2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10611420

Semi-Supervised Learning for Visual Bird's Eye View Semantic Segmentation

Junyu Zhu[1*], Lina Liu[2*], Yu Tang[3], Feng Wen[3], Wanlong Li[3†] and Yong Liu[1†]

***Abstract*— Visual bird's eye view (BEV) semantic segmentation helps autonomous vehicles understand the surrounding environment only from front-view (FV) images, including static elements (e.g., roads) and dynamic elements (e.g., vehicles, pedestrians). However, the high cost of annotation procedures of full-supervised methods limits the capability of the visual BEV semantic segmentation, which usually needs HD maps, 3D object bounding boxes, and camera extrinsic matrixes. In this paper, we present a novel semi-supervised framework for visual BEV semantic segmentation to boost performance by exploiting unlabeled images during the training. A consistency loss that makes full use of unlabeled data is then proposed to constrain the model on not only semantic prediction but also the BEV feature. Furthermore, we propose a novel and effective data augmentation method named conjoint rotation which reasonably augments the dataset while maintaining the geometric relationship between the FV images and the BEV semantic segmentation. Extensive experiments on the nuScenes dataset show that our semi-supervised framework can effectively improve prediction accuracy. To the best of our knowledge, this is the first work that explores improving visual BEV semantic segmentation performance using unlabeled data. The code is available at https://github.com/Junyu-Z/Semi-BEVseg.**

I. Introduction

Bird's eye view (BEV) semantic segmentation is a powerful representation of the surrounding environment, which can assist mobile robots such as autonomous vehicles in perceiving the surroundings of static road layouts and dynamic objects(e.g., vehicles, pedestrians). With rich information and absolute scales, BEV semantic segmentation can directly connect with many downstream tasks, such as path planning and motion control. Recently, vision-based methods [1], [2], [3], [4], [5] that infer BEV semantic segmentation only from cameras have been developed to reduce the cost of sensors.

A visual BEV semantic segmentation model generally consists of three components [6]: a backbone network as a visual feature extractor, a view transformer module for getting the BEV feature from the front-view (FV) feature, and a segmentation decoder to predict semantic segmentation from the BEV feature. And most of the existing BEV semantic segmentation methods are full-supervised, mainly focusing on exploring new view transform approaches [7], [1], [8], integrating temporal cues [9], [10], and designing more complex segmentation decoders [3], [11]. However, these methods rely heavily on the accessibility and quantity of labeled data that needs high costs for constructing HD maps, annotating 3D object bounding boxes, and capturing camera extrinsic parameters. Compared with annotation, collecting unlabeled images requires less labor. Therefore, in this work, we are motivated to study semi-supervised learning based BEV semantic segmentation from monocular images to boost the performance by exploiting unlabeled data.

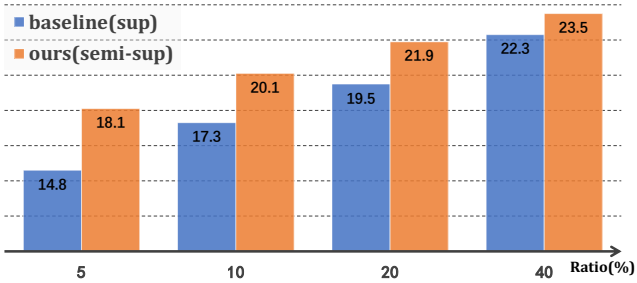

Fig. 1. mIoU(%) on the nuScenes dataset between our semi-supervised framework and supervised baseline using different label ratios.

While many works have explored semi-supervised learning for conventional 2D semantic segmentation, semi-supervised visual BEV semantic segmentation is rather underexplored. Following the common consistency regularization in semi-supervised learning, we propose a consistency loss that restricts the model with perturbations on unlabeled images. Inspired by [12], in addition to semantic segmentation consistency, we use additional consistency of BEV feature for further improvement. And to excavate the spatial consistency of the BEV feature, we use horizontal flipping as the perturbation rather than color jitter which is typical for semi-supervised 2D semantic segmentation.

Apart from using the above consistency regularization on unlabeled data, we also explore improving the quantity and diversity of the dataset for better performance. Although several well-designed and effective data augmentation methods [13], [14] have been proposed for 2D/3D semantic segmentation, there is no relevant research in the visual BEV semantic segmentation field. Unlike pixel-aligned 2D/3D semantic segmentation, the complex geometric relationship of the projection between FV images and BEV semantic segmentation maps makes data augmentation harder. Through geometric intuition and mathematical analysis, we propose a novel data augmentation method called conjoint rotation for this task. And it benefits not only our semi-supervised framework but also the full-supervised model.

Following the conventions in semi-supervised tasks, we conduct experiments on nuScenes [15] dataset with different

[1]Junyu Zhu and Yong Liu are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China. E-mail:junyuzhu@zju.edu.cn, yongliu@iipc.zju.edu.cn.
[2]Lina Liu is with China Mobile Research Institute, Beijing, China. E-mail: liulina0601@gmail.com.
[3]Wanlong Li, Yu Tang, and Feng Wen are with Noah's Ark Lab, Huawei Technologies, Beijing, China. E-mail:{liwanlong, tangyu17, wenfeng3}@huawei.com.
*Equal contribution. †Corresponding author.

ratios of labels and demonstrate that our semi-supervised framework can effectively improve performance by relatively >10% on average with the unlabeled data as shown in Fig. 1. Moreover, extensive ablation studies are also conducted to prove the effectiveness of each component. We hope this work can be a stepping-stone for future research in this field.

To summarize, our main contributions are as follows:

- We dig into visual BEV semantic segmentation with limited labels and offer the first semi-supervised BEV semantic segmentation framework that enhances the performance using unlabeled data.
- We propose a consistency loss exploiting unlabeled data to restrict the model on semantic segmentation and the BEV feature.
- We design a novel data augmentation method for visual BEV semantic segmentation, and it works well on our semi-supervised framework and full-supervised model.
- The proposed framework achieves relatively >10% average improvements over the full-supervised baseline on the nuScenes.

II. Related works

A. Visual BEV Semantic Segmentation

Visual BEV semantic segmentation is a task of using FV images to predict BEV semantic segmentation. Via homography transformations, [16], [17] use inverse perspective mapping (IPM) to map FV images/features onto the BEV plane. This approach relies heavily on the plane hypothesis, so it easily fails for objects that lie above the BEV plane, such as cars and pedestrians. VED [5] uses the fully-connected bottleneck layer to realize the feature transformation from the front view to the BEV. Due to the lack of available ground truth data, early methods rely on various weak supervision. And with the emergence of the nuScenes dataset [15] that contains HD maps, 3D object bounding boxes, and much image data from six calibrated cameras in different scenes, visual BEV semantic segmentation develops rapidly. Based on view transformation (VT) strategies, different methods can be divided into the following categories:

MLP-based VT [1], [3], [2] is based on the geometric correspondence between the vertical lines in the image and polar rays in BEV. 2D-to-3D-based VT [18], [7] gets BEV feature by explicit or implicit depth estimation. 3D-to-2D-based VT [19], [10], [20] projects 3D points from the BEV plane onto the 2D image plane to get corresponding features. Transformer-based VT [4], [21], [22], [10] is another ready solution for transforming features from the front view to the BEV by implicit geometric reasoning.

Although impressive results have been achieved by recent fully-supervised methods, requiring time-consuming and laborious labeling is a common shortcoming. Gao et al. [23] present a framework that can be trained with both labeled and unlabeled data but fails to improve performance with unlabeled data. And their work focuses on estimating road layout but no dynamic elements. In this work, under a more challenging setting, we dig into underexplored semi-supervised learning in visual BEV semantic segmentation to improve performance by exploiting unlabeled data.

B. Semi-Supervised 2D Semantic Segmentation

Inspired by the progress of semi-supervised learning in the image classification field, semi-supervised semantic segmentation for the 2D image has been explored by many works in these years. To force the decision boundary to lie in the low-density area, many works [24], [25], [26], [27] utilize a common strategy, consistency regularization. Pseudo-labeling [28], [29], [30] is another effective technique.

In this work, we apply consistency regularization to the semi-supervised visual BEV semantic segmentation task and propose a consistency loss that acts on both semantic segmentation and the BEV feature.

C. Data Augmentation

Data augmentation is a practical approach to improving generalization ability and has been explored in many fields, including image classification [31], [32], 3D point cloud semantic segmentation [14], and 2D semantic segmentation [13]. In the visual BEV semantic segmentation field, there is no relevant work currently, to the best of our knowledge.

In this work, out of geometric intuition and mathematical analysis, we propose a new data augmentation named conjoint rotation that is effective for this task.

III. Method

For the visual BEV semantic segmentation task, we need to predict a semantic segmentation map Y from the given FV image I with its corresponding camera intrinsic matrix K. In this paper, each pixel of $Y \in p^{C\times Z\times X}$ describes the probability of C categories, such as drivable area, walkway, pedestrian, and car. And different types in a BEV semantic map may appear in the same pixel, which is different from the setting of some existing works [7], [10], [4]. Under the semi-supervised setting, the training set consists of a labeled set $D_L = \{(I_L^0, K_L^0, \hat{Y}^0), (I_L^1, K_L^1, \hat{Y}^1), ..., (I_L^i, K_L^i, \hat{Y}^i), ...\}$ and an unlabeled set $D_U = \{(I_U^0, K_U^0), (I_U^1, K_U^1), ..., (I_U^i, K_U^i), ...\}$. And we aim to exploit $D_L \cup D_U$ to train a model that performs better than only trained on D_L. An overview of the proposed framework is illustrated in Fig. 2.

In this work, we follow the Mean Teacher [33] that is originally proposed for image classification and extend it to the more taxing task of visual BEV semantic segmentation. We design a segmentation consistency loss L_{sc} and a feature consistency loss L_{fc} for consistency regularization. Furthermore, we propose a novel data augmentation method called conjoint rotation to improve performance.

In the following subsections, we first give a brief introduction to the visual BEV semantic segmentation model in Sec. III-A. In Secs. III-B,III-C,III-D, we successively describe the supervised loss L_{sup}, segmentation consistency loss L_{sc}, and BEV feature consistency loss L_{fc}. And we present our proposed conjoint rotation in Sec. III-E. Finally, Sec. III-F summarizes the training process.

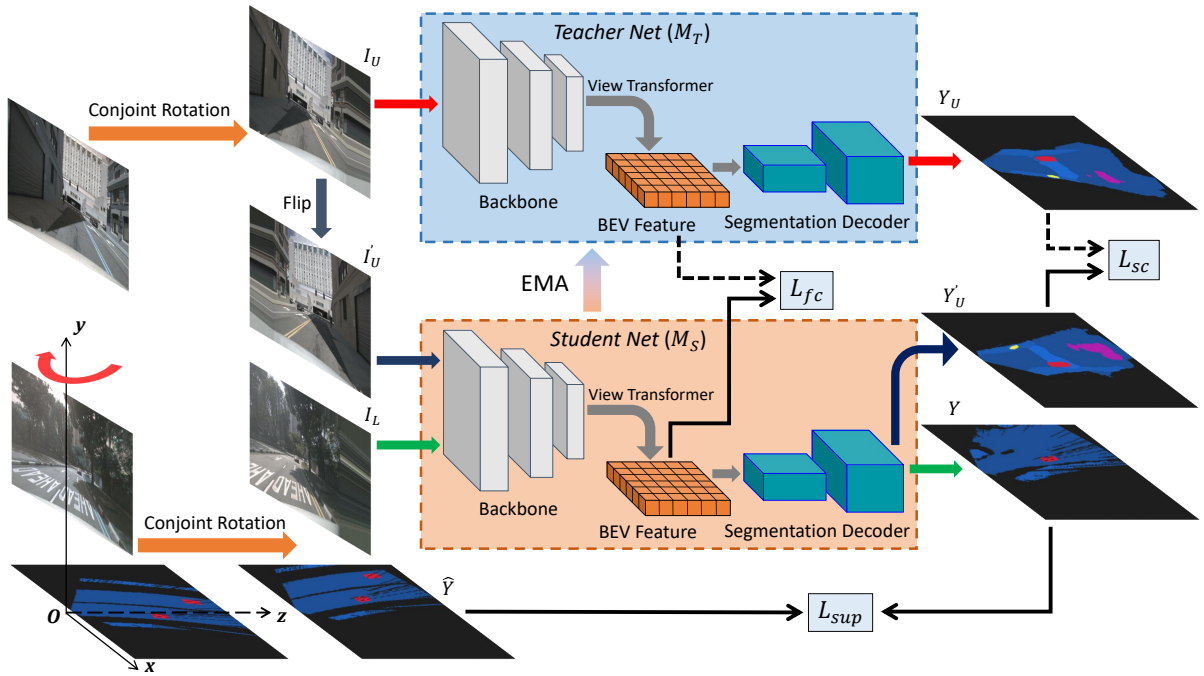

Fig. 2. Framework overview. By our proposed conjoint rotation, the labeled and unlabeled data are first augmented to get I_L, $\hat{Y}$, and I_U. Immediately after that, Y and Y_U are predicted by the Student Net M_S and Teacher Net M_T, respectively. Meanwhile, M_S predicted $Y_U^{'}$ from flipped image $I_U^{'}$. Note that the view transformer of M_S and M_T needs the camera intrinsic matrix K as input, and K would also be changed when flipping the image. The feature consistency loss L_{fc} is computed from the L2 loss of BEV features of I_U and $I_U^{'}$. And the segmentation consistency loss L_{sc} is computed from the L2 loss of BEV semantic segmentation, Y_U and $Y_U^{'}$. Also, the supervised loss L_{sup} is computed between Y and $\hat{Y}$. After M_S is updated with gradient descent using the above losses, M_T is updated as an exponential moving average (EMA) of M_S. The Teacher Net can perform better than the Student Net after the training with proper hyper-parameters.

A. Visual BEV semantic segmentation Model

Generally, a visual BEV semantic segmentation model M first uses a backbone network to extract the FV feature from the given FV image $I \in R^{3\times H\times W}$. The model gets the BEV feature from the FV feature through a view transformer that is usually related to the camera intrinsic matrix $K \in R^{3\times 3}$. Finally, using a segmentation decoder, the model predicts BEV semantic segmentation $Y \in p^{C\times Z\times X}$ from the BEV feature. In this work, our framework uses two models with the same structure called Teacher Net M_T and Student Net M_S. Their parameters are separately randomly initialized except the pretrained backbone network, and the M_T performs better after the training process.

B. Supervised Loss

Following state-of-the-art methods [9], [3], [2], we use the same Dice loss as the supervised loss L_{sup} for labeled data. The L_{sup} is defined over C classes and N pixels:

$$L_{sup} = 1 - \frac{1}{|C|}\sum_{k=1}^{C}\frac{2\sum_i^N \hat{y}_i^k y_i^k}{\sum_i^N \hat{y}_i^k + y_i^k + \epsilon}. \tag{1}$$

where $\hat{y}_i^k$ is the target binary variable grid cell of $\hat{Y}$, y_i^k is the predicted probability variable of Y, and ϵ is set as 1e-5 to prevent the denominator from being zero.

C. Segmentation Consistency Loss

For the unlabeled data, we calculate the segmentation consistency loss L_{sc} between semantic segmentations, Y_U and $Y_U^{'}$, from the unlabeled data $\{I_U, K_U\}$ and the horizontally flipped version $\{I_U^{'}, K_U^{'}\}$. The Y_U and $Y_U^{'}$ are the output class probability matrixes after the last sigmoid function. Because of the geometric relationship between the FV and the BEV, the consistency of BEV semantic segmentation of the original image and the flipped one is natural. Using L2 distance $\|\cdot\|_2$ and horizontally flipping operation Φ, the L_{sc} is formulated as:

$$L_{sc} = \|\widetilde{Y}_U - \Phi(Y_U^{'})\|_2. \tag{2}$$

where $\widetilde{Y}_U$ means that the gradient of Y_U is detached.

D. BEV Feature Consistency Loss

In BEVDet [12], Huang et al. conducted common 2D augmentation operations, including random flipping, scaling and rotating on both the BEV feature and the 3D object detection targets for boosting the detection performance. Their augmentation strategy actually indicates spatial correspondence between the BEV feature and BEV position. And we further find the spatial correspondence between the BEV feature and BEV semantic segmentation can also be established. In other words, when two semantic segmentation maps are symmetric, their BEV features should also be symmetric. Thus apart from applying consistency in BEV semantic segmentation, we design a feature consistency loss L_{fc} for the BEV feature to refine the consistency:

$$L_{fc} = \|\widetilde{F}_U - \Phi(F_U^{'})\|_2. \tag{3}$$

where F_U and $F_U^{'}$ are the BEV features of I_U and $I_U^{'}$. $\|\cdot\|_2$ is the L2 distance. Φ denotes the horizontally flipping operation and $\widetilde{F}_U$ means that the gradient of F_U is detached.

E. Conjoint Rotation for Data Augmentation

Data augmentation can effectively improve the quantity and diversity of the training set to boost performance. There are no specially designed data augmentation methods on the visual BEV semantic segmentation, and only some simple methods, e.g., horizontal flipping and color jitter are used, to the best of our knowledge. it's mainly because the geometric relationship between the FV image and the BEV semantic segmentation is more complex than in the pixel-aligned 2D/3D tasks. The existing methods can easily destroy the spatial position relationship of corresponding pixels, making the view transformer more challenging to be trained.

We find that conjointly rotating the FV image and the GT BEV semantic segmentation map can reasonably augment the dataset without damaging the geometric relationship. As shown in Fig. 3, with the random angle α sampled from a pre-determined interval $[-\alpha_{max}, \alpha_{max}]$, we rotate the GT BEV semantic segmentation map and the FV image along a y-axis that is vertical to the BEV plane and passes through the origin of camera coordinate system.

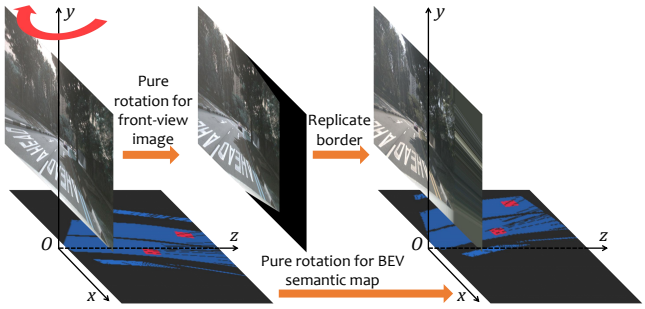

Fig. 3. Illustration of conjoint rotation.

The above rotation is a pure rotation for the camera, so the rotated FV image can be obtained using a homography transformation H_1 that merely relates to the camera intrinsic matrix K and angle α by forward warping operation. And according to [34], the transformation brought by H_1 can be expressed using the follow equations:

$$\begin{cases} u_2 = (h_{11}u_1 + h_{12}v_1 + h_{13})/(h_{31}u_1 + h_{32}v_1 + h_{33}) \\ v_2 = (h_{21}u_1 + h_{22}v_1 + h_{23})/(h_{31}u_1 + h_{32}v_1 + h_{33}) \end{cases}, \tag{4}$$

where the (u_1, v_1) and (u_2, v_2) respectively denotes the pixel coordinate in the original and transformed images, and h_{ij} is only determined by K and α. The above forward warping operation will introduce black edges in the transformed image. Such black edges can lower the improvement of the conjoint rotation, and we find replicating the border after the homography transformation can work better, as shown in our experiments in Sec. IV-E.

Furthermore, the perpendicular relationship between the BEV plane and the y-axis makes the rotation of the BEV semantic segmentation map equivalent to a rotation in the x-z plane around the coordinate origin O. The rotated GT BEV semantic segmentation map can be obtained by inverse warping operation with a 2D rotation matrix H_2:

$$H_2 = \begin{bmatrix} cos(\alpha) & -sin(\alpha) \\ sin(\alpha) & cos(\alpha) \end{bmatrix}. \tag{5}$$

Note that if the above y-axis is not vertical to the BEV plane, it is impossible to get the rotated BEV segmentation map because of the unpredictable occlusion.

Conjoint rotation acts concurrently on the FV images and the GT BEV segmentation map for labeled data while only acting on the FV images for unlabeled data.

F. Training Process

Each training batch consists of half labeled data and half unlabeled data, which is then augmented by the proposed conjoint rotation to get I_L and I_U. And $I_U^{'}$ is got by horizontal flipping the I_U. The Student Net M_S is used to predict BEV semantic segmentation maps Y and $Y_U^{'}$. The Teacher Net M_T is used to predict Y_U. We update θ_t, the parameters of M_S at training step t using following overall loss:

$$L = L_{sup} + \lambda_1 L_{sc} + \lambda_2 L_{fc}, \tag{6}$$

Then, following Mean-Teacher [33], we update $\theta_t^{'}$, the parameters of M_T at training step t using exponential moving average (EMA):

$$\theta_t^{'} = \alpha\theta_{t-1}^{'} + (1-\alpha)\theta_t. \tag{7}$$

where the EMA decay factor α is set as 0.999 empirically. And after the training, we use the M_T for evaluation.

IV. Experiments

A. Datasets

We conduct experiments on the nuScenes [15]. Following [1], We use the same data generation process and the same data split. The training set and testing set contain 168048 images and 35886 images, respectively. The resolution of input images is 600×800, and the output BEV semantic segmentation map has a resolution of 196×200, with each pixel representing $0.25m \times 0.25m$ in the real world.

Following the conventions in semi-supervised tasks, we divide the training set into labeled and unlabeled subsets with different ratios. Specifically, we use the first 5%, 10%, 20%, and 40% samples of each sequence of nuScenes as the labeled set and assume the remaining samples as the unlabeled set.

B. Network Architecture

Our model has the same architecture as PON [1], a milestone work in the visual BEV semantic segmentation field. The mode uses a ResNet-50 with an FPN [35] as the backbone. The view transformer is implemented by an MLP. And the segmentation decoder consists of a stack of residual blocks and a sigmoid activation function at the last layer.

C. Implementation Details

Our work is implemented in Pytorch on 8 NVIDIA V100 GPUs. We train the models using the Adam optimizer with 25 epochs and a batch size of 32. The initial learning rate is set as 1×10^{-4} and decays to 1×10^{-5} after 15 epochs. The weight λ_1 and λ_2 in Eq. 6 are empirically set as 2×10^{-3} and 2×10^{-4} respectively. Besides, we set $\alpha_{max} = 35°$ for the proposed conjoint rotation augmentation. With a 50% chance, we apply the conjoint rotation on the FV images

TABLE I

IOU(%) ON NUSCENES WITH DIFFERENT RATIOS OF LABELS. "C.V.": CONSTRUCTION VEHICLES, "PED.": PEDESTRIAN, "MOTOR": MOTORCYCLE.

Ratio	Method	Mean	drivable	crossing	walkway	carpark	car	trunk	bus	trailer	C.V.	ped.	motor.	bike	cone	barrier
5%	sup-only	14.8	57.7	25.7	30.3	24.9	29.5	9.8	5.1	4.3	0.6	2.7	0.6	1.4	6.1	8.0
	Π-Model [36]	15.1	57.1	26.2	29.7	24.2	29.4	10.8	5.5	7.3	**1.7**	2.7	0.7	1.7	5.4	8.2
	MT [33]	15.4	56.7	26.8	30.4	25.3	30.4	11.4	6.8	7.5	0.9	3.0	0.5	1.2	6.9	8.4
	CPS [29]	14.5	57.0	25.4	29.4	24.0	29.3	10.5	5.9	6.6	0.3	2.0	0.1	0.3	5.0	7.8
	UniMatch [37]	14.4	57.0	25.2	29.3	24.0	29.2	10.4	5.7	6.4	0.4	1.8	0.1	0.3	4.8	7.2
	Ours	**18.1**	**59.3**	**29.8**	**33.8**	**26.1**	**34.6**	**14.7**	**9.8**	**10.7**	1.5	**5.9**	**1.6**	**2.7**	**9.6**	**12.5**
10%	sup-only	17.3	58.6	26.8	32.5	**28.5**	33.2	14.9	9.9	10.3	1.2	4.9	2.1	2.3	7.5	10.2
	Π-Model [36]	17.4	59.3	30.0	32.7	27.2	33.0	13.2	8.7	9.5	1.4	5.8	1.8	3.1	8.0	10.4
	MT [33]	17.8	58.6	29.7	31.8	27.4	33.7	15.6	8.9	11.4	1.9	5.1	1.8	2.8	9.4	10.8
	CPS [29]	16.4	58.7	27.0	30.3	26.0	32.0	14.6	7.7	10.1	1.2	3.0	0.5	1.3	8.0	8.7
	UniMatch [37]	16.6	58.8	27.2	30.2	26.4	32.4	14.7	8.0	10.0	1.7	3.2	0.7	1.2	8.5	8.9
	Ours	**20.1**	**60.8**	**31.9**	**35.7**	27.4	**36.4**	**17.3**	**13.8**	**13.9**	**2.8**	**7.8**	**4.0**	**5.2**	**11.4**	**12.9**
20%	sup-only	19.5	61.0	32.4	34.8	27.7	36.8	15.8	14.0	11.9	2.3	7.1	3.8	3.5	7.8	13.7
	Π-Model [36]	19.8	60.7	32.2	35.2	26.8	36.5	16.7	13.6	11.4	1.9	6.8	4.7	4.7	10.6	15.0
	MT [33]	20.3	60.4	32.8	36.0	29.8	36.2	17.6	11.9	12.9	**4.4**	7.8	3.7	4.4	**11.1**	15.1
	CPS [29]	18.4	59.4	31.7	35.2	29.2	35.0	17.0	10.9	12.0	3.6	7.0	2.7	3.5	10.2	13.6
	UniMatch [37]	18.2	59.4	31.3	35.0	29.0	35.0	16.8	10.5	11.8	3.5	6.8	2.6	3.2	10.0	13.6
	Ours	**21.9**	**61.5**	**34.0**	**37.0**	**30.6**	**38.5**	**20.4**	**16.8**	**14.3**	3.4	**9.7**	**6.7**	**6.9**	10.7	**15.8**
40%	sup-only	22.3	61.3	34.9	37.5	30.9	38.9	20.5	17.8	16.4	3.0	10.5	6.1	6.3	**11.5**	15.9
	Π-Model [36]	22.6	61.8	35.1	37.9	30.8	38.2	20.6	21.1	15.5	4.7	9.9	6.4	7.1	10.4	16.4
	MT [33]	22.6	61.5	34.9	37.9	**31.6**	38.4	20.0	18.5	16.2	3.2	10.5	**7.6**	**8.3**	10.9	**16.7**
	CPS [29]	20.6	59.5	33.0	36.0	29.5	37.3	18.0	17.2	13.1	1.6	8.5	5.7	6.2	8.9	15.0
	UniMatch [37]	20.5	59.6	33.0	35.7	29.4	37.3	18.0	17.0	12.8	1.6	8.4	5.6	6.0	8.9	15.0
	Ours	**23.5**	**62.8**	**36.0**	**38.9**	31.5	**39.6**	**22.7**	**21.6**	**18.3**	**5.1**	**11.1**	6.8	7.6	11.1	16.1

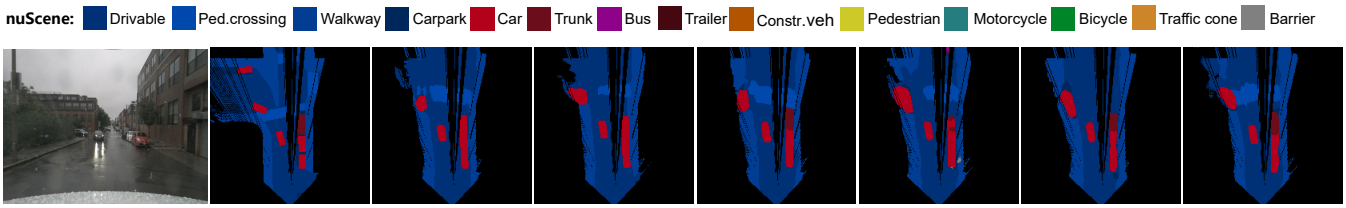

Fig. 4. **Qualitative results with 20% labels.** We follow the color scheme in PON [1] and use the visibility mask (black) for visualization.

before being resized and fed to the network. For evaluation, following [1], we use the IoU as our evaluation metric, and those invisible pixels are ignored during evaluation.

For sufficient comparison, we extend several classical and state-of-the-art semi-supervised 2D segmentation methods to this task. Especially, we look into Π-Model [36], Mean-Teacher (MT) [33], CPS [29] and UniMatch [37]. And their weights of consistency loss are respectively set as 5×10^{-3}, 2×10^{-3}, 1×10^{-3} and 1×10^{-3}.

D. Main Results

Tab. I presents the class-wise IoU scores on the nuScenes dataset. With various ratios of labeled data, our semi-supervised framework can significantly outperform the supervised-only baselines in almost all categories, indicating that our framework is able to utilize unlabeled data to effectively enhance performance. Given 10% labeled data and 90% unlabeled data, our framework can even outperform the full-supervised baseline using 20% labels. A similar improvement is also achieved under the setting of 5% labels. This implies that our semi-supervised framework can enhance the efficiency of data utilization. Compared with extended 2D semi-supervised methods [36], [33], [29], [37], our framework achieves better scores. Thus the superiority of our framework is presented. And interestingly, state-of-the-art 2D semi-supervised methods [29], [37] perform worse than classical methods [36], [33] on this task. We conjecture that the operation of generating pseudo labels in [29] and [37] is not applicable to this multi-label classification task.

Furthermore, the qualitative results on the nuScenes with 20% labels are shown in Fig. 4. They also prove that by exploiting unlabeled data, our framework helps improve the semantic segmentation quality.

TABLE II

MIOU(%) FOR ABLATION STUDIES. CR DENOTES CONJOINT ROTATION.

L_{sup}	L_{sc}	L_{fc}	CR	5%	10%	20%	40%
✓				14.8	17.3	19.5	22.3
✓	✓			15.4(↑0.6)	17.8(↑0.5)	20.3(↑0.8)	22.6(↑0.3)
✓	✓	✓		15.5(↑0.7)	17.9(↑0.6)	20.5(↑1.0)	22.3(↑0.0)
✓			✓	17.5(↑2.7)	19.2(↑1.9)	21.1(↑1.6)	22.8(↑0.5)
✓	✓		✓	18.1(↑3.3)	20.1(↑2.8)	21.4(↑1.9)	23.3(↑1.0)
✓	✓	✓	✓	**18.1(↑3.3)**	**20.1(↑2.8)**	**21.9(↑2.4)**	**23.5(↑1.2)**

E. Ablation Study

To better understand the effect of each component of our framework, we conduct an ablation study as presented in Tab. II. The results show that when all components are combined together, the performance is the best. Note that we only report mIoU scores in this section, due to the abundance of experiments.

We conduct more detailed ablation studies to get deeper insights into our framework. The following experiments are conducted with 20% labels unless otherwise specified.

Benefits of Consistency losses. The supervised loss L_{sup} gives the model the primary supervisory signal. However, when labels are limited, information of unlabeled data cannot be excavated with only L_{sup}. As shown in Tab. II, compared with the full-supervised model, better scores can be gained by introducing segmentation consistency loss L_{sc}. Furthermore,

Fig. 5. **Different border modes.** (a)Original FV image. Augmented FV image with (b)zero border, (c)reflect border, and (d)replicate border.

applying consistency constraints on the BEV feature by L_{fc}, model performance can be further refined in almost all cases.

Fig. 6. **Sensitivity of θ_{max} of conjoint rotation augmentation.**

Effectiveness of conjoint rotation. The results in Tab. II show that in all cases, conjoint rotation can significantly improve performance. Thus, the conjoint rotation is effective for both full-supervised and semi-supervised models thanks to maintaining the 3D geometric relationship even though information on image edges may be lost. As the unique hyper-parameter for conjoint rotation, the role of α_{max} needs to be explored. And we conduct sensitivity experiments on α_{max} and show results in Fig. 6. According to Fig. 6, we choose $35°$ as θ_{max} for better performance. And the results also show that the improvement is remarkable in a wide range of θ_{max} $(15° - 55°)$, which validates the robustness of conjoint rotation. The bordering mode is also essential for conjoint rotation. We compare the performance using different bordering modes(Fig. 5(b) to (d)) and present the results in Tab. III. With zero border, image black edges brought by forward warping operation can lower the improvement of conjoint rotation. And replicating border can make the improvement more significant. But interestingly, there is no improvement when reflecting border. Moreover, we make a comparison with other augmentation methods to demonstrate our effectiveness in Tab. IV. The unsatisfaction with Cutout [38] and Random Erasing [31] may lie in the damage to the geometric relationship between the FV images and the BEV semantic segmentation maps. Although BEV-Space data augmentation [12] can improve detection performance, the performance degradation shown in Tab. IV proves that it's not applicable to this task.

TABLE III

ABATION STUDY ON BORDER MODE.

Replicate Border (Ours)	Zero Border	Reflect Border
21.9	21.3	21.3

Perturbation strategy. Different perturbation strategies may bring different results for consistency-based semi-supervised learning. We make a comparison between our horizontal flip and color jitter, a common perturbation strategy in the semi-supervised 2D semantic segmentation field. Results of the first two rows in Tab. V show that our framework without L_{fc} performs slightly better when using color jitter as the perturbation. But when feature consistency loss L_{fc} is applied, performance can be further improved with the horizontal flip while almost unchanged with color jitter. It indicates that L_{fc} improves the performance by effectively constraining the spatial consistency that is perturbed by horizontal flip. And the results in the third row imply that the color jitter can destroy such consistency.

TABLE IV

COMPARISON WITH OTHER AUGMENTATION METHODS.

Augmentation method	mIoU(%)
Cutout [38]	20.1
Random Erasing [31]	20.5
BEV-space Data Augmentation(Rotate) [12]	20.4
BEV-space Data Augmentation(Flip) [12]	20.4
Conjoint Rotation	**21.9**

TABLE V

MIOU(%) SCORES WITH DIFFERENT PERTURBATION STRATEGY.

Horizontal Flip	Color Jitter	Ours w/o L_{fc}	Ours
✓		21.4	**21.9**
	✓	21.5	21.5
✓	✓	21.3	21.6

Improvements with 3D-to-2D-based VT. To verify the effectiveness of our framework, we further use the 3D-to-2D-based VT [10] to conduct the experiments. mIoU scores in Tab. VI validate that our framework can still effectively exploit unlabeled data to improve performance.

TABLE VI

IMPROVEMENTS WITH 3D-TO-2D-BASED VT ON THE NUSCENES

Method	5%	10%	20%	40%
3D-to-2D(sup-only)	14.6	16.4	17.6	19.9
3D-to-2D(semi-sup)	**16.2**	**17.4**	**18.9**	**20.7**

V. CONCLUSION

In this work, we delve into the visual BEV semantic segmentation with limited labels and present a novel semi-supervised framework to utilize unlabeled data to improve performance. We propose restricting the model using consistency on semantic segmentation and the BEV feature to use unlabeled data fully. Moreover, we design a novel data augmentation method based on the ingenious geometric relationship. Experiment results demonstrate that our framework can effectively improve performance and data utilization even when using different view transformer. And the effectiveness of our contributions is proved by extensive ablation studies. In the future, we will investigate extending our contributions to BEV detection and 3D semantic occupancy tasks.

REFERENCES

[1] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *CVPR*, 2020.

[2] S. Gong, X. Ye, X. Tan, J. Wang, E. Ding, Y. Zhou, and X. Bai, "Gitnet: Geometric prior-based transformation for birds-eye-view segmentation," in *ECCV*, 2022.

[3] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *ICRA*, 2022.

[4] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *CVPR*, 2022.

[5] 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.

[6] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, E. Xie, Z. Li, H. Deng, H. Tian, *et al.*, "Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe," *arXiv preprint arXiv:2209.05324*, 2022.

[7] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *ECCV*, 2020.

[8] H. Zhou, Z. Ge, Z. Li, and X. Zhang, "Matrixvt: Efficient multi-camera to bev transformation for 3d perception," *arXiv preprint arXiv:2211.10593*, 2022.

[9] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Enabling spatio-temporal aggregation in birds-eye-view vehicle estimation," in *ICRA*, 2021.

[10] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *ECCV*, 2022.

[11] N. Gosala and A. Valada, "Bird's-eye-view panoptic segmentation using monocular frontal view images," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1968–1975, 2022.

[12] J. Huang, G. Huang, Z. Zhu, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[13] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *CVPR*, 2021.

[14] A. Xiao, J. Huang, D. Guan, K. Cui, S. Lu, and L. Shao, "Polarmix: A general data augmentation technique for lidar point clouds," in *NeurIPS*, 2022.

[15] 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 *CVPR*, 2020.

[16] S. Sengupta, P. Sturgess, L. Ladickỳ, and P. H. Torr, "Automatic dense visual semantic mapping from street-level imagery," in *IROS*, 2012.

[17] Y. B. Can, A. Liniger, O. Unal, D. Paudel, and L. Van Gool, "Understanding bird's-eye view of road semantics using an onboard camera," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3302–3309, 2022.

[18] I. Dwivedi, S. Malla, Y.-T. Chen, and B. Dariush, "Bird's eye view segmentation using lifted 2d semantic features," in *BMVC*, 2021.

[19] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," *arXiv preprint arXiv:1811.08188*, 2018.

[20] Z. Liu, S. Chen, X. Guo, X. Wang, T. Cheng, H. Zhu, Q. Zhang, W. Liu, and Y. Zhang, "Vision-based uneven bev representation learning with polar rasterization and surface estimation," in *CoRL*, 2022.

[21] F. Bartoccioni, E. Zablocki, A. Bursuc, P. Perez, M. Cord, and K. Alahari, "Lara: Latents and rays for multi-camera bird's-eye-view semantic segmentation," in *CoRL*, 2022.

[22] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *CVPR*, 2023.

[23] 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. 11974–11981, 2022.

[24] G. French and M. Mackiewicz, "Colour augmentation for improved semi-supervised semantic segmentation," *arXiv preprint arXiv:2110.04487*, 2021.

[25] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *CVPR*, 2021.

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

[27] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *Neural Networks*, vol. 145, pp. 90–106, 2022.

[28] T. Kalluri, G. Varma, M. Chandraker, and C. Jawahar, "Universal semi-supervised semantic segmentation," in *CVPR*, 2019.

[29] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *CVPR*, 2021.

[30] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *CVPR*, 2020.

[31] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI*, 2020.

[32] 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 *ICCV*, 2019.

[33] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *NeurIPS*, 2017.

[34] J.-W. Bian, H. Zhan, N. Wang, T.-J. Chin, C. Shen, and I. Reid, "Auto-rectify network for unsupervised indoor depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9802–9813, 2021.

[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.

[36] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *ICLR*, 2016.

[37] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *CVPR*, 2023.

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