S2G2: Semi-Supervised Semantic Bird-Eye-View Grid-Map Generation Using a Monocular Camera for Autonomous Driving

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

Abstract—**Semantic bird-eye-view (BEV) grid map is a straightforward data representation for semantic environment perception. It can be conveniently integrated with downstream tasks, such as motion planning, trajectory prediction, etc. Most existing methods of semantic BEV grid-map generation adopt supervised learning, which requires extensive hand-labeled ground truth to achieve acceptable results. However, there exist limited datasets with hand-labeled ground truth for semantic BEV grid map generation, which hinders the research progress in this field. Moreover, manually labeling images is tedious and labor-intensive, and it is difficult to manually produce a semantic BEV map given a front-view image. To provide a solution to this problem, we propose a novel semi-supervised network to generate semantic BEV grid maps. Our network is end-to-end, which takes as input an image from a vehicle-mounted front-view monocular camera, and directly outputs the semantic BEV grid map. We evaluate our network on a public dataset. The experimental results demonstrate the superiority of our network over the state-of-the-arts.**

Index Terms—**Semi-supervised learning, semantic BEV grid maps, view transformation, autonomous driving.**

I. Introduction

DATA representation for semantic environment perception is critical in autonomous driving. In recent years, semantic bird-eye-view (BEV) grid maps have attracted increasing attention in the robotics research community. Compared to the common data representation (i.e., front-view semantic segmentation images), semantic BEV grid maps are more straightforward to use. In semantic BEV maps, geometric relationships between ego-vehicle and obstacles are explicitly illustrated in a natural view. This advantage makes them more suitable for downstream tasks, such as motion planning [1], [2], [3], trajectory prediction, etc. Moreover, many networks of these downstream tasks are trained with visual images in simulation environments (e.g., CARLA). They suffer from the domain gap issue when transferred to real-world environments, since simulation environments lack real texture details compared to real-world environments. Using semantic BEV grid maps instead of visual images could alleviate this issue, because semantic maps almost have the same style in both simulation and real-world environments.

Different from semantic segmentation algorithms that label front-view camera images pixel-wisely into front-view semantic maps, our task is more like a generation process which generates semantic BEV maps from front-view camera images. To achieve this goal, some works [4], [5] first generate standard front-view semantic maps from front-view camera images, and then project the segmentation maps into the bird eye view using view transforming algorithms, such as the inverse perspective mapping (IPM) algorithm. However, the IPM algorithm suffers from the flat ground assumption [6], and the pipeline accumulates errors through the two steps. The issues make this stream of methods less generalizable. To address these issues, recent methods [7], [8], [9] resort to generating semantic BEV maps in an end-to-end manner, which could avoid using the IPM algorithm and alleviate the error propagation issue. However, most of the existing end-to-end methods adopt supervised learning to train their networks, which requires a large amount of ground-truth images to achieve acceptable results. The datasets with hand-labeled ground truth for semantic BEV grid maps are limited. In addition, manually labeling images is tedious and labor-intensive, and manually drawing semantic BEV maps according to the front-view camera images is difficult for humans.

To provide a solution to this problem, we propose a novel Semi-Supervised semantic BEV Grid-map Generation (S2G2) network, which requires only a small amount of labeled data and a large amount of unlabeled data to achieve superior performance. Our network is end-to-end. It consists of two major components: view transformation from front-view to bird eye view, and semantic labeling on the bird eye view.

To the best of our knowledge, our network S2G2 is the first solution to generate semantic BEV maps in a semi-supervised manner. We implement multiple baselines to perform extensive comparative studies on a public dataset [9]. The results demonstrate our superiority. The contributions of this work are summarized as follows:

Manuscript received 6 April 2022; accepted 14 August 2022. Date of publication 21 September 2022; date of current version 3 October 2022. This letter was recommended for publication by Associate Editor H. Myung and Editor H. Moon upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grants 61876054 and 62003286, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010116, in part by the Open Research Project of Zhejiang Lab under Grant 2021NL0AB01, and in part by the Start-up Fund of The Hong Kong Polytechnic University under Grant P0034801. *(Corresponding author: Shuang Gao.)*

Shuang Gao is with the Department of Control Science and Engineering, Harbin Institute of Technology, 150001 Harbin, China, and also with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: shuang.gao@connect.polyu.hk).

Qiang Wang is with the Department of Control Science and Engineering, Harbin Institute of Technology, 150001 Harbin, China (e-mail: wangqiang@hit.edu.cn).

Yuxiang Sun is with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: yx.sun@polyu.edu.hk, sun.yuxiang@outlook.com).

Digital Object Identifier 10.1109/LRA.2022.3208377

1) We propose S2G2[1], a novel semi-supervised semantic BEV generation network that can be trained with unlabeled data.
2) We introduce a new dual-attention view transformation module to transform the front-view input into the bird-eye-view feature maps.
3) We create several semi-supervised baseline methods and compare our network with the baselines and the state-of-the-art supervised methods.

II. Related Work

Our related work mainly involves semantic segmentation, semantic BEV grid map generation, and semi-supervised learning. We review several representative works in these fields here.

A. Semantic Segmentation

Semantic segmentation aims to label each pixel of a given image into individual classes. Badrinarayanan et al. [10] designed SegNet, which shows the potential of the deep learning in semantic segmentation. The network consists of an encoder and a decoder. The encoder is used to extract features from input images, and the decoder is used to restore the spatial resolution and produce the segmentation map. Different from SegNet, U-Net [11] also uses the encoder-decoder architecture, but it introduced a set of shortcut connections to pass the feature maps from the encoder to the decoder. Chen et al. [12] proposed DeepLab V3+, in which the atrous convolution and atrous spatial pyramid pooling were proposed to improve the segmentation performance.

B. Semantic BEV Grid Map Generation

1) Point Cloud-Based Methods: Some methods use point clouds produced by radar or LiDAR for semantic grid map generation. For example, Sless et al. [13] proposed a learnable inverse sensor model which maps the sparse and noisy Radar data into binary occupancy grid map in a data-driven manner. RadarNet [14] exploited both radar and LiDAR sensors for perception and designed a two-stage fusion module to deal with the problem of noisy data and measurement ambiguities. Isele et al. [15] transferred deep learning-based LiDAR segmentation approaches into the radar point-cloud segmentation. Their work encapsulated the semantic information into a polar-coordinate map. Kempen et al. [16] developed an end-to-end learning framework that can quantify the first- and second-order uncertainty, producing a reliable occupancy grid map.

2) Visual Image-Based Methods: Some methods use images produced by visual cameras. For example, Roddick et al. [17] proposed a pyramid occupancy network that predicts semantic grid maps by a set of multiscale dense transformers and a top-down module. Dwivedi et al. [18] utilized the monocular depth estimation to facilitate the BEV segmentation task, using the pseudo-LiDAR point cloud generated by the depth prediction. Lu et al. [9] designed a modified variational encoder-decoder network to get the semantic grid map in bird-eye view by utilizing the hallucination ability of CNNs. MonoLayout [19] predicted the BEV road layout and complemented the occlusion parts in a single RGB image via adversarial feature learning. Yang et al. [8] developed a cross-view transformation module to perform the perspective changing from a frontal view into the bird-eye view in order to generate a BEV binary semantic map for driving surroundings.

C. Semi-Supervised Learning

For semi-supervised learning, we generally divide the current methods into two categories: contrastive learning [20], [21] and transfer learning [22]. Contrastive learning penalizes the consistency loss between the outputs from two identical networks that take as input the diversely augmented versions of the same image. In transfer learning, one teacher network is pre-trained off-line with the labeled data to obtain a satisfactory performance. The knowledge from the teacher network is then transferred to the student network by training with the pseudo-labels of unlabeled data generated by the teacher network. Both the two categories of semi-supervised methods have limitations. For example, the two identical parallel networks in the former would lead to high computational cost, and the latter requires off-line training. Grounded on the idea of contrastive learning, we propose a new semi-supervised framework, but avoid using two identical parallel networks, so that the computational cost could not be increased.

D. Difference From Previous Works

Different from the aforementioned methods, we propose an end-to-end network that takes as input the front-view images from a monocular camera and generates multi-class semantic BEV grid maps in a semi-supervised manner. The most close work to ours is the method proposed by Yang et al. [8], but their network can only predict a binary map and their training process relies on supervised learning. In contrast, we design a semi-supervised structure, alleviating the requirement of the hand-labeled data during training. Moreover, our method can generate multi-class labels, not just binary labels.

III. The Proposed Network

A. The Overall Architecture

The motivation of this work is to generate semantic BEV grid maps from input front-view monocular images in an end-to-end manner. A contrastive learning-based semi-supervised learning framework with double branches is proposed and its network architecture is illustrated in Fig. 1. As we can see, our S2G2 takes as input both labeled and unlabeled front-view images. It consists of a feature extractor, a dual-attention view transformation (DVT) module, and a double branch generator (DBG). We employ the EfficientNet [23] as the backbone of the feature extractor to extract front-view features from the input images. This encoder is shared for both passive and active branches. The network first extracts front-view feature maps $\mathcal{F}_{front}$ from the input front-view images, then transforms the viewpoint from front-view to BEV through the DVT module. Two distinct BEV feature maps, $\mathcal{F}_I$ and $\mathcal{F}_C$, originating from the same input image, denoted as the homologous features, are generated by the dual-attention block in the DVT module. The contrastive learning strategy adjusts the network parameters by minimizing

[1] Our code and dataset are available at https://github.com/lab-sun/S2G2

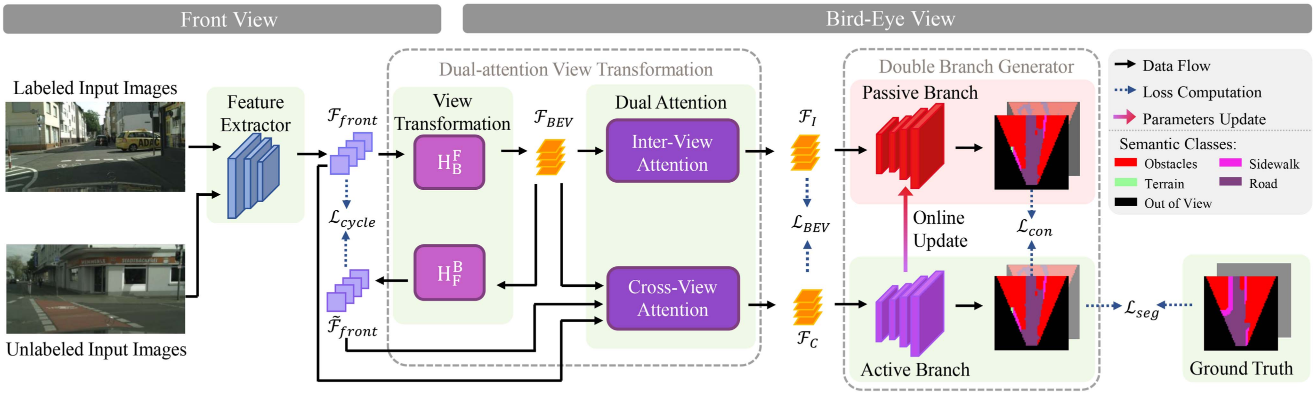

Fig. 1. The overall architecture of the proposed S2G2. We aim to generate semantic BEV grid maps in a semi-supervised manner. The network mainly consists of a feature extractor, a dual-attention view transformation (DVT) module, and a double branch generator (DBG). We feed both the labeled and unlabeled images into the feature extractor to get the front-view feature maps, $\mathcal{F}_{front}$. In DVT, based on cycle consistency, the front-view features are first projected into BEV features $\mathcal{F}_{BEV}$. Then the dual-attention block refines the BEV feature maps and produces a pair of homologous BEV feature maps, $\mathcal{F}_I$ and $\mathcal{F}_C$. The DBG takes as input the two BEV feature maps and produces the generation results. The semi-supervised learning is achieved by penalizing the consistency loss between the active and passive branches. In the DBG module, the active branch updates its weights by gradient descent. Based on the weights of the active branch, the weights in the passive branch are updated by using the exponential moving average (EMA) prediction. The modules involved in the training process are marked with green background. The module not involved in training is marked with red background. The solid lines represent the data flow in our network and the dotted lines represents loss computation. The figure is best viewed in color.

TABLE I
THE NUMBERS OF THE CHANNEL OF FRONT-VIEW FEATURE MAP $\mathcal{F}_{front}$ AFTER THE FEATURE EXTRACTION WITH DIFFERENT EFFICIENTNET VARIANTS AS THE BACKBONE, RANGING FROM EFFICIENTNET-B0 TO EFFICIENTNET-B7. EFFNET IS THE SHORT FOR EFFICIENTNET

	EffNet-B0	EffNet-B1	EffNet-B2	EffNet-B3
Channels	320	320	352	384
	EffNet-B4	EffNet-B5	EffNet-B6	EffNet-B7
Channels	448	512	576	640

the consistency loss between two identical networks. Therefore, the homologous features from the DVT module serve as the diversely augmented versions of one original image, which are the naturally suitable inputs for this semi-supervised scheme. With the semantic heads in both branches, a semantic BEV grid map can be generated.

B. The Feature Extractor

A pre-trained CNN model, EfficientNet [23], is used as our feature extractor. Different from the existing contrastive learning methods [20], we employ a shared encoder to extract the low-level features from both labeled and unlabeled images rather than using a parallel architecture with two identical networks, which results in a larger network model. With the increase of stages in EfficientNet, the receptive fields are enlarged, the backbone gradually reduces the feature-map resolution but increases the number of feature-map channels. The output feature map of the encoder is denote as $\mathcal{F}_{front}$. Since the EfficientNet has various variants, the number of channels of $\mathcal{F}_{front}$ can be different. Detailed channel numbers are display in Table I

C. The Dual-Attention View Transformation Module

To transform feature maps from front-view to BEV, we design the DVT module. According to the frontal features, this module predicts the corresponding feature maps in bird-eye view. The DVT module includes a view transformation block and a dual attention block. The former is designed to perform the view projection in a learning-based approach and the later will strengthen the transformed results.

1) View Transformation Block: Inspired by [24], the view transformation can be realized by training a transformation module H_B^F that transforms the feature maps from the front-view to BEV, $\mathcal{F}_{BEV} = H_B^F(\mathcal{F}_{front})$. Another transformation function H_F^B is the inverse of H_B^F. H_F^B transforms the $\mathcal{F}_{BEV}$ back to the front view, $\tilde{\mathcal{F}}_{front} = H_F^B(\mathcal{F}_{BEV})$. To train the view transformation block, a cycle consistency loss is introduced here:

$$\mathcal{L}_{cycle} = ||H_F^B\left(H_B^F\left(\mathcal{F}_{front}\right)\right) - \mathcal{F}_{front}||_1. \qquad (1)$$

Minimizing $\mathcal{L}_{cycle}$ encourages the re-transformed front-view feature map $\tilde{\mathcal{F}}_{front}$ to be similar to the original one, $\mathcal{F}_{front}$. The input of the module H_B^F is the front-view feature map and a corresponding BEV feature map is the output. Here, we use double convolutional layers to fit the transformation module H_B^F and H_F^B. The convolutional operation focuses on the local features and preserves the spatial information. The designed double-layer convolution enlarges the receptive field layer by layer until covering the whole input feature, $\mathcal{F}_{front}$. This could allow our view transformation block considering both local and global information during the view transformation.

2) Dual Attention Block: Based on the work [8], we design a dual attention block to improve the view transformation results. We keep the cross-view attention part unchanged, following [8], which takes $\mathcal{F}_{front}$, $\mathcal{F}_{BEV}$ and $\tilde{\mathcal{F}}_{front}$ as inputs to infer the attention score between the front-view and BEV. The cross-view attention emphasizes the relationship between the two different views. However, the internal relationship within the generated BEV feature map is also worth noting. Since the convolution layer can be seen as a feature extractor, the feature maps produced from multi-layer convolutions already gathered different kinds of features, stacking in the channel dimension. Moreover, the salient features are located differently in each feature layer

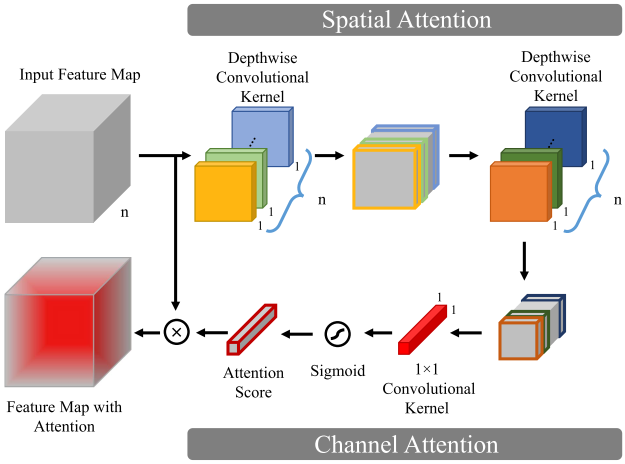

Fig. 2. The structure of the inter-view attention block. The block enables the network to focus on both spatial attention and channel attention of the input BEV feature map. The two types of attention are extracted sequentially. After normalization with the Sigmoid function, the attention score is multiplied with the input to get the final attention feature map.

at the spatial dimension. Therefore, we design an inter-view attention sub-block, which combines both channel and spatial attention to highlight the internal relationship of the BEV feature map, $\mathcal{F}_{BEV}$. The inter-view attention sub-block and cross-view attention sub-block together form the dual attention block and complement to each other.

Fig. 2 demonstrates the proposed inter-view attention sub-block. In this sub-block, we first perform n depthwise convolution kernels [25] on n-channel input feature map separately without changing the depth. This operation reduces the resolution of the feature map and the salient features located in the spatial dimension can be learned as training. We repeat depthwise convolution twice in our proposed S2G2. An 1×1 convolution is applied to the intermediate features sequentially, which exploits the channel relationship of the feature maps. Both types of convolution are followed with a max-pooling operation. In short, the inter-view attention is computed as:

$$S_A = \text{sigmoid}\{C_1[C_d(\mathcal{F}_{in})]\}, \tag{2}$$

$$\mathcal{F}_{out} = \mathcal{F}_{in} \otimes S_A, \tag{3}$$

where C_d and C_1 denote the depthwise convolution and 1×1 convolution respectively. After two successive attention extraction, a Sigmoid function is applied to map the convolution output into the range from 0 to 1. Then, an attention score, S_A is produced. $\otimes$ represents element-wise multiplication. Through multiplication, the internal attention is spread into the input feature map.

In order to make the inter-view attention and cross-view attention complementary, we introduce a cross entropy loss function, $\mathcal{L}_{BEV}$, between the outputs of the separate attention block. The DVT module maintains the same dimension in the input and output feature maps, so it can be inserted into any existing network seamlessly.

D. The Double Branch Generator

Grounded on the contrastive learning strategy, we propose a double branch generator, which is composed of an active branch and a passive branch to implement semi-supervised learning. The main idea behind contrastive learning is that similar data are clustered together and different data are pushed away. This assumes that the network should generate consistent outputs, given similar inputs. In such a way, the unlabeled data can be utilized to boost the training process. Therefore, the performance of the contrastive learning-based semi-supervised methods relies largely on the generation of the homologous data. The existing contrastive learning approaches perform strong data augmentation combinations, as such Mixup [26], Cutout [27], and CutMix [28] to generate diverse versions of the same data. In our work, the output feature map is not aligned with the input image due to the view transformation task. Therefore, those data augmentation techniques that require the alignment between the input and output, do not apply to our network.

With the dual attention block, we get two outputs, $\mathcal{F}_I$ and $\mathcal{F}_C$. The two feature maps concentrate on the inter-view relationship and cross-view relationship, respectively. The two attention-included outputs originate from the same input but differ from each other. Therefore, we take them as inputs for the active branch and passive branch, naturally. To endow the network with the ability to output similar predictions for the similar inputs, we introduce a consistency loss, $\mathcal{L}_{con}$ that calculates the differences between the predictions from the active branch and the passive branch with mean squared error. The consistency loss can be written as:

$$\mathcal{L}_{con} = ||P_{act}(\mathcal{F}_c, \omega_a) - P_{pas}(\mathcal{F}_I, \omega_p)||_2, \tag{4}$$

where $P_{act}(\cdot)$ and $P_{pas}(\cdot)$ are the predictions from active branch and passive branch. The weights for the two branches are ω_a and ω_p, respectively.

At each training step, the active branch updates via the gradient descent from the weighted sum of the segmentation loss, $\mathcal{L}_{seg}$ and consistency loss, $\mathcal{L}_{con}$. We define the segmentation loss as the cross-entropy loss for the labeled images. The consistency loss is used for both labeled and unlabeled images. The weights in passive branch (ω_p) are updated with an Exponential Moving Average (EMA) strategy instead of the gradient descent manner, which is formulated as:

$$\omega_p^i = \lambda\omega_p^{i-1} + (1 - \lambda)\omega_a^i, \tag{5}$$

where the superscript i represents the *i-th* training step. λ is a hyperparameter for EMA decay, and it is set as 0.999 in our experiment.

E. Loss Function

We add losses from different modules together. We train our S2G2 in an end-to-end manner. The overall loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{seg} + \alpha\mathcal{L}_{cycle} + \beta\mathcal{L}_{BEV} + \gamma\mathcal{L}_{con}, \tag{6}$$

where $\mathcal{L}_{seg}$ is the major loss for our network. α, β, and γ are the weighted coefficient to balance each loss. In practice, we empirically set α, β both equal to 1, and let consistancy weight, γ, be adjusted in a self-adaption way. We will discuss the details of those hyperparameters in the experiment section.

IV. Experimental Results and Discussions

A. The Dataset

We conducted our experiments using the dataset from MonoOccupancy [9]. This dataset includes 2600, 375 and 500 labeled images for training, validation and test. The dataset preserves the input images of the public dataset, Cityscapes [29]. But the authors create their own semantic BEV ground truth via semi-global matching (SGM) method [30], using the disparity maps provided by Cityscapes. The ground truth contains 4 semantic classes, which are obstacles, sidewalk, terrain, and road. To evaluate the semi-supervised architecture, we build three groups with different ratios of unlabeled images. The ratios of the unlabeled images are 10%, 40% and 80%. Note that we use all the labeled images in the training set. The input images are normalized to 256×512 and the output size is 64×64.

We randomly shuffle the input training data before feeding them to the network. Because the input image and generated semantic maps are in different perspectives, we apply random flipping and random brightness changing to perform the data augmentation, maintaining the relative position of the content in the images.

B. Training Details

The training is performed with an NVIDIA GeForce RTX 3060 GPU. Due to the limited memories of the GPU, we set the batch size to 4, which consists of 2 labeled images and 2 unlabeled images. We train our network for 200 epochs with the stochastic gradient descent (SGD) optimizer. The momentum and weight decay are set to 0.9 and 1×10^{-4}, respectively. We initialize the learning rate to 5×10^{-5} and adopt an exponential decay scheme to adjust the learning rate as training. The decay coefficient of the learning rate is 0.995.

Specifically, we adopt EfficientNet-B4 with the pre-trained weights as our backbone, the rest of our network parameters are randomly initialized. We employ a ramp-up tuning scheme to adjust the consistency weight, γ, following the practice of the Mean Teacher [20]. The ramp-up scheme ensures that γ increases to the set value gradually until the end of the ramp-up phase. Through an extensive ablation study, we set the consistency weight, γ, to 1 and the ramp-up step to 100. The selection details will be elaborated in the ablation study.

C. Ablation Study

We conduct several ablation experiments to verify the effectiveness of the structure and parameters used in our S2G2. To assess the performance of our S2G2, mIoU and mAP are adopted as evaluation metrics. According to the different amounts of the unlabeled images in the training set, we conduct our ablation experiments with 3 different sets, which respectively include 10%, 40%, and 80% of the unlabeled images.

1) Ablation on the Feature Extraction Module: In our S2G2, the implementation of semi-supervised learning depends on a double branch generator which enlarges the scale of network parameters, compared with the fully supervised network. Therefore, to make our network be efficient and effective in terms of training speed and memory usage, a compact and powerful backbone should be selected for the feature extraction module. EfficientNet is a network that focuses both on accuracy and efficiency. The EfficientNet family includes 8 variants, which are named as EfficientNet-B0 to EfficientNet-B7, respectively. Those variants are different from each other in the depth, width, and resolution.

TABLE II
The Ablation Study Results (%) of the Variants of the EfficientNet Family. According to the Different Amounts of the Unlabeled Images in the Training Set, We Conduct Our Ablation Study Into 3 Groups, Which Contain 10%, 40%, and 80% Unlabeled Images, Respectively

Variants	10%		40%		80%	
	mIou	mAP	mIou	mAP	mIou	mAP
EfficientNet-B0	0.5468	0.6410	0.5834	0.7006	0.5795	0.6941
EfficientNet-B1	0.5847	0.6745	0.5703	0.6635	0.5863	0.7066
EfficientNet-B2	0.5774	0.6766	0.5807	0.6902	0.5839	0.7020
EfficientNet-B3	0.5840	**0.7112**	0.5747	0.6784	0.5877	0.7099
EfficientNet-B4	**0.5894**	0.7003	0.5889	0.6956	**0.5879**	**0.7110**
EfficientNet-B5	0.5852	0.7044	**0.5890**	0.6954	0.5830	0.7098
EfficientNet-B6	0.5854	0.7046	0.5880	**0.7162**	0.5879	0.6952
EfficientNet-B7	0.5794	0.6959	0.5852	0.6937	0.5835	0.6928

In the ablation study, we first compare the performance of the proposed network with different EfficientNet variants. In our network, we only keep the feature extraction part of the EfficientNet and remove the average pooling of the last layer, as well as the classification head. The modified EfficientNet variants produce a front-view feature map with a fixed size of 8×16 but with diverse numbers of channels. The different channels of the output feature map are listed in Table I.

Table II displays the results of the ablation study on the different EfficientNet variants, including EfficientNet-B0 to EfficientNet-B7. The obvious raising trends can be seen when increasing the complexity of EfficientNet from B0 to B4. But after B4, the mIou and mAP of the prediction performance stay in a relatively stable range. To trade off performance and computation cost, we select EfficientNet-B4 as the backbone of our S2G2.

2) Ablation on the Dual-Attention Block: To verify the effectiveness of the dual-attention block, we conduct two groups of tests with and without a certain attention block. We first only keep the inter-view attention block and discard the cross-view attention block. We term this variant as Only Inter-View Attention (OIVA). Then, we remove the inter-view attention block instead and get the Only Cross-View Attention (OCVA) variant. According to the results of the previous ablation study, our network gets the best performance with EfficientNet-B4. But EfficientNet-B7 is the most complicated variant with the most number of parameters, it should perform better in OIVA or OCVA. So, we chose the B4 and B7 variants as our feature extraction module to conduct this ablation study.

Moreover, we also exchange the input order to the Double Branch Generator module, which leads to two different structures. The first one takes the output feature from the inter-view attention, $\mathcal{F}_I$ as the input of passive branch and we denote this variant as S2G2-IPCA (Inter-view attention feature map for Passive branch and Cross-view attention feature map for Active branch). Note that S2G2-IPCA is the same as the proposed S2G2. For the second one, we let the feature map, $\mathcal{F}_C$ be the input of passive branch and term this as S2G2-CPIA (Cross-view attention feature map for Passive branch and Inter-view attention feature map for Active branch).

TABLE III
The Ablation Study Results (%) on Dual-Attention Block. OIVA Stands for the Variant That Only Keeps the Inter-View Attention Sub-Block and OCVA Means the Module That Only Have the Cross-View Attention Sub-Block. B4 and B7 Present the Experiments are Conducted With the EfficientNet-B4 and EfficientNet-B7 as Their Backbone

Variants	10%		40%		80%	
	mIou	mAP	mIou	mAP	mIou	mAP
OIVA(B4)	0.5610	0.6632	0.5371	0.5750	0.5597	0.6725
OCVA(B4)	0.5277	0.6820	0.5372	0.6551	0.5605	0.6673
S2G2-B4	**0.5894**	**0.7003**	**0.5889**	**0.6956**	**0.5879**	**0.7110**
OIVA(B7)	0.5327	0.5957	0.5467	0.6368	0.5593	0.6550
OCVA(B7)	0.5425	0.4809	0.5308	0.5793	0.5471	0.6173
S2G2-B7	**0.5794**	**0.6959**	**0.5852**	**0.6937**	**0.5835**	**0.6928**

TABLE IV
The Ablation Study Results (%) on the Different Input Orders to the Finial Double Branch Generator. S2G2-CPIA Module Feeds the Cross-View Attention Feature Map, $\mathcal{F}_C$, to the Passive Branch and the Inter-View Attention Feature Map, $\mathcal{F}_I$, to the Active Branch. S2G2-IPCA is the Opposite Version of S2G2-CPIA

Variants	mIoU	mAP
S2G2-CPIA	0.5695	0.6657
S2G2-IPCA	**0.5894**	**0.7003**

TABLE V
The Comparative Results (%) on the Baseline Methods. The Various Semantic Segmentation Methods are Integrated Into the Mean Teacher Framework to Perform Semi-Supervised Learning. The Random Gaussian Noise is Added to the Input Images Before Fed Into the Separate Networks. The Bold Font Highlight the Best Results in Each Column. Our Proposed S2G2 Outperforms the Others

Methods	10%		40%		80%	
	mIoU	mAP	mIoU	mAP	mIoU	mAP
SegNet	0.5084	0.6138	0.5224	0.5205	0.5184	0.5561
U-Net	0.4413	0.5417	0.4554	0.5050	0.4535	0.4996
RTFNet	0.5256	0.6343	0.5346	0.5985	0.5258	0.5801
HRNet	0.5539	0.6670	0.5568	0.5970	0.5535	0.5938
DeepLab V3+	0.5144	0.6060	0.5145	0.5808	0.5024	0.5923
MonoOccupancy	0.5262	0.6787	0.5338	0.6575	0.5329	0.6148
S2G2 (ours)	**0.5894**	**0.7003**	**0.5889**	**0.6956**	**0.5879**	**0.7110**

Table III demonstrates the effectiveness of the dual-attention block. From the table, the dual-attention that combines the inter-view and cross-view attention together, gets a superior performance against its counterparts. Table IV shows the comparative results of different input orders to the active branch and passive branch. The results indicate that the IPCA variant has a higher mIoU with 58.94%, compared with the CPIA variant, 56.95%. This is also true for the metric mAP.

3) Ablation on the Double Branch Generator Module: For the double branch generator module, we test different sets of parameters to check the impacts on the intensity of contrastive learning. Specifically, the consistency loss is linked to the generation of similar outputs from the active branch and passive branch. The learning effectiveness of the unlabeled images of the network is affected by the consistency loss-related hyperparameters, including consistency weight γ and the ramp-up step. Therefore, we set 6 different values for the consistency weight (0.05, 0.1, 0.5, 1, 5, 10) and 5 different ramp-up step (50, 75, 100, 125, 150) in this experiment.

TABLE VI
The Comparative Results (%) on the Test Dataset From [9]. All the Comparative Methods Predict the Semantic BEV Map in a Supervised Manner. The Table Shows That Our Semi-Supervised Approach Achieves the Best Performance

Methods	mIoU	mAP
PYVA [8] (CVPR 2021)	0.5066	0.6219
PON [17] (CVPR 2020)	0.4883	0.6332
MonoLayout [19] (WACV 2020)	0.5307	0.6776
MonoOccupancy [9] (RA-L 2019)	0.5786	0.6513
S2G2 (ours)	**0.5886**	**0.7023**

The process of parameter tuning is presented in Fig. 3. We find that when the weight of the consistency loss equals to 1, and the ramp-up step is set as 100, the network produces the best performance. These two parameters impose an effect on how well the network can learn from the unlabeled data. We also find that a small consistency loss weight and a big ramp-up step can lead to insufficient contrastive learning due to less punishment towards the consistency loss. This means that the network could not predict the consistent outputs for the homologous features. But large weight value and too quick ramp-up may force the assimilation of the two branches, still resulting in a deficient learning capacity.

D. Comparative Results

1) Comparison With Baseline Methods: As our S2G2 is the first method that generates the semantic BEV grid map in a semi-supervised manner, we develop several baseline methods to perform the comparative experiments. Our target outputs are still in the image domain, so to form our baselines, we take advantage of the popular semantic segmentation methods, including U-Net [11], DeepLab V3+ [12], RTFNet [31], SegNet [10], HRNet [32]. Other than those semantic segmentation methods, we also take the MonoOccupancy [9] as consideration. However, the above methods are all trained in a supervised manner. In order to train the networks with the unlabeled images, we integrate the mentioned segmentation networks into the Mean Teacher [20] framework, which is originally designed for semi-supervised classification.

Specifically, we modify the aforementioned methods by adding an aspect-ratio changing layer and adjusting the output size of their decoders, because the input resolution (256×512) is not the same as that of the output (64×64). The Mean Teacher framework depends on the two identical networks to perform contrastive learning. So we duplicate the aforementioned methods as the two parallel networks in the Mean Teacher framework. To follow the idea of contrastive learning, we apply the random Gaussian noise to make the input image a pair of homologous similar ones. Then the noise-injected images are fed into the two networks of the Mean Teacher framework, respectively.

We report the quantitative comparative results for the baseline methods in Table V. The results show that our proposed S2G2 achieves the best performance in terms of mIoU and mAP across all the networks. From the table, we can see that the MonoOccupancy gets the second-best results. MonoOccupancy is also a semantic BEV grid map generator.

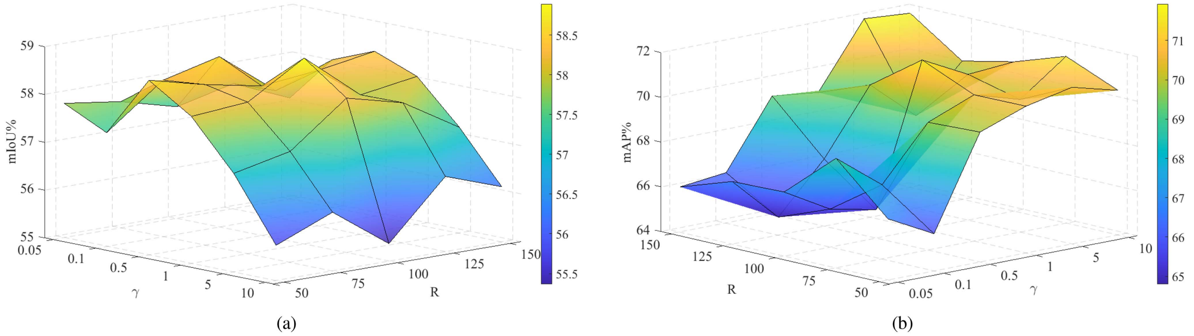

Fig. 3. Impacts of the ramp-up steps (R) and the weighted coefficient of consistency loss (γ) on the mIoU and mAP. We take the training group with 40% unlabeled images as example. We set 0.05, 0.1, 0.5, 1, 10 as γ and 50, 75, 100, 125 and 150 as R in different tests. (a) is the mIoU result of different settings. (b) is the mAP performance. The figure is best viewed in color.

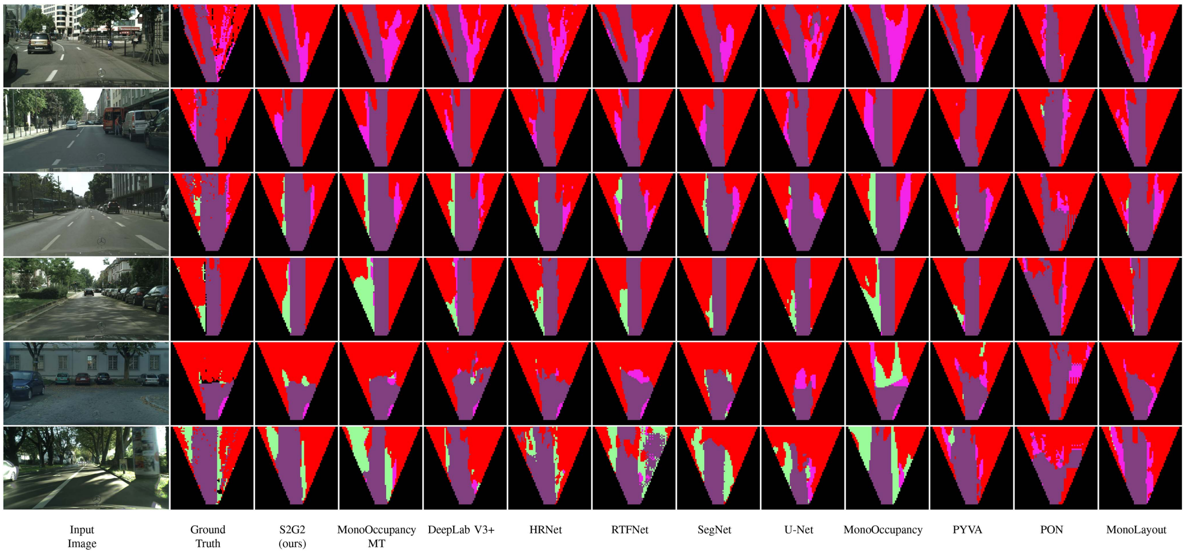

Fig. 4. Example qualitative performances for the semantic BEV grid-map generation networks. Every row shows the results for different networks testing with the same input images. MonoOccupancy in the mean teacher framework is marked with MT to distinguish it from the original one. Note that the ground truth may contain noise. The last two rows show the unusual driving condition and the road under strong uneven sunlight. The comparative results demonstrate the precision and superiority of our network. The figure is best viewed in color.

We conjecture the reason for the inferior performance of MonoOccupancy is the flattened operation in its supervised variational automatic encoder (VAE) structure, which converts the 2D feature map into 1D vector, dropping the spatial information. Although the generated semantic BEV grid map is a form of an image, the better performance of our proposed S2G2 and MonoOccupancy indicates that there are still great gaps laying between the tasks of semantic BEV grid map generation and the classical semantic segmentation. Therefore the semantic BEV grid map generation needs a specially designed structure.

2) Comparison With the State-of-The-Art Methods: We also evaluate the performances of our S2G2 together with some of the state-of-the-art supervised learning-based methods, including PYVA [8], PON [17], MonoLayout [19], and MonoOccupancy [9]. It can be seen from Table VI, testing on the dataset provided by MonoOccupancy, our proposed S2G2 outperforms all the previous networks with 58.86% in mIoU and 70.23% in mAP. The second best results were produced by MonoOccupancy. We attribute it to the fact that the other methods could not adapt well to the noisy ground truth provided by the training dataset.

Moreover, we compared the performance of the MonoOccupancy in the mean teacher framework and the original one. We find that the results of the former one are inferior to the latter. The reason for this case may be that the semi-supervised semantic generation requires strong perturbations to produce a qualified homologous similar input pair. We refer readers to [21] for more details.

3) The Qualitative Demonstrations: Sample qualitative semantic BEV generation results are shown in Fig. 4. In general, our S2G2 generates a more precise and clear semantic BEV grid map. Note that the ground truth contains noise since they are produced by the SGM method. Even so, our S2G2 can still generate the compelling semantic BEV map. According to the results, we can see that our S2G2 is more sensitive to obstacles compared to the other methods, which is crucial for safe navigation. The last two rows display the more complicated driving environments due to the unusual road conditions and the uneven sunlight. Most other methods fail to predict the correct semantic classes, but our S2G2 still provides relatively clear and accurate semantic boundaries.

V. Conclusions and Future Work

Semantic BEV grid map is a kind of promising data representations for semantic environment perception in autonomous driving. We presented here a novel semi-supervised semantic BEV grid-map generation network that takes as input both labeled and unlabeled front-view images from a monocular camera and directly outputs semantic BEV grid maps. The proposed network can be trained in an end-to-end manner. We conducted extensive ablation experiments to determine the appropriate architecture and hyperparameters. The network was evaluated and tested on a real-world dataset. We demonstrated its superiority over several semi-supervised baseline methods. Although our S2G2 can generate satisfying semantic results, our method is still limited by the narrow field of view, which means that our method can only generate cone-shaped semantic grid map at the current stage. In the future, we will use sequential multiple images as input to generate grid maps with 360° full view. In addition, we plan to enhance our semantic BEV generation network to produce more information, for example, the graph structure of road layouts.

References

[1] P. Cai, H. Wang, Y. Sun, and M. Liu, "DIGNet: Learning scalable self-driving policies for generic traffic scenarios with graph neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 8979–8984.

[2] H. Wang, P. Cai, Y. Sun, L. Wang, and M. Liu, "Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 13731–13737.

[3] H. Wang, P. Cai, R. Fan, Y. Sun, and M. Liu, "End-to-end interactive prediction and planning with optical flow distillation for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 2229–2238.

[4] Ö. Erkent, C. Wolf, and C. Laugier, "Semantic grid estimation with occupancy grids and semantic segmentation networks," in *Proc. IEEE 15th Int. Conf. Control, Automation, Robot. Vis.*, 2018, pp. 1051–1056.

[5] S. Richter, J. Beck, S. Wirges, and C. Stiller, "Semantic evidential grid mapping based on stereo vision," in *Proc. IEEE Int. Conf. Multisensor Fusion Integration Intell. Syst.*, 2020, pp. 179–184.

[6] Y. Zhou, Y. Takeda, M. Tomizuka, and W. Zhan, "Automatic construction of lane-level HD maps for urban scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 6649–6656.

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

[8] 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.*, 2021, pp. 15536–15545.

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

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, Springer, 2015, pp. 234–241.

[12] 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.*, 2018, pp. 801–818.

[13] L. Sless, B. El Shlomo, G. Cohen, and S. Oron, "Road scene understanding by occupancy grid learning from sparse radar clusters using semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 867–875.

[14] B. Yang, R. Guo, M. Liang, S. Casas, and R. Urtasun, "RadarNet: Exploiting radar for robust perception of dynamic objects," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 496–512.

[15] S. T. Isele, F. Klein, M. Brosowsky, and J. M. Zöllner, "Learning semantics on radar point-clouds," in *Proc. IEEE Intell. Veh. Symp.*, 2021, pp. 810–817.

[16] R. Van Kempen, B. Lampe, T. Woopen, and L. Eckstein, "A simulation-based end-to-end learning framework for evidential occupancy grid mapping," in *Proc. IEEE Intell. Veh. Symp.*, 2021, pp. 934–939.

[17] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11138–11147.

[18] I. Dwivedi, S. Malla, Y.-T. Chen, and B. Dariush, "Bird's eye view segmentation using lifted 2D semantic features," in *Proc. Brit. Mach. Vis. Conf.*, 2021, pp. 6985–6994.

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

[20] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1195–1204.

[21] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," 2019, *arXiv:1906.01916*.

[22] Y. Zou et al., "PseudoSeg: Designing pseudo labels for semantic segmentation," 2020, *arXiv:2010.09713*.

[23] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 6105–6114.

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[25] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807.

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv: 1710.09412*.

[27] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[28] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6022–6031.

[29] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[30] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

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

[32] K. Sun et al., "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*.