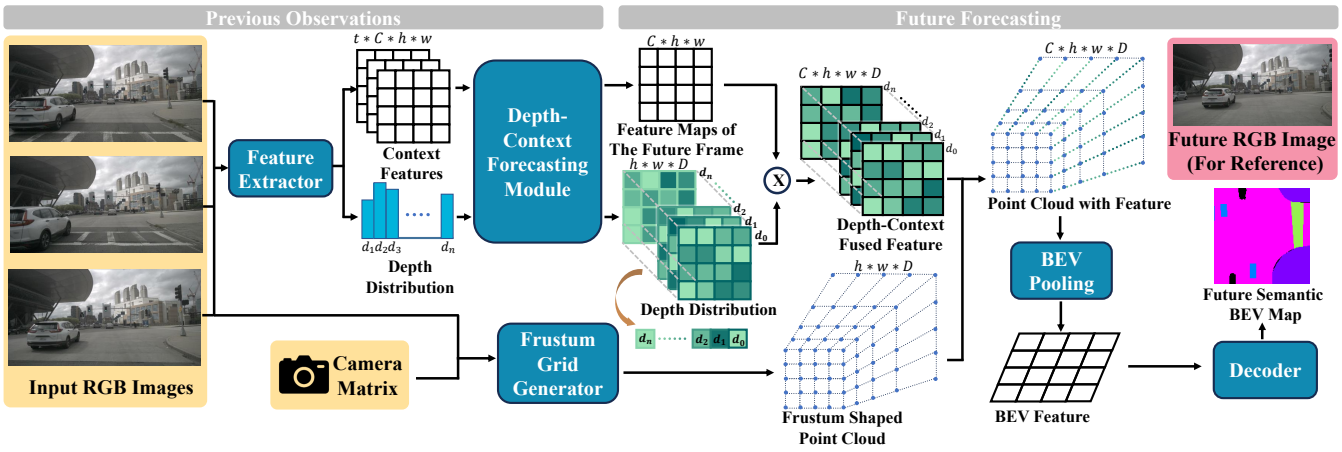

Fig. 1: The overall architecture of the proposed semantic forecasting network.

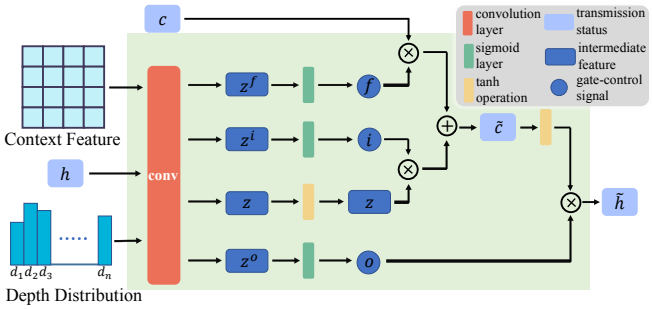

Fig. 2: The structure of the proposed depth-context forecasting module.

$(x_w, y_w, z_w, 1)$ is the world coordinate of a point P_w, corresponding to pixel p. z_c is the distance between the real-world point P_w and the camera, namely the depth of the pixel p. Note that the depth z_c is uncertain for a monocular image.

To get the corresponding depths of the pixels in the given monocular image, we base on [22], assigning n possible depths to each pixel in the I_t. The possible depth $D = \{d_0, d_1, ..., d_n\}$ is a set of equidistant discrete values. Thus, the 2D image can be projected into a frustum-shaped point cloud with depth D. The depth distribution probability α is predicted, which can be considered as the confidence score at the different depths. At the same time, the context feature F_c of the input is extracted through the backbone network. Then, we get the depth-context feature for each point in the frustum-shaped point cloud, f_{dc} by combining the depth confidence score and the context feature:

$$f_{dc} = \alpha \cdot f_c \tag{2}$$

We use the sum pooling to produce the BEV feature maps by projecting the point feature into the BEV grids.

C. Future Semantic Forecasting

The depth-context forecasting module is displayed in Fig. 2. We design this module based on the convLSTM. The operations in the green background compose the forecasting block. Here, we simultaneously forecast the depth and context features for the future frame. First, for the same frame, its depth and context features are fed into a convolutional layer to get 4 intermediate feature maps. Those feature maps change into the different gates (the information, forget, output gate) via the sigmoid function. The gates control the transmission of the information to the next forecasting block. Every frame in the input sequence goes through this forecasting block and predicts the next status for the fusion with the next frame. This module is able to process the sequential input and can be inserted into other existing networks seamlessly.

TABLE I: The ablation study results (%) of the variants of the EfficientNet Family. Eff is the short for the EfficientNet. The seven semantic classes are divided into static and dynamic categories, and the mIoU and mAP for those two categories, as well as the mean results across the seven classes, are reported respectively. The best results are highlighted in bold font.

Variants	Statics		Dynamics		mIoU	mAP
	mIoU	mAP	mIoU	mAP		
Eff-B0	36.55	62.17	8.21	24.76	27.85	46.14
Eff-B1	42.77	61.62	7.71	28.14	27.74	47.27
Eff-B2	42.61	60.01	8.26	26.41	27.89	45.61
Eff-B3	42.68	60.70	8.48	25.71	28.02	45.71
Eff-B4	42.94	62.51	8.26	25.46	28.08	46.63
Eff-B5	**43.48**	**63.70**	7.36	25.98	28.00	47.54
Eff-B6	43.03	62.01	8.30	23.59	28.14	45.54
Eff-B7	43.22	63.22	**8.61**	**28.45**	**28.39**	**48.32**

D. The Semantic BEV Head

After getting the BEV feature map, a semantic BEV head is introduced to generate the semantic BEV map. This module first conducts the feature learning from the BEV space by a structure that contains the first three layers of the ResNet18 [34]. The size of the BEV feature map shrinks after those layers, and then the upsampling operation is used to recover the output size of the semantic BEV map.

IV. Experimental Results and Discussions

A. The Dataset

In our experiments, we use a public autonomous driving dataset, nuScenes [35], to evaluate the performance of our network for semantic BEV map forecasting. There are 850 annotated scenes in the nuScenes dataset. The annotations include the 3D object bounding boxes, the high-definition (HD) maps, and the camera matrix for every frame. Using those annotations, we create the sequential input images for semantic forecasting and the future semantic BEV map as the ground truth labels. We annotate the 7 semantic classes, including the background, drivable area, pedestrian crossing, walkway, obstacle, vehicle and pedestrian. Note that some ground truth labels may not be properly generated due to the limitation of the flat ground assumption. To train the network, we randomly split the whole dataset into 548 training sets, 150 validation sets, and 148 test sets, excluding the sets that contain incorrect semantic BEV labels. For the input sequence, we choose the 3 consecutive frames as the input and take the $4th$ or $6th$ frame for future forecasting. The size of the input images is 256×512, and the output future semantic BEV map contains 150×150 grids, whose resolution is 0.2 m.

B. Training Details

Our proposed network is implemented on an NVIDIA GeForce RTX 3090 (24 GB RAN) graphics card. Taking the computation cost and the time consumption into consideration, we set the batch size to 16. We train our network for 30 epochs with the Adam optimizer. The initial learning rate is 1×10^{-4} and the weight decay rate is 1×10^{-5}.

C. Ablation Study

We conduct ablation studies to verify the effectiveness of the proposed network. In our experiments, the mean Intersection over Union (mIoU) and the mean Average Precision (mAP) are used as the evaluation metrics to assess the network performance.

1) Ablation on the Backbone Network: Since we chose the F2F strategy to forecast the future semantics, it is important to select a powerful backbone network to extract the front-view features from the previously observed images. EfficientNet [33] is known for its accuracy and efficiency. The EfficientNet includes 8 variants, whose structures mainly differ in depth, channel and width. The names of the different variants range from Efficient-B0 to Efficient-B7. This ablation study compares the performance of the network equipped with different EfficientNet variants.

We report the ablation study results in Tab. I. The correct predictions of the road layout and the objects on the road are both critical for autonomous driving. For the convenience of comparison, we divide the 7 semantic classes into the static and dynamic categories. The former includes the background, drivable area, pedestrian crossing, and walkway; the latter includes the obstacle, vehicle and pedestrian. The table shows an obvious rising trend in mIoU and mAP when the backbone changes from a simple structure to a complex one. Therefore, we chose EfficientNet-B7 as our backbone for the proposed network.

TABLE II: The ablation study results (%) of the semantic forecasting. The experiment is separated into two groups, forecasting the $1st$ and $3rd$ future frame, respectively. To further verify the semantic forecasting ability, we set three different inputs for each group. I_n stands for the number of previously observed frames, and O_f indicates which frame is predicted in the future. The best results are highlighted in bold font for forecasting $1st$ and $3rd$ future frame, respectively.

I_n	O_f	Statics		Dynamics		mIoU	mAP
		mIoU	mAP	mIoU	mAP		
1	$1st$	41.09	**63.96**	6.71	26.48	26.36	47.90
3	$1st$	**43.22**	63.22	**8.61**	**28.45**	**28.39**	**48.32**
5	$1st$	41.61	61.44	7.37	26.03	26.93	46.27
1	$3rd$	36.74	56.64	5.07	21.89	23.10	41.75
3	$3rd$	38.24	56.22	6.03	24.36	24.43	42.57
5	$3rd$	**38.84**	**58.00**	**7.22**	**26.94**	**25.29**	**44.69**

2) Ablation on the Semantic Forecasting: In this section, we compare the forecasting performance of the network with different input and output conditions. This ablation study is separated into two groups, which predict the semantic BEV map for the $1st$ and $3rd$ future frame, respectively. Furthermore, we also set different numbers of the previously observed frames as input for each group. Specifically, the 1, 3, and 5 past front-view images are fed into the network to forecast the future one or three frames.

Tab. II displays the results of this ablation study. We find that the network forecasting the future one frame performed best when taking as input 3 past observations, while worse having 5 inputs. The situation is changed for forecasting the $3rd$ future frame. The table shows that the best performance can be achieved by taking 5 past frames. We conjecture the reason for this change in the results is the information redundancy due to the increasing numbers of input. Using too many past frames to forecast the near future frame is redundant, whereas longer-term future prediction requires more past information to perform better.

D. Comparative Study

As the proposed network is the first method to forecast the semantic BEV map in the F2F manner, we create several baseline methods to perform the comparative experiments. The networks we chose for baseline comparison are specific to the semantic BEV map generation task. The networks include VPN [17], VED [16], PYVA [19]. All those networks can only predict the current semantics without the forecasting ability. So, to test the semantic forecasting performance of the proposed module, we integrate the feature forecasting module into those networks. The feature forecasting module is inserted behind the feature extractors to keep the original network structures unchanged. In this experiment, we also conduct the test with different input and output conditions to compare the precision of semantic forecasting. Meanwhile,

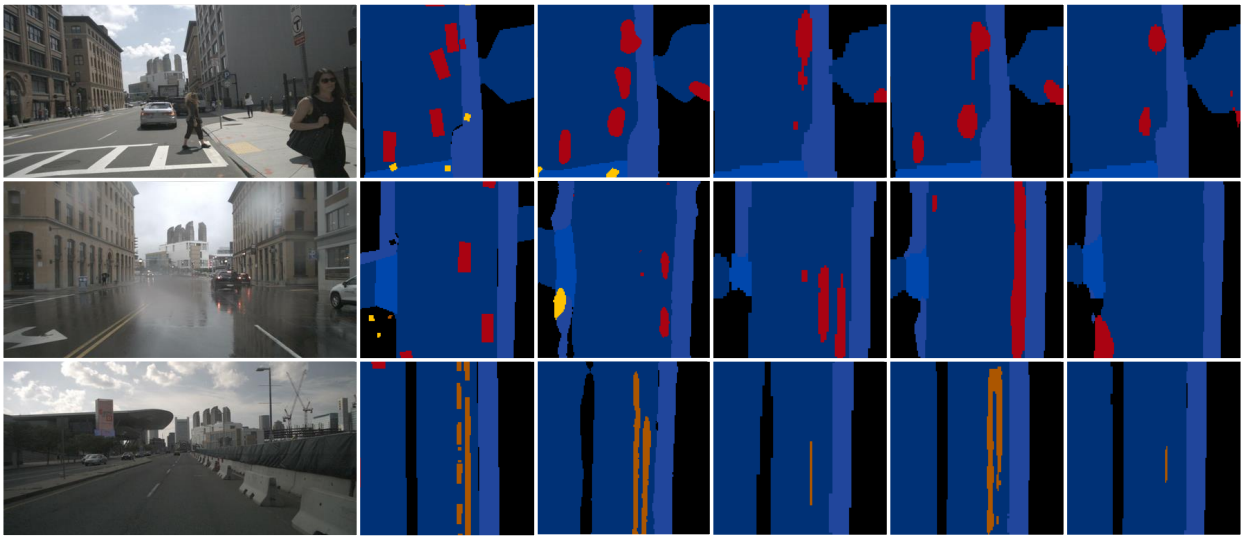

Fig. 3: Sample qualitative demonstrations for the semantic BEV map forecasting networks. The displayed results are the next future frame predicted by each network testing with the three previously observed frames as input. The results demonstrate the superiority of our network. The figure is best viewed in color.

TABLE III: The comparative results (%) with the baseline methods. We conduct different groups of experiments to test the performance of the selected network with the proposed semantic forecasting module. Each network takes as input 1 or 3 past frames and forecasts the next frame or the 3*rd* frame in the future. I_n and O_n represent the numbers of the input images and which frame is predicted in the future, respectively. We bold the best results according to the different input-output conditions for each method.

Network	I_n / O_n	Background		Drivable Area		Ped. Crossing		WalkWay		Obstacle		Vehicle		Pedestrian		mIoU	mAP
		IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP	IoU	AP		
VPN	1/1	53.05	61.17	61.12	**80.92**	11.24	**71.84**	28.93	39.91	0.54	**34.49**	2.59	51.07	0.0	0.0	22.50	**48.48**
	3/1	**56.21**	**71.06**	**67.65**	73.87	**22.58**	67.24	**29.31**	**63.10**	0.0	0.0	7.66	**38.84**	0.0	0.0	**26.20**	44.87
	3/3	52.38	62.21	62.27	76.69	19.35	56.57	27.18	47.26	**2.75**	32.10	**8.22**	31.41	0.0	0.0	24.59	43.75
VED	1/1	56.32	**71.99**	67.36	74.40	27.31	59.74	33.57	60.22	0.0	0.0	3.60	53.80	0.0	0.0	26.88	45.74
	3/1	**61.06**	69.56	**70.88**	**78.63**	**34.49**	**63.61**	**36.13**	**63.91**	0.95	0.0	**6.67**	**51.04**	0.0	0.0	**27.63**	46.69
	3/3	54.48	66.75	65.41	75.36	22.66	61.08	31.47	56.15	**2.20**	**35.60**	4.84	42.62	0.0	0.0	25.87	**48.22**
PYVA	1/1	54.33	**70.31**	66.20	75.56	28.27	55.12	31.64	51.18	3.08	36.81	7.60	36.90	0.0	0.0	27.3	42.91
	3/1	**58.53**	68.86	**68.98**	**80.36**	**29.87**	**61.27**	**33.79**	**53.98**	0.0	0.0	6.74	**42.14**	0.0	0.0	**28.27**	**43.80**
	3/3	54.66	67.73	65.43	76.41	25.03	53.06	30.30	50.60	**5.21**	**36.65**	6.92	32.09	0.0	0.0	26.79	45.22
OURS	1/1	50.31	**62.99**	**62.94**	72.71	25.15	58.50	25.97	**61.65**	10.17	41.48	9.87	**37.33**	0.10	0.62	26.36	47.90
	3/1	**51.45**	62.31	62.87	**77.12**	**27.57**	**59.62**	**30.98**	53.84	**13.64**	**53.42**	**11.71**	29.15	0.48	2.78	**28.39**	**48.32**
	3/3	50.27	61.95	61.49	74.90	25.61	54.18	29.07	54.74	11.42	49.89	9.94	24.85	**0.74**	**3.35**	26.93	46.27

we use the original network structures to predict the semantic BEV map for the next future frame with the 1 frame input as a baseline.

1) Quantitative Results: The comparative results are shown in Tab. III. Taking the three previously observed frames, the proposed network achieves the best forecasting performance, with 28.39% in mIoU and 48.32% in mAP. From the table, we can see that all the networks inserted with the forecasting module get better forecasting results compared with the origin structure (marked with $1/1$ for the I_n/O_n term). This verifies the effectiveness of our semantic forecasting module. In addition, the table shows that our network performs best in predicting small dynamic objects, such as obstacles, vehicles, and pedestrians, illustrating the superiority of our network.

2) Qualitative Demonstrations: Some sample qualitative results are shown in Fig. 3. The networks take three previous frames as input and forecast the next future semantic BEV map. In general, our network achieves the best performance for the semantic forecasting task. Compared with the other networks, Our network is sensitive to small objects like obstacles and pedestrians (labeled in brown and yellow, respectively).

V. Conclusions and Future Work

Semantic forecasting could provide the prior information for the other tasks in autonomous driving. In this work, we attempt to forecast the future semantic BEV maps in an F2F manner. The proposed network takes as input the previously observed image and outputs the future semantics in BEV. The network was evaluated and tested on the public dataset. We demonstrated our network's superiority over the baselines. Although the proposed network has satisfying forecasting performance, the precision of the small objects prediction is still not as good as that of the road layout. In the future, we will explore a better way to improve the segmentation performance of small objects.

References

[1] S. Teng, X. Hu, P. Deng, B. Li, Y. Li, Y. Ai, D. Yang, L. Li, Z. Xuanyuan, F. Zhu, and L. Chen, "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 6, pp. 3692–3711, Jun. 2023.

[2] P. Cai, Y. Sun, H. Wang, and M. Liu, "Vtgnet: A vision-based trajectory generation network for autonomous vehicles in urban environments," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 3, pp. 419–429, 2021.

[3] Y. Feng, W. Hua, and Y. Sun, "Nle-dm: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 9, pp. 9780–9791, 2023.

[4] W. Ma, S. Huang, and Y. Sun, "Triplet-graph: Global metric localization based on semantic triplet graph for autonomous vehicles," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3155–3162, 2024.

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

[6] W. Ma, H. Yin, L. Yao, Y. Sun, and Z. Su, "Evaluation of range sensing-based place recognition for long-term urban localization," *IEEE Transactions on Intelligent Vehicles*, pp. 1–12, 2024.

[7] P. S. Chib and P. Singh, "Recent advancements in end-to-end autonomous driving using deep learning: A survey," *IEEE Transactions on Intelligent Vehicles*, 2023.

[8] Z. Feng, Y. Guo, and Y. Sun, "Segmentation of road negative obstacles based on dual semantic-feature complementary fusion for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, pp. 1–11, 2024.

[9] Y. Feng and Y. Sun, "Polarpoint-bev: Bird-eye-view perception in polar points for explainable end-to-end autonomous driving," *IEEE Transactions on Intelligent Vehicles*, pp. 1–11, 2024.

[10] L. Xiong, Y. Zhang, Y. Liu, H. Xiao, and C. Tang, "Integrated decision making and planning based on feasible region construction for autonomous vehicles considering prediction uncertainty," *IEEE Transactions on Intelligent Vehicles*, 2023.

[11] L. Wang, C. Fernandez, and C. Stiller, "High-level decision making for automated highway driving via behavior cloning," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 923–935, 2022.

[12] A. Bhattacharyya, M. Fritz, and B. Schiele, "Bayesian prediction of future street scenes using synthetic likelihoods," *arXiv preprint arXiv:1810.00746*, 2018.

[13] M. Rochan *et al.*, "Future semantic segmentation with convolutional lstm," *arXiv preprint arXiv:1807.07946*, 2018.

[14] J.-F. Hu, J. Sun, Z. Lin, J.-H. Lai, W. Zeng, and W.-S. Zheng, "Apanet: Auto-path aggregation for future instance segmentation prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3386–3403, 2021.

[15] Z. Lin, J. Sun, J.-F. Hu, Q. Yu, J.-H. Lai, and W.-S. Zheng, "Predictive feature learning for future segmentation prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7365–7374.

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

[17] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

[18] 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 Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 974–11 981, 2022.

[19] W. Yang, Q. Li, W. Liu, Y. Yu, Y. Ma, S. He, and J. Pan, "Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 536–15 545.

[20] L. Hoyer, P. Kesper, A. Khoreva, and V. Fischer, "Short-term prediction and multi-camera fusion on semantic grids," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

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

[22] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[24] M. Oršić and S. Šegvić, "Efficient semantic segmentation with pyramidal fusion," *Pattern Recognition*, vol. 110, p. 107611, 2021.

[25] Y. Liu, C. Ding, Y. Tian, G. Pang, V. Belagiannis, I. Reid, and G. Carneiro, "Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1151–1161.

[26] P. Hu, F. Perazzi, F. C. Heilbron, O. Wang, Z. Lin, K. Saenko, and S. Sclaroff, "Real-time semantic segmentation with fast attention," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 263–270, 2020.

[27] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[28] J. Zhang, Y. Zhang, and X. Xu, "Objectaug: object-level data augmentation for semantic image segmentation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[29] P. Luc, C. Couprie, Y. Lecun, and J. Verbeek, "Predicting future instance segmentation by forecasting convolutional features," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 584–599.

[30] H.-k. Chiu, E. Adeli, and J. C. Niebles, "Segmenting the future," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4202–4209, 2020.

[31] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485.

[32] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li, "Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1486–1494.

[33] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[35] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.