MWSIS: Multimodal Weakly Supervised Instance Segmentation with 2D Box Annotations for Autonomous Driving

Guangfeng Jiang[1], Jun Liu[1*], Yuzhi Wu[1], Wenlong Liao[2], Tao He[3], Pai Peng[3*]

[1]Department of Electronic Engineering and Information Science, University of Science and Technology of China
[2]Shanghai Jiao Tong University
[3]COWAROBOT
jgf1998@mail.ustc.edu.cn, junliu@ustc.edu.cn, pengpai@cowarobot.com

Abstract

Instance segmentation is a fundamental research in computer vision, especially in autonomous driving. However, manual mask annotation for instance segmentation is quite time-consuming and costly. To address this problem, some prior works attempt to apply weakly supervised manner by exploring 2D or 3D boxes. However, no one has ever successfully segmented 2D and 3D instances simultaneously by only using 2D box annotations, which could further reduce the annotation cost by an order of magnitude. Thus, we propose a novel framework called Multimodal Weakly Supervised Instance Segmentation (MWSIS), which incorporates various fine-grained label correction modules for both 2D and 3D modalities, along with a new multimodal cross-supervision approach. In the 2D pseudo label generation branch, the Instance-based Pseudo Mask Generation (IPG) module utilizes predictions for self-supervised correction. Similarly, in the 3D pseudo label generation branch, the Spatial-based Pseudo Label Generation (SPG) module generates pseudo labels by incorporating the spatial prior information of the point cloud. To further refine the generated pseudo labels, the Point-based Voting Label Correction (PVC) module utilizes historical predictions for correction. Additionally, a Ring Segment-based Label Correction (RSC) module is proposed to refine the predictions by leveraging the depth prior information from the point cloud. Finally, the Consistency Sparse Cross-modal Supervision (CSCS) module reduces the inconsistency of multimodal predictions by response distillation. Particularly, transferring the 3D backbone to downstream tasks not only improves the performance of the 3D detectors, but also outperforms fully supervised instance segmentation with only 5% fully supervised annotations. On the Waymo dataset, the proposed framework demonstrates significant improvements over the baseline, especially achieving 2.59% mAP and 12.75% mAP increases for 2D and 3D instance segmentation tasks, respectively. The code is available at https://github.com/jiangxb98/mwsis-plugin.

Introduction

In the field of autonomous driving, instance segmentation of images and point clouds are two important research directions, each providing fine-grained perception results in their respective modalities. However, the fully supervised instance segmentation approaches (He et al. 2017; Tian, Shen, and Chen 2020; Cheng, Schwing, and Kirillov 2021; Jiang et al. 2020; Wu et al. 2022) rely on pixel-wise or point-wise instance annotations, which are labor-intensive and costly (Behley et al. 2019; Cheng, Parkhi, and Kirillov 2022). As a result, many weakly supervised approaches (Hsu et al. 2019; Lee et al. 2021; Shen et al. 2021; Tian et al. 2021; Wang et al. 2021; Cheng, Parkhi, and Kirillov 2022; Chibane et al. 2022; Dong et al. 2022) using weak labels have emerged, e.g., BoxInst (Tian et al. 2021) is trained with 2D box annotations to predict 2D masks in the image space, and Box2Mask(Chibane et al. 2022) is trained with 3D box annotations to predict 3D masks in the point cloud space. However, the 3D box annotation is also costly.

These early weakly supervised works are mostly limited to a single modality. Thanks to the development of multimodal datasets, e.g., nuScenes (Caesar et al. 2020) and Waymo (Sun et al. 2020), which offer millions of images and hundreds of thousands of LiDAR scans. To the best of our knowledge, LWSIS (Li et al. 2023) is the first to focus on multimodal instance segmentation, which employs LiDAR points inside 3D box annotations to guide the 2D instance segmentation. However, LWSIS only utilizes multimodal information to obtain a single-modal segmentor and uses costly 3D box annotations. This leads us to the question: **Is it possible to leverage well synchronized & aligned camera and LiDAR data, along with only much simpler and more cost-effective 2D box annotations, to train both image and point cloud segmentors simultaneously via weak supervision?**

Motivated by this question, we propose a novel framework called Multimodal Weakly Supervised Instance Segmentation (MWSIS). It trains both 2D and 3D segmentors using only 2D box annotations, leveraging complementary information from different modalities through cross-modal distillation.

However, our framework inevitably faces two challenges: one is how to deal with the low signal-to-noise ratio introduced by 2D box annotations as weak supervision signals, and the other is how to implement cross-modal distillation. To address these challenges, we devise several modules for pseudo label generation and correction at various granularity, such as the Instance-based Pseudo Mask Gen-

[*]Corresponding authors.

eration (IPG) module, the Spatial-based Pseudo Label Generation (SPG) module, the Point-based Voting Label Correction (PVC) module, and the Ring Segment-based Label Correction (RSC) module, to improve the quality of pseudo labels and filter out noise. Additionally, we design a Consistency Sparse Cross-modal Supervision (CSCS) module to achieve cross-modal supervision. The detailed implementation of the above modules is described in the method section.

In short, our contributions are summarized in four folds:

- To the best of our knowledge, we are the first to use the 2D box annotations as the sole external supervision signal to train both image and point cloud instance segmentors simultaneously.
- We propose various fine-grained label correction modules for different modalities, including instance-based, spatial-based, point-based, and ring segment-based modules. These modules not only enhance the instance segmentation performance, but also improve the quality of the pseudo label.
- We propose a novel cross-modal supervision method, named CSCS, which exploits the complementary properties of the point cloud and image modalities. This method improves the performance of the segmentors.
- Our framework can be used as a pre-training method to improve the performance of 3D downstream tasks such as semantic segmentation, instance segmentation, and object detection.

Related Work

The development of multimodal datasets, such as KITTI (Geiger, Lenz, and Urtasun 2012), nuScenes (Caesar et al. 2020), and Waymo (Sun et al. 2020), has greatly contributed to the progress of weakly supervised methods in various tasks, including semantic segmentation (Wei et al. 2020; Hu et al. 2021; Unal, Dai, and Gool 2022), instance segmentation (Chibane et al. 2022; Li et al. 2023), and object detection (Qin, Wang, and Lu 2020; Wei et al. 2021; Peng et al. 2022; Liu et al. 2022b,a). The weakly supervised instance segmentation task is to extract objects from an image or point cloud using simpler annotations than mask annotations. These simpler annotations include image-level labels (Cholakkal et al. 2019; Ge et al. 2019; Shen et al. 2021), points (Lee, Kim, and Sull 2021; Cheng, Parkhi, and Kirillov 2022; Li et al. 2023), scribble (wei Li et al. 2023; Chen et al. 2023), and boxes (Hsu et al. 2019; Lee et al. 2021; Wang et al. 2021; Chibane et al. 2022).

2D Weakly Supervised Instance Segmentation

Among models trained with image-level supervision, PDSL (Shen et al. 2021) adopts self-supervised learning to learn class-independent foreground segmentation. Among works using box-level supervision, BBTP (Hsu et al. 2019) is the first one that uses 2D boxes to generate instance masks in the formulation of multiple instance learning, leveraging the tightness of boxes to predict instance masks. BBAM (Lee et al. 2021) utilizes the semantics extracted by a trained detector to generate pseudo labels, which are then used to train semantic segmentation and instance segmentation networks. BoxInst (Tian et al. 2021) makes use of the local color consistency constraints and designs projection loss and pairwise similarity loss to supervise the mask branch of CondInst (Tian, Shen, and Chen 2020). Among point-level supervised methods, PointSup (Cheng, Parkhi, and Kirillov 2022) adds random sampling points as segmentation annotations based on 2D box annotations. LWSIS (Li et al. 2023) is the first work to incorporate LiDAR points, which utilizes more precise guidance of points inside the 3D box as a supervision signal for 2D instance segmentation, and has achieved significant improvement without introducing additional network parameters.

3D Weakly Supervised Instance Segmentation

There is only a small amount of work on the weakly supervised 3D instance segmentation. As one of the methods using 3D box annotations, Box2Mask (Chibane et al. 2022) is inspired by classical Hough voting, in which each point directly votes for a 3D box, and it uses the IoU-guided NMC method to obtain the clustered box, which is then back-project to the point cloud to obtain the instance segmentation result. RWSeg (Dong et al. 2022) employs self-attention and random walk to propagate semantic and instance information to unknown regions, respectively.

Basically, these methods mentioned above can be divided into two categories: single-modal weakly supervised models like (Tian et al. 2021; Lee et al. 2021; Chibane et al. 2022; Dong et al. 2022) and multimodal weakly supervised models for a single-modal task like (Li et al. 2023). The aforementioned single-modal approaches utilize the information only from a single modality and do not take advantage of complementary information from multiple modalities, while the multimodal approaches leverage information from other modalities to acquire single-modal segmentors. In this paper, we examine the first attempt at multimodal weakly supervised instance segmentation to acquire the multimodal segmentors simultaneously, while using only 2D box annotations.

Method

In this section, we provide the implementation details of our MWSIS framework. The framework consists of two separate modal branches and a multimodal cross-supervision module. Each branch includes a trainable student segmentor and a teacher segmentor updated by EMA (Exponential Moving Average). The cross-supervision module utilizes the output of the teacher segmentor to perform cross-supervision on the student segmentor of another modality through response distillation. As shown in Fig. 1, in the image branch (top), we choose the BoxInst as the baseline and employ the instance-based IPG module for self-supervision. In the point cloud branch (bottom), we employ a plain voxel-based SparseUNet (Shi et al. 2020) segmentor as the baseline, and we introduce the SPG module based on the spatial priors of point clouds, such as depth and Euclidean distance, to generate high-quality pseudo instance masks for the point cloud from 2D box annotations. We also present the point-based

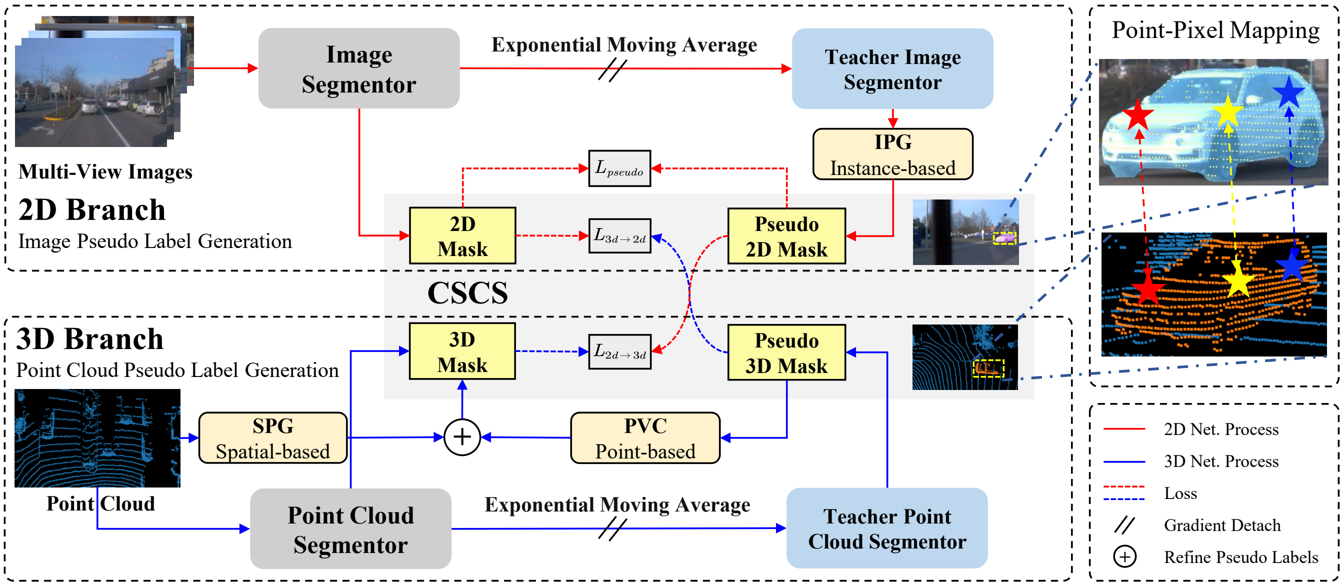

Figure 1: Overview of MWSIS framework. The framework mainly consists of three parts: image pseudo label generation branch, point cloud pseudo label generation branch, and CSCS module. In the 2D branch, the pseudo 2D masks generated by the IPG module self-supervise the 2D masks. In the 3D branch, the label is refined by the SPG module and PVC module to supervise the 3D Masks. Finally, the pseudo masks from the teacher model are used for cross-supervision in the CSCS module.

PVC module based on voting using historical results from the teacher segmentor to refine the pseudo labels, and the ring segment-based RSC module to refine the student segmentor predictions. Finally, the CSCS module incorporates the mean teacher (Tarvainen and Valpola 2017) and CPS (Chen et al. 2021) principles and employs response distillation to ensure consistency between multimodal masks.

Preliminary

We define the input multimodal data as $\{\mathbf{P}, \mathbf{I}\}$, where $\mathbf{P} \in \mathbb{R}^{N_{in}\times C_{in}}$ denotes N_{in} points with C_{in}-dimensional input features (e.g., 3D cordinates and reflectance) and $\mathbf{I} \in \mathbb{R}^{N_{cam}\times 3\times H_{in}\times W_{in}}$ denotes the multi-view RGB images of dimension $3 \times H_{in} \times W_{in}$ obtained from N_{cam} cameras. Using sensor calibration of the dataset, we can project 3D points onto 2D images, and obtain the point-pixel mapping relationship between spatial points $\mathbf{P}_{3d} \in \mathbb{R}^{N_{in}\times 3}$ and pixel points $\mathbf{P}_{2d} \in \mathbb{R}^{N_{in}\times 2}$.

2D – Image Pseudo Label Generation

Instance-based Pseudo Mask Generation (IPG). We chose BoxInst (Tian et al. 2021) as our 2D baseline. In anchor-free methods (Tian, Shen, and Chen 2020; Tian et al. 2019), multiple positive samples can be assigned to the same ground truth (GT) instance during the assignment, resulting in inconsistency predictions, as shown by the different colors of stars in Fig. 2. The masks with higher confidence scores and closer to the GT boxes center are considered to be more reliable. In this view, we propose the instance-based IPG module, as illustrated in Fig. 2. Specifically, we calculate the IoUs between the predicted boxes and the GT boxes

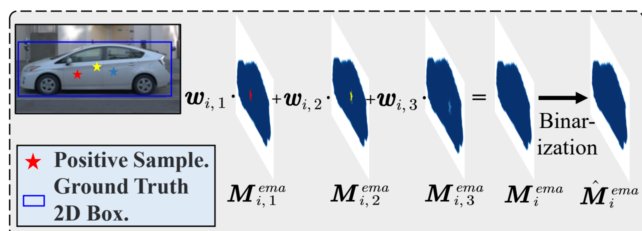

Figure 2: IPG pseudo mask generation process.

and weight the corresponding predicted masks based on the IoUs and scores. The formula is defined as:

$$\mathbf{M}_i^{ema} = \sum_{j}^{N_i} w_{i,j}\mathbf{M}_{i,j}^{ema}$$
$$w_{i,j} = \frac{s_{i,j}e^{kIoU_{i,j}}}{\sum_j^{N_i} s_{i,j}e^{kIoU_{i,j}}} \tag{1}$$

where $\mathbf{M}_i^{ema}$ is the predicted probability map corresponding to the i-th GT box, N_i is the number of positive samples assigned to the i-th GT box, $\mathbf{M}_{i,j}^{ema}$ is the j-th mask corresponding to the i-th GT box, k is a hyperparameter, $s_{i,j}$ and $IoU_{i,j}$ are the confidence score and IoU of the j-th predicted box to the i-th GT box, respectively.

After obtaining the weighted probability map, we set two thresholds τ_{low} and τ_{high} to obtain the pseudo mask $\hat{\mathbf{M}}_i^{ema}$. Then we apply the pseudo mask $\hat{\mathbf{M}}_i^{ema}$ as the self-supervision signal. The self-supervision loss function is de-

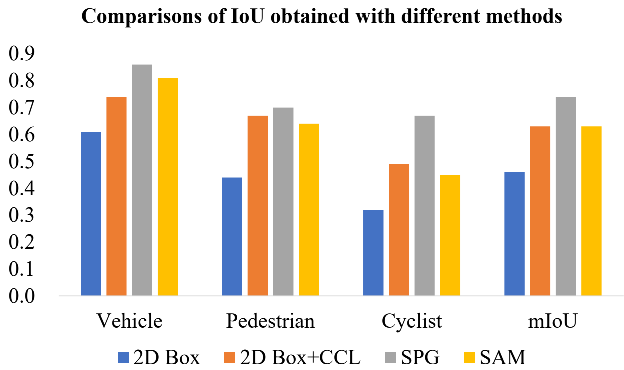

Figure 3: Comparisons of IoU obtained with different methods on Waymo validation dataset. SAM means the process of obtaining masks through the use of SAM (Kirillov et al. 2023), where 2D boxes are employed as prompts.

fined as:

$$L_{pseudo} = L_{pseudo}\left(\mathbf{M}_{pred}, \hat{\mathbf{M}}^{ema}\right) \tag{2}$$

where L consists of two terms: binary cross-entropy loss L_{bce} and dice loss L_{dice}, $\mathbf{M}_{pred}$ is the predicted masks.

3D – Point Cloud Pseudo Label Generation

The input point cloud $\mathbf{P}$ can be roughly partitioned into potential foreground points $\mathbf{P}_{in}$ and background points $\mathbf{P}_{out}$ by 3D frustums bounded by 2D boxes. However, as shown in Fig. 3 and Fig. 4(b), the foreground partition still contains a large number of background points, which is suboptimal to be treated as the pseudo label directly. Thus, we adopt a simple and efficient CCL (Hennequin et al. 2018) clustering algorithm to partition the potential foreground points into clusters and select the largest connected component as the foreground points corresponding to the instance. To further improve the quality of the pseudo instance masks, we propose three label correction modules: the SPG, the PVC, and the RSC.

Spatial-based Pseudo Label Generation (SPG). We notice that LiDAR points obtained from each laser beam exhibit a certain pattern. If the depth changes suddenly, it is likely that a different object is scanned. In Fig. 4(c), points of the same color denote smooth depth variations, such as ground points, while points of different colors denote non-smooth depth variations. To leverage the prior information about the depth variation of the point cloud, we propose the Depth Clustering Segment (DCS) algorithm whose details are given in the following.

A LiDAR sensor returns an $M \times N$ measurement matrix in a single scan, where M denotes the number of beams and N denotes the number of measurements. The idea of the DCS algorithm is to traverse each row and record the depth of each column. If the depth changes smoothly, the current point belongs to the same ring segment. Otherwise, it belongs to a different ring segment. The algorithm is shown in Appendix A.1.

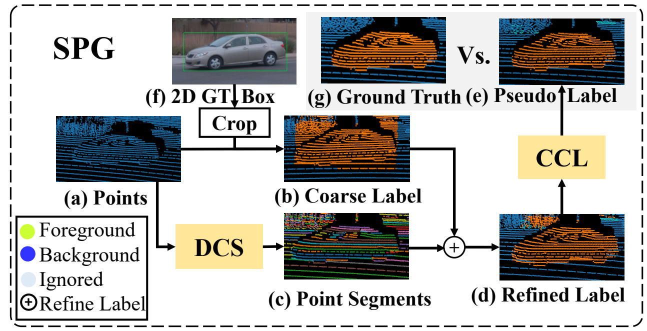

Figure 4: The overview of SPG.

Fig. 4 shows the process of generating 3D pseudo labels by the SPG. Firstly, points are cropped by 2D GT boxes to obtain coarse pseudo labels in Fig. 4(b), where the yellow and blue points denote the foreground points $\mathbf{P}_{in}$ and the background points $\mathbf{P}_{out}$, respectively. At the same time, the DCS algorithm is applied to cluster the point cloud into different ring segments $\mathbf{P}^{id}$ in Fig. 4(c). By comparing Fig. 4(b) and Fig. 4(c) we observe that some points belonging to the same segment are distributed inside and outside the 2D box, denoted by $\mathbf{P}^{id}_{i\_box}$ and $\mathbf{P}^{id}_{o\_box}$, respectively. We refine the labels by comparing the relative proportions of $\mathbf{P}^{id}_{i\_box}$ and $\mathbf{P}^{id}_{o\_box}$. If the number of points outside the box exceeds the number of points inside the box, we classify the $\mathbf{P}^{id}_{i\_box}$ as background points. The specific label classification is shown in the following formula:

$$\begin{aligned} \hat{y}^{ref}_{id} &= \begin{cases} 0, & \text{if } prop > 0.5 \\ 1, & \text{if } prop < 0.1 \\ -1, & \text{otherwise} \end{cases} \\ prop &= \frac{\left|\mathbf{P}^{id}_{o\_box}\right|}{\left|\mathbf{P}^{id}_{i\_box}\right| + \left|\mathbf{P}^{id}_{o\_box}\right|} \end{aligned} \tag{3}$$

where $\hat{y}^{ref}_{id}$ represents the pseudo labels of the points belonging to the id-th ring segment, $|\cdot|$ denotes the number of points. We iterate over all the points to get the refined pseudo labels $\hat{\mathbf{Y}}_{ref} = \{\hat{y}^{ref}_1, \ldots, \hat{y}^{ref}_{N_{in}}\} \in \{-1, 0, 1\}^{N_{in}}$ in Fig. 4(d), where the numbers -1, 0, 1 represent the ignored, background, and foreground labels of points, respectively.

After performing the aforementioned label refinement operation, we apply the CCL to all pseudo foreground points $\mathbf{P}_{fg}$ within the 2D boxes $\mathbf{B} = \{B_1, \ldots, B_{N_{box}}\}$, where N_{box} is the number of 2D boxes. According to Eq. 4, we can obtain the maximum cluster $\mathbf{P}'_{\{fg,b\}}$ of the foreground points $\mathbf{P}_{\{fg,b\}}$ in the b-th box, and get the corresponding point pseudo labels $\hat{\mathbf{Y}}_{\{fg,b\}}$. This process generates the final semantic labels $\hat{\mathbf{Y}}_{sem} = \{\hat{y}^{sem}_1, \ldots, \hat{y}^{sem}_{N_{in}}\} \in \{-1, 0, \ldots, N_{cls}\}^{N_{in}}$ and instance labels $\hat{\mathbf{Y}}_{inst} = \{\hat{y}^{inst}_1, \ldots, \hat{y}^{inst}_{N_{in}}\} \in \{0, 1, \ldots, N_{box}\}^{N_{in}}$, where N_{cls} is the number of classes, $\hat{y}^{sem}_i$ and $\hat{y}^{inst}_i$ are the i-th point semantic and instance labels, respectively.

$$\mathbf{P}'_{\{fg,b\}} = Max\left(CCL\left(\mathbf{P}_{\{fg,b\}}\right)\right) \tag{4}$$

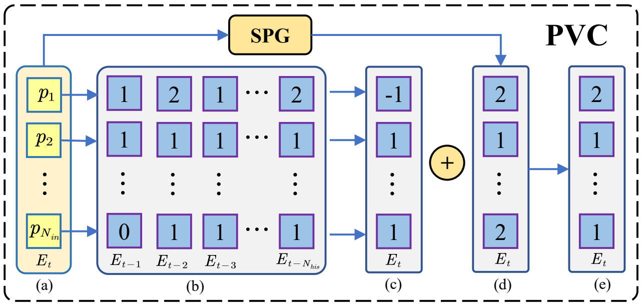

Figure 5: The PVC module architecture. For the current epoch in Fig. 5(a), historical predictions in Fig. 5(b) are used to vote and obtain correction pseudo labels in Fig. 5(c). Then the pseudo labels generated by SPG in Fig. 5(d) are corrected to obtain the final pseudo labels in Fig. 5(e).

Point-based Voting Label Correction (PVC). To further improve the accuracy of the pseudo label generated by the SPG, we propose the PVC module. The purpose of this module is to leverage the network generalization ability to correct the pseudo labels. Specifically, we establish a matrix space $\mathbf{H}$ of dimension $N_{num} \times N_{his} \times N_{in}$ to store the teacher model predictions from the previous N_{his} epochs, where N_{num} is the number of the samples in the training split. As shown in Fig. 5, for a frame of point cloud $\mathbf{P}$, when the current training epoch E_t reaches the defined epoch E_s, we use the predictions from the previous N_{his} epochs to conduct voting. In the voting process, we select the points with segmentation scores greater than τ_h or less than τ_l as reliable foreground and background, respectively, and ignore the rest. If the pseudo label of a point is deemed reliable for more than T_h samples, we override its pseudo segmentation label by the voted majority.

Ring Segment-based Label Correction (RSC). The RSC module conducts voting inside every laser ring segment produced by the DCS algorithm to further refine the predicted masks. Our RSC algorithm is shown in Appendix A.1. In this algorithm, if the proportion of points in the current ring segment that belongs to the background exceeds a threshold T_1, we override the corresponding predicted labels as background. Similarly, if the proportion of points in the current segment for foreground exceeds a threshold T_2, we override the predicted labels to the current foreground class.

Consistency Sparse Cross-modal Supervision (CSCS)

Inspired by the mean teacher(Tarvainen and Valpola 2017) and CPS (Chen et al. 2021), we design the CSCS module that combines the advantages of multimodal information by response distillation, as shown in Fig. 6. For a frame of point cloud and corresponding multi-view images, we obtain two distinct point-wise predictions from the student segmentor and teacher segmentor. Then we introduce a consistency sparse loss L_{cscs} to supervise the alignment between the two

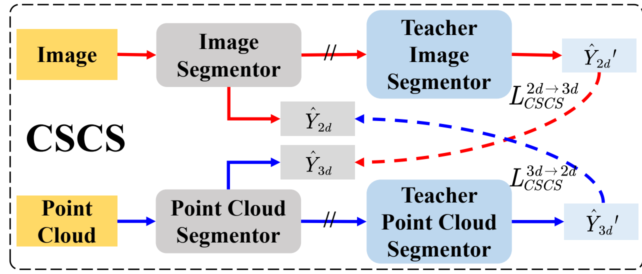

Figure 6: The architecture of CSCS module.

modalities. The CSCS loss function is defined as:

$$L_{cscs} = -\frac{1}{N_{in}N_{cls}} \sum_{i}^{N_{in}} \sum_{j}^{N_{cls}} [p_{i,j}^{t} \log\left(q_{i,j}^{s}\right) + \left(1 - p_{i,j}^{t}\right) \log\left(1 - q_{i,j}^{s}\right)] \tag{5}$$

$$L_{cscs}^{2d\to3d} = L_{cscs}\left(T_{net} = 2d, S_{net} = 3d\right) \tag{6}$$

$$L_{cscs}^{3d\to2d} = L_{cscs}\left(T_{net} = 3d, S_{net} = 2d\right) \tag{7}$$

where $p_{i,j}$ and $q_{i,j}$ are the scores of the j-th class for the i-th point in different modal branches, p^t and q^s denote the output of teacher network (T_{net}) and student network (S_{net}), respectively. The symbols $2d$ and $3d$ represent the 2D segmentor and 3D segmentor, respectively.

Loss

The overall loss function of the MWSIS is defined as:

$$L_{total} = L_{2d} + L_{3d} \tag{8}$$

In the 2D branch, we augment the BoxInst loss $L_{boxinst}$ with the self-supervised loss L_{pseudo} and the consistency sparse loss $L_{cscs}^{3d\to2d}$. In the 3D branch, we use the classification loss L_{cls}, the regression loss L_{vote}, and the consistency sparse loss $L_{cscs}^{2d\to3d}$.

$$L_{2d} = \alpha_1 L_{boxinst} + \alpha_2 L_{pseudo} + \alpha_3 L_{cscs}^{3d\to2d} \tag{9}$$

$$L_{3d} = \alpha_4 L_{cls} + \alpha_5 L_{vote} + \alpha_6 L_{cscs}^{2d\to3d} \tag{10}$$

where $\alpha_1 \sim \alpha_6$ are set as 1.0, 1.0, 0.5, 100.0, 1.0, 2.0 to balance loss terms, respectively.

Experiments

Waymo Open Dataset

We conduct our experiments on version 1.4.0 of the Waymo dataset, which includes both well synchronized & aligned LiDAR points and images. Waymo released a panoptic segmentation dataset (Mei et al. 2022), which includes 61,480 images for training and 9,405 images for validation. For the 3D segmentation, the dataset contains 23,691 and 5,976 frames for training and validation, respectively. Due to the differences between 2D and 3D segmentation tasks on training datasets, we train and evaluate the corresponding segmentation task on two datasets, respectively. We specifically focus on the vehicle, pedestrian, and cyclist classes.

Supervision	Annotation	Model	Epoch / Iter	mAP			AP		
				AP	AP50	AP75	Veh.	Ped.	Cyc.
Full	2D Mask	CondInst (2020)	24e	45.35	69.74	48.87	64.23	39.61	32.21
Weak	2D Box	BoxInst (2021)	24e	34.61	65.35	32.45	48.48	27.65	27.69
	2D+3D Box	LWSIS (2023)	90k	**37.77**	-	-	-	-	-
	2D Box	MWSIS (ours)	24e	37.20	67.03	38.07	**51.13**	29.05	31.42
			90k	37.74	**67.35**	**41.38**	49.72	**30.18**	**33.32**

Table 1: Performance comparisons of 2D instance segmentation on Waymo val. dataset. Abbreviations: vehicle (Veh.), pedestrian (Ped.), cyclist (Cyc.). 2D Box and 3D Box representations use boxes as weakly supervised annotations.

Supervision	Annotation	Model	mAP			AP			IoU			
			AP	AP50	AP75	Veh.	Ped.	Cyc.	mIoU	Veh.	Ped.	Cyc.
Full	3D Mask	SparseUNet (2020)	57.59	65.52	59.99	80.30	57.18	35.29	80.22	96.67	82.47	61.51
Weak	3D Box		48.29	63.49	55.42	68.95	45.94	29.99	72.70	89.68	73.58	54.85
	SAM		43.34	54.90	45.98	**64.01**	39.89	26.13	75.59	**93.47**	77.98	55.31
	2D Box		34.18	49.95	38.01	44.18	36.43	21.94	63.38	73.69	71.55	44.91
		MWSIS (ours)	**46.93**	**56.58**	**49.78**	62.46	**45.73**	**32.59**	**75.93**	90.45	**79.00**	**58.33**

Table 2: Performance comparisons of 3D instance and semantic segmentation on Waymo val. dataset. SAM represents the instance masks obtained using 2D GT boxes as prompts, which are adopted as weakly supervised annotations.

Implementation Details

Evaluation Metric. For 2D and 3D instance segmentation, we follow the COCO evaluation method using AP (average precision over IoU thresholds), AP50 (IoU is 0.50), and AP75 (IoU is 0.75) to evaluate our method. For 3D semantic segmentation, we use the standard mIoU to evaluate the results.

Training Setting. We train our model for 24 epochs with a batch size of 8 on 4 A6000 GPUs. Each batch contains 1 frame of the point cloud and 5 images. The 2D network adopts the SGD optimizer at a learning rate of 0.01, while the 3D network employs the AdamW optimizer with a one-cycle learning rate policy, setting the maximum rate to 0.001. Our framework is based on the mmdetection framework.

Data Augmentation. In our framework, the 2D and 3D modalities are mutually independent, and we can decouple the data augmentation for multimodal data through the point-pixel mapping relationship. For the 2D branch, we employ resizing, random flipping, normalization, and padding as data augmentation methods, while for the 3D branch, we apply global rotation, global translation, global scaling, random flipping, and shuffling as data augmentation methods.

Results

2D Instance Segmentation. We compare our method with competitive fully supervised, and weakly supervised instance segmentation methods on Waymo dataset. As shown in Tab. 1, our method achieves a 2.59% mAP improvement over BoxInst and outperforms the fully supervised approach in the cyclist class. Moreover, our method achieves comparable performance to the LWSIS, which utilizes more precise guidance of points inside the 3D box.

3D Instance and Semantic Segmentation. We compare the performance of our MWSIS with full supervision and other weak annotations. As shown in Tab. 2, we use the data processed by clustering as the baseline for comparison in row 4. The MWSIS improves the baseline by 12.75% mAP on the instance segmentation task. On the semantic segmentation task, our method achieves 94.65% of the fully supervised performance, and surpasses the performance based on 3D box annotations in all classes.

SAM with 3D Segmentation. SAM (Kirillov et al. 2023) uses points, boxes, masks, and texts as prompts to segment objects in the image. We use 2D boxes as our prompts to obtain masks. The mapping relationship between the point cloud and image pixels is employed to obtain masks for the point cloud. To further obtain final mask labels, we apply the CCL algorithm to the point cloud. Our model exhibits 3.59% mAP higher than that using SAM alone.

The Potential for Downstream Tasks. To verify the potential of our 3D label correction method, we pre-train the 3D backbone on Waymo training set and then fine-tune it on downstream tasks. In Fig. 7, we compare the instance segmentation performance after fine-tuning with different proportions of full supervision data. The network outperforms full supervision with only 5% of the full supervision data. Additionally, it is shown in Tab. 3 that our pre-training backbone can improve the performance of 3D object detectors like FSD (Fan et al. 2022).

Ablation Studies

3D Pseudo Label Generation. Fig. 3 shows comparisons of pseudo labels generated by the SPG with other methods. It can be seen that the quality of pseudo labels generated by the SPG is higher than that of other methods, even SAM (Kirillov et al. 2023). In rows 4 and 5 of Tab. 4, the SPG performance achieves 8.14% mAP higher than that using the CCL directly (baseline), showing that the SPG can effectively improve the quality of pseudo labels by using the spatial prior information of the point cloud. In rows 5 and 6 of Tab. 4, the PVC has 2.32% mAP higher than that using the SPG, which shows that the PVC corrects the noise labels in the network by voting. Compared with the 3rd and 4th rows, 8th and 9th

Model	Method	mAP/mAPH		*Veh.* AP/APH		*Ped.*AP/APH		*Cyc.*AP/APH	
	Pre-training	L1	L2	L1	L2	L1	L2	L1	L2
FSD	-	78.7/76.3	71.9/69.7	77.8/77.3	68.9/68.5	81.9/76.4	73.2/68.0	76.5/75.2	73.8/72.5
(2022)	✓	**79.2/76.8**	**72.4/70.2**	**78.1/77.6**	**69.3/68.9**	81.8/76.3	73.0/67.9	**77.6/76.5**	**74.8/73.7**

Table 3: 3D detection performance on Waymo val. dataset. Our pre-training backbone improves the detection performance of the FSD.

Supervision	Annotation	Model	Module					Metric	
			CCL	SPG	PVC	RSC	CSCS	mAP	mIoU
Weak	SAM	SparseUNet (2020)	-	-	-	-	-	28.57	70.39
			✓	-	-	-	-	43.34	75.59
			✓	✓	✓	-	-	45.06	76.63
			✓	✓	✓	✓	-	45.54	77.41
	2D Box		-	-	-	-	-	4.63	42.94
			✓	-	-	-	-	34.18	63.38
			✓	✓	-	-	-	42.32	73.87
			✓	✓	✓	-	-	44.71	74.75
			✓	✓	✓	✓	-	45.76	**76.29**
			✓	✓	✓	✓	✓	**46.93**	75.93

Table 4: Ablation studies for 3D instance and semantic segmentation on Waymo val. dataset.

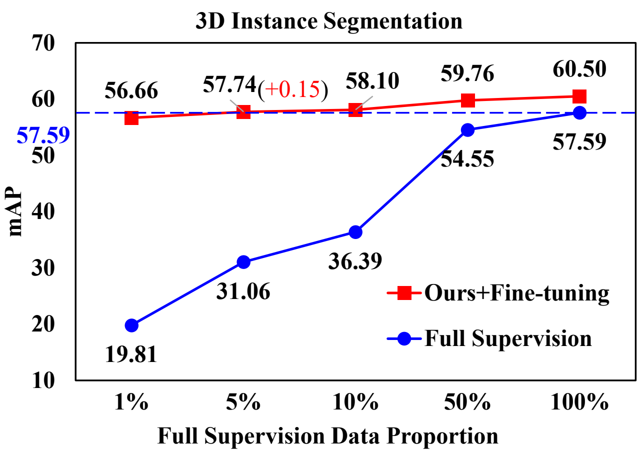

Figure 7: 3D instance segmentation on Waymo val. dataset. The weakly supervised pre-training backbone for fine-tuning compared to the fully supervised performance.

Model	Task	$L_{cscs}^{2d\rightarrow 3d}$	$L_{cscs}^{3d\rightarrow 2d}$	mAP
BoxInst (2021) + IPG	2D	-	-	36.77
		-	✓	36.99
		✓	✓	**37.20**
SparseUNet (2020)	3D	-	-	45.76
		✓	-	46.37
		✓	✓	**46.93**

Table 5: Comparison of different supervision methods.

rows of Tab. 4, the performance improves by 0.48% mAP and 1.05% mAP, respectively, after incorporating the RSC module. These experiments reveal that the depth prior ring segment can effectively correct the predictions.

Consistency Sparse Cross-modal Supervision (CSCS). In Tab. 5, after applying the CSCS, the segmentation performances of 2D and 3D networks are improved by 0.43% mAP and 1.17% mAP, respectively, which confirms the effectiveness of multimodal response distillation.

Conclusion

This paper proposes a novel multimodal instance segmentation framework with weak annotations. It explores two questions: one is how to correct pseudo labels with solely 2D box annotations, and the other is how to train multimodal segmentors simultaneously. To address the former challenge, we introduce various fine-grained pseudo label generation and correction methods to improve the quality of pseudo labels. For the second problem, our framework decouples the two modal branches and trains the multimodal segmentors simultaneously. Additionally, a cross-modal supervision module CSCS is proposed to utilize the teacher segmentors' output for response distillation by the point-pixel mapping relationship. By these methods, the multimodal segmentors achieve efficient training and reliable performance. Compared to the baseline, our method achieves 2.59% mAP and 12.75% mAP improvement on the 2D and 3D instance segmentation tasks, respectively. Moreover, the 3D segmentor outperformes the full supervision using only 5% of the fully supervised annotations, and can also serve as a pre-training method to improve the performance on the other 3D downstream tasks.

Acknowledgments

This work was supported in part by the Youth Innovation Promotion Association CAS (CX2100060053), and the Natural Science Foundation of Anhui Province under Grant 2208085J17. We sincerely appreciate all the reviewers. They give positive and high-quality comments on our paper with a lot of constructive feedback to help us improve the paper.

References

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. A Dataset for Semantic Segmentation of Point Cloud Sequences. *ArXiv preprint*, abs/1904.01416.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 11618–11628. IEEE.

Chen, X.; Cheung, Y. S. J.; Lim, S. N.; and Zhao, H. 2023. ScribbleSeg: Scribble-based Interactive Image Segmentation. *ArXiv preprint*, abs/2303.11320.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-Supervised Semantic Segmentation With Cross Pseudo Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2613–2622. Computer Vision Foundation / IEEE.

Cheng, B.; Parkhi, O.; and Kirillov, A. 2022. Pointly-Supervised Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2607–2616. IEEE.

Cheng, B.; Schwing, A. G.; and Kirillov, A. 2021. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 17864–17875.

Chibane, J.; Engelmann, F.; Tran, T. A.; and Pons-Moll, G. 2022. Box2Mask: Weakly Supervised 3D Semantic Instance Segmentation using Bounding Boxes. In *European Conference on Computer Vision*.

Cholakkal, H.; Sun, G.; Khan, F. S.; and Shao, L. 2019. Object Counting and Instance Segmentation With Image-Level Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 12397–12405. Computer Vision Foundation / IEEE.

Dong, S.; Li, R.; Wei, J.; Liu, F.; and Lin, G. 2022. RWSeg: Cross-graph Competing Random Walks for Weakly Supervised 3D Instance Segmentation. *ArXiv preprint*, abs/2208.05110.

Fan, L.; Wang, F.; Wang, N.; and ZHANG, Z.-X. 2022. Fully sparse 3d object detection. *Advances in Neural Information Processing Systems*, 35: 351–363.

Ge, W.; Huang, W.; Guo, S.; and Scott, M. R. 2019. Label-PEnet: Sequential Label Propagation and Enhancement Networks for Weakly Supervised Instance Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 3344–3353. IEEE.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, 3354–3361. IEEE Computer Society.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2980–2988. IEEE Computer Society.

Hennequin, A.; Lacassagne, L.; Cabaret, L.; and Meunier, Q. 2018. A new Direct Connected Component Labeling and Analysis Algorithms for GPUs. In *2018 Conference on Design and Architectures for Signal and Image Processing (DASIP)*. Porto, Portugal.

Hsu, C.; Hsu, K.; Tsai, C.; Lin, Y.; and Chuang, Y. 2019. Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 6582–6593.

Hu, Q.; Yang, B.; Fang, G.; Guo, Y.; Leonardis, A.; Trigoni, N.; and Markham, A. 2021. SQN: Weakly-Supervised Semantic Segmentation of Large-Scale 3D Point Clouds with 1000x Fewer Labels. In *European Conference on Computer Vision*.

Jiang, L.; Zhao, H.; Shi, S.; Liu, S.; Fu, C.; and Jia, J. 2020. PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 4866–4875. IEEE.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. B. 2023. Segment Anything. *ArXiv preprint*, abs/2304.02643.

Lee, J.; Kim, C.; and Sull, S. 2021. Weakly Supervised Segmentation of Small Buildings with Point Labels. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 7386–7395. IEEE.

Lee, J.; Yi, J.; Shin, C.; and Yoon, S. 2021. BBAM: Bounding Box Attribution Map for Weakly Supervised Semantic and Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2643–2652. Computer Vision Foundation / IEEE.

Li, X.; Yin, J.; Shi, B.; Li, Y.; Yang, R.; and Shen, J. 2023. Lwsis: Lidar-guided weakly supervised instance segmentation for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1433–1441.

Liu, C.; Qian, X.; Huang, B.; Qi, X.; Lam, E. Y.; Tan, S.-C.; and Wong, N. 2022a. Multimodal Transformer for Automatic 3D Annotation and Object Detection. In *European Conference on Computer Vision*.

Liu, C.; Qian, X.; Qi, X.; Lam, E. Y.; Tan, S.-C.; and Wong, N. 2022b. MAP-Gen: An Automated 3D-Box Annotation Flow with Multimodal Attention Point Generator. *2022 26th*

International Conference on Pattern Recognition (ICPR), 1148–1155.

Mei, J.; Zhu, A. Z.; Yan, X.; Yan, H.; Qiao, S.; Zhu, Y.; Chen, L.-C.; Kretzschmar, H.; and Anguelov, D. 2022. Waymo Open Dataset: Panoramic Video Panoptic Segmentation. *ArXiv preprint*, abs/2206.07704.

Peng, L.; Yan, S.; Wu, B.; Yang, Z.; He, X.; and Cai, D. 2022. WeakM3D: Towards Weakly Supervised Monocular 3D Object Detection. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Qin, Z.; Wang, J.; and Lu, Y. 2020. Weakly supervised 3d object detection from point clouds. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4144–4152.

Shen, Y.; Cao, L.; Chen, Z.; Zhang, B.; Su, C.; Wu, Y.; Huang, F.; and Ji, R. 2021. Parallel Detection-and-Segmentation Learning for Weakly Supervised Instance Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 8178–8188. IEEE.

Shi, S.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE transactions on pattern analysis and machine intelligence*, 43(8): 2647–2664.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; Vasudevan, V.; Han, W.; Ngiam, J.; Zhao, H.; Timofeev, A.; Ettinger, S.; Krivokon, M.; Gao, A.; Joshi, A.; Zhang, Y.; Shlens, J.; Chen, Z.; and Anguelov, D. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2443–2451. IEEE.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 1195–1204.

Tian, Z.; Shen, C.; and Chen, H. 2020. Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 282–298. Springer.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 9626–9635. IEEE.

Tian, Z.; Shen, C.; Wang, X.; and Chen, H. 2021. BoxInst: High-Performance Instance Segmentation With Box Annotations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 5443–5452. Computer Vision Foundation / IEEE.

Unal, O.; Dai, D.; and Gool, L. V. 2022. Scribble-Supervised LiDAR Semantic Segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2687–2697.

Wang, X.; Feng, J.; Hu, B.; Ding, Q.; Ran, L.; Chen, X.; and Liu, W. 2021. Weakly-Supervised Instance Segmentation via Class-Agnostic Learning With Salient Images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 10225–10235. Computer Vision Foundation / IEEE.

Wei, J.; Lin, G.; Yap, K.; Hung, T.; and Xie, L. 2020. Multi-Path Region Mining for Weakly Supervised 3D Semantic Segmentation on Point Clouds. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 4383–4392. IEEE.

Wei, Y.; Su, S.-C.; Lu, J.; and Zhou, J. 2021. FGR: Frustum-Aware Geometric Reasoning for Weakly Supervised 3D Vehicle Detection. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 4348–4354.

wei Li, Z.; Zheng, Y.; Luo, X.; Shan, D.; and Hong, Q. 2023. ScribbleVC: Scribble-supervised Medical Image Segmentation with Vision-Class Embedding. *ArXiv preprint*, abs/2307.16226.

Wu, Y.; Shi, M.; Du, S.; Lu, H.; Cao, Z.; and Zhong, W. 2022. 3D Instances as 1D Kernels. In *European Conference on Computer Vision*.