LISD: AN EFFICIENT MULTI-TASK LEARNING FRAMEWORK FOR LIDAR SEGMENTATION AND DETECTION

Jiahua Xu[⋆]   *Si Zuo*[†]   *Chenfeng Wei*[⋆✉]   *Wei Zhou*[‡✉]

[⋆]Wuxi Intelligent Control Research Institute, Hunan University, China
[†]College of Mechanical and Vehicle Engineering, Hunan University, China
[‡]School of Computer Science and Informatics, Cardiff University, United Kingdom

ABSTRACT

With the rapid proliferation of autonomous driving, there has been a heightened focus on the research of lidar-based 3D semantic segmentation and object detection methodologies, aiming to ensure the safety of traffic participants. In recent decades, learning-based approaches have emerged, demonstrating remarkable performance gains in comparison to conventional algorithms. However, the segmentation and detection tasks have traditionally been examined in isolation to achieve the best precision. To this end, we propose an efficient multi-task learning framework named LiSD which can address both segmentation and detection tasks, aiming to optimize the overall performance. Our proposed LiSD is a voxel-based encoder-decoder framework that contains a hierarchical feature collaboration module and a holistic information aggregation module. Different integration methods are adopted to keep sparsity in segmentation while densifying features for query initialization in detection. Besides, cross-task information is utilized in an instance-aware refinement module to obtain more accurate predictions. Experimental results on the nuScenes dataset and Waymo Open Dataset demonstrate the effectiveness of our proposed model. It is worth noting that LiSD achieves the state-of-the-art performance of 83.3% mIoU on the nuScenes segmentation benchmark for lidar-only methods.

Index Terms— multi-task learning, semantic segmentation, object detection

1. INTRODUCTION

Semantic segmentation and 3D object detection tasks play pivotal roles in autonomous driving, serving as foundational components for establishing a comprehensive environmental perception system, which is crucial for mitigating labor costs and ensuring traffic safety. To facilitate advancements in this field, large-scale databases such as nuScenes [1] and Waymo Open Dataset (WOD) [2] have been created, which are invaluable resources for guiding the refinement and evaluation of the perception algorithms designed to handle complicated
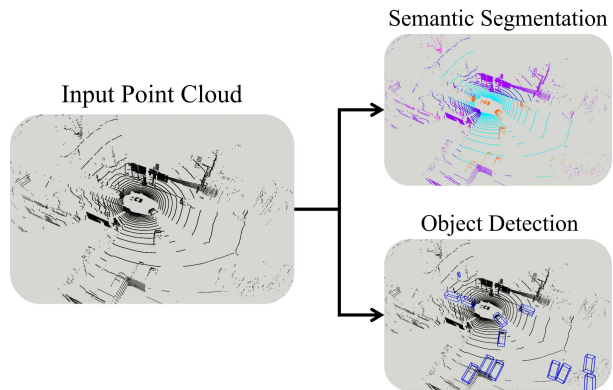

Fig. 1. The proposed LiSD model adopts point cloud data as its input and simultaneously produces semantic segmentation and object detection results.

road conditions. Based on the aforementioned databases, semantic segmentation and object detection tasks are conventionally researched independently to accomplish the best accuracy, for example, frameworks such as Cylinder3D are purposefully tailored for semantic segmentation, which can not yield remarkable performance compared to frameworks explicitly designed for object detection [3]. Hence, there exists an urgent need to investigate unified frameworks for achieving optimal performance across both semantic segmentation and object detection tasks. Concurrently, the generation of segmentation and detection outcomes within a singular inference process proves time-saving, presenting an advantage over the separate execution of distinct tasks [4].

Recently, deep learning frameworks have achieved considerable success in the domain of lidar perception, which can be roughly categorized into point-based, voxel-based, Range View (RV) based, Bird's Eye View (BEV) based, and hybrid methodologies. Voxel-based methods have emerged as the predominant paradigm in segmentation and detection tasks, owing to the development of sparse convolution techniques [5]. Standard sparse convolution yields output points when there existing related points in the receptive field. In contrast,
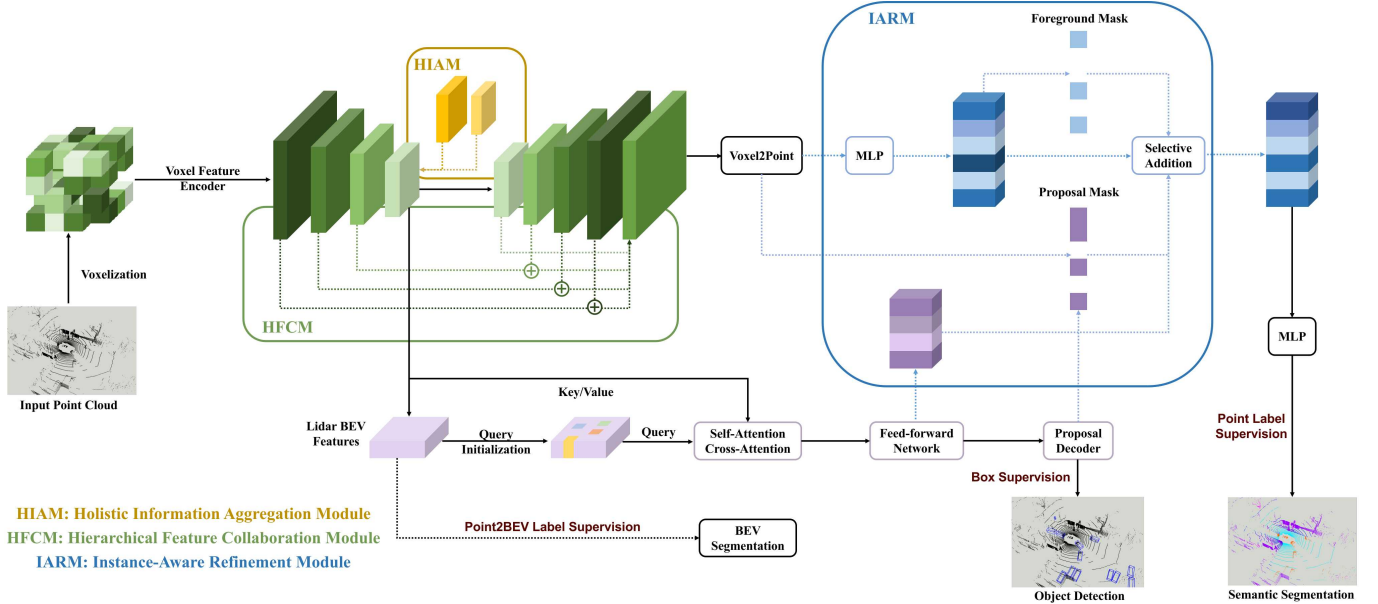
Fig. 2. The architecture of our proposed LiSD is outlined, where the point cloud serves as input for both segmentation and detection tasks. HIAM and HFCM are adopted in the voxel-based encoder-decoder to integrate holistic and hierarchical information for both tasks. Additionally, IARM is introduced to directly refine the point-level feature representation and indirectly exert influence on the box regression.

submanifold sparse convolution limits the output location as active only if the corresponding input location is active [6]. Submanifold convolution becomes indispensable in 3D networks to reduce memory consumption, which implies the receptive field is constrained when adopting submanifold convolution [5]. VoxelNeXt [7] incorporates additional downsampling and sparse height compression to generate robust feature representation with sufficient receptive fields. Nevertheless, these operations will induce alterations in the density of the input sparse feature, consequently increasing the difficulty in the implementation of inverse sparse convolution. To address this problem, Ye *et al.* propose the global context pooling (GCP) in the multi-task framework LidarMulti-Net [8]. GCP transforms the 3D sparse features into 2D BEV dense features to extract the global information. Moreover, a cross-space transformer module is adopted in LiDARFormer [9] to learn long-range information in the BEV feature. However, direct extraction of the global information from the conversion of 2D BEV dense features leads to an increase in memory consumption, given the storage of inactive points.

Regarding the cross-task information interaction during multi-task learning, LidarMTL [4] adopts a conventional approach in which only low-level features are shared. In Lidar-MultiNet [8], a second-stage refinement module is introduced to enhance the first-stage semantic segmentation and produce the panoptic segmentation results. Zhou *et al.* adopt a cross-task module [9] to transfer high-level features through cross-task attention mechanisms with high computational complex-

ity.

In this paper, to reduce memory consumption and computational complexity while keeping precision, we propose an efficient multi-task learning framework, denoted as LiSD, for lidar semantic segmentation and object detection as shown in Fig. 1. Instead of the direct placement of voxels from various scales onto the ground used in VoxelNeXt, we introduce a memory-friendly holistic information aggregation module, which interpolates the high-level features to the relevant active positions of low-level features. Through this methodology, sparsity is preserved with the acquisition of global information. Besides, the hierarchical feature collaboration is adopted in our LiSD to enhance the voxel feature representation. Moreover, in contrast to the aforementioned cross-task information interaction methodologies, we propose a straightforward yet effective instance-aware refinement module. This module is specifically designed to enhance the feature representation of foreground points through the incorporation of proposal features. LiSD is evaluated on two databases, namely nuScenes and WOD, demonstrating competitive performance for both segmentation and detection tasks. Notably, LiSD attains a leading segmentation performance of 83.3% mIoU on nuScene, outperforming all the lidar-based methods currently ranked on the leaderboard.

The main contributions can be listed as follows:

- We propose an efficient multi-task learning framework LiSD for lidar semantic segmentation and object detection.

- We introduce a memory-friendly holistic information aggregation module (HIAM) to integrate global information suitable for segmentation and detection tasks, and a hierarchical feature collaboration module (HFCM) to enhance the voxel feature representation.

- We present an instance-aware refinement module (IARM) to improve the foreground point feature representation with the assistance of object proposals.

- The proposed LiSD achieves competitive performance in segmentation and detection tasks on the nuScenes and WOD datasets.

2. METHOD

In this section, we delineate the structure of our multi-task learning framework, LiSD, which seamlessly integrates three perception tasks, namely, semantic segmentation, object detection, and the auxiliary BEV segmentation through a single feed-forward pass, as illustrated in Fig. 2.

2.1. Overview

Given the input point cloud $P = \{p_i \mid p_i \in \mathbb{R}^{3+c}\}_{i=1}^N$, the proposed LiSD yields semantic segmentation labels $L = \{l_i \mid l_i \in (1 \cdots K)\}_{i=1}^N$ and object detection bounding boxes $B = \{b_i \mid b_i \in \mathbb{R}^9\}_{i=1}^M$, where N represents the number of points, M, K denote the number of predicted boxes and semantic classes, respectively. Each point is endowed with $(3 + c)$-dimensional features, e.g., 3D coordinate, intensity, elongation, timestamp, etc. The predicted boxes are characterized by the center coordinates, sizes, orientations, and velocities.

Firstly, the voxelized point cloud is fed into the Voxel Feature Encoder (VFE) for producing the sparse voxel feature representation $F_{v_j} = mean(p_i), i = 1 \cdots n$ through an average pooling layer, where p_i denotes the feature representation of the i-th point within the voxel v_j, and v_j contains a total of n points. Then, the voxel-based encoder-decoder formed from 3D sparse convolution is adopted to generate voxel and BEV feature representation for segmentation and detection tasks. The encoder comprises four stages of sparse convolution blocks to downsample the spatial resolution, thereby acquiring high-level voxel features for the detection head. Conversely, the decoder is equipped with four symmetrical stages of sparse inverse convolution blocks to recover to the original voxel size for the segmentation head. The holistic information aggregation module and the hierarchical feature collaboration module are introduced in the encoder-decoder to enlarge the receptive field and enhance the feature representation. Ultimately, the segmentation and detection heads are employed to produce the semantic labels and object bounding boxes. An instance-aware refinement module is designed
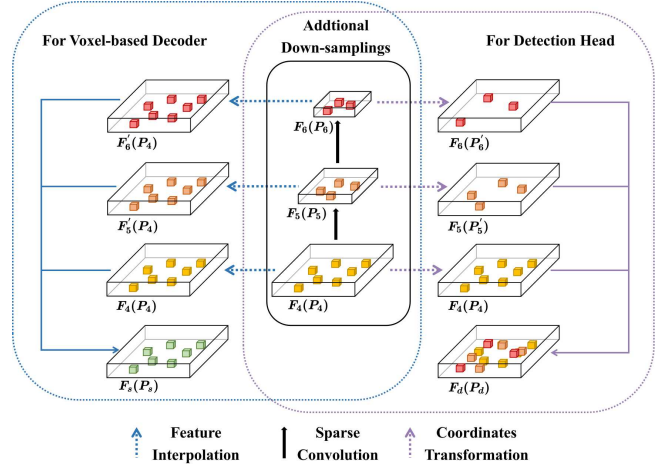

Fig. 3. The detailed structure of the proposed HIAM. Additional down-samplings are adopted to acquire holistic information. Subsequently, feature interpolation is employed to aggregate the information for the voxel-based decoder preceding the segmentation head, while coordinates transformation is utilized to integrate the information for the detection head.

to integrate cross-task information to improve the accuracy of the prediction results.

2.2. Holistic Information Aggregation Module

As inspired by VoxelNeXt [7], sufficient receptive fields are required to ensure correct predictions when dealing with sparse voxel features. Additional two stages of down-sampling are introduced in VoxelNeXt to generate features $\{F_5, F_6\}$ with strides $\{16, 32\}$, and the multi-scale features of the original encoder with strides $\{1, 2, 4, 8\}$ are denoted as $\{F_1, F_2, F_3, F_4\}$. The enhanced feature representation F_d for the detection head is obtained as follows:

$$F_d = F_4 \cup F_5 \cup F_6$$
$$P_d = P_4 \cup P_5^{'} \cup P_6^{'} \tag{1}$$

where P_d represents the position of the enhanced voxel feature, and P_4, P_5, P_6 correspond to the positions of F_4, F_5, F_6, respectively. $P_5^{'}$ is aligned to P_4 by doubling the 3D coordinates $(x_{p_5}, y_{p_5}, z_{p_5})$ of the position $p_5 \in P_5$, and the same as $P_6^{'}$. Nevertheless, the produced feature F_d exhibits a significantly higher density compared to F_4. The different sparsity of F_d and F_4 poses challenges in the implementation of inverse sparse convolution.

To solve this problem, we take another approach to integrate holistic information for the voxel-based decoder, as depicted in Fig. 3. Specifically, the voxel features in $F_5^{'}$ corresponding to the position P_4 are interpolated with neighboring voxel features in F_5 to maintain sparsity, and this process is

replicated for F_6'. The enhanced feature representation F_s for the voxel-based decoder is denoted as follows:

$$F_s = F_4 + F_5' + F_6'$$
$$P_s = P_4 \tag{2}$$

where P_s denotes the position of the refined voxel feature, identical to P_4. Consequently, the implementation of inverse sparse convolution becomes straightforward. Note that F_d is further projected to the BEV feature map $\overline{F}_d$ by putting all voxels onto the ground and summing up features in the same positions. Leveraging the HIAM, the proposed LiSD could significantly extend the receptive fields, which is a crucial aspect for improving semantic segmentation. Furthermore, the BEV feature map obtained through HIAM is further leveraged by object detection and BEV segmentation tasks.

2.3. Hierarchical Feature Collaboration Module

As indicated in [10], hierarchical features possess robust semantic information across various scales, which is beneficial for the semantic segmentation task. Apart from the 3D U-Net architecture, we design an additional hierarchical feature collaboration module to augment the voxel feature representation for the segmentation head. As previously noted, the multi-scale features of the encoder and decoder with strides $\{1, 2, 4, 8\}$ are represented as $\{F_1, F_2, F_3, F_4\}$ and $\left\{ F_1', F_2', F_3', F_s \right\}$, respectively. The enhanced feature $\overline{F}_s$ is obtained via the top-down pathway and lateral connections as follows:

$$\overline{F}_s = cat \left(\theta_1(F_1 + F_1'), \theta_2(F_2 + F_2'), \theta_3(F_3 + F_3'), \theta_s F_s \right) \tag{3}$$

where θ_1, θ_2, θ_3, θ_s represent the encoding and up-scaling functions applied to hierarchical features, and cat denotes the concatenation operation across feature channels.

2.4. Instance-Aware Refinement Module

Given that the devised LiSD is a multi-task learning framework, there exists the potential for the integration of cross-task information between the segmentation and detection heads. As illustrated in Fig. 2, voxel features undergo an initial conversion to point features for foreground probability estimation, which is guided by semantic segmentation labels during the training phase. The foreground mask m_{fi} of the i-th point is denoted as:

$$m_{fi} = \begin{cases} 1, & p_{fi} \geq \delta_f \\ 0, & p_{fi} < \delta_f \end{cases} \tag{4}$$

where p_{fi} represents the foreground probability of the i-th point, and δ_f indicates the probability threshold to distinguish between foreground and background points, e.g. $\delta_f = 0.5$ in this paper. Simultaneously, the detection head predicts M

initial boxes, and the proposal mask m_{bij} of the i-th point (p_i) to j-th box (b_j) is computed as:

$$m_{bij} = \begin{cases} 1, & p_i \ inside \ b_j \\ 0, & p_i \ outside \ b_j \end{cases} \tag{5}$$

Then, the refined point feature f_{pi}' of the i-th point is obtained as:

$$f_{pi}' = f_{pi} + \sum_{j=0}^{M} m_{fi} \cdot m_{bij} \cdot \mathrm{MLP}(f_{bj}) \tag{6}$$

where f_{pi} represents the feature of the i-th point before instance-aware refinement, f_{bj} represents the feature of the j-th box generated from the prediction head. MLP denotes the Multi-Layer Perceptron to adjust the feature dimension of f_{bj} to align with that of f_{pi}. Benefiting from the IARM, the feature representation of the foreground points is enhanced with the incorporation of the proposal features, and constraints are applied to the box regression process concurrently.

2.5. Joint Training

LiSD is trained in an end-to-end manner via a multi-task loss function. Specifically, for the semantic and BEV segmentation tasks, the optimization is guided by the cross-entropy loss and Lovasz loss [11]. For the detection task, binary cross-entropy loss, L_1 loss, and IoU loss [12] are employed to minimize the classification, regression, and IoU cost, respectively. Subsequently, the final loss L is defined as a weighted sum of the task-specific losses:

$$L = \sum_{i \in \{seg, bev, det\}} \frac{1}{2\sigma_i^2} L_i + \frac{1}{2} \log \sigma_i^2 \tag{7}$$

where σ_i denotes the noise parameter for task i to compute task-dependent uncertainty [13]. Hence, as the uncertainty increases for task i, the contribution of the task-specific loss L_i to L diminishes.

3. EXPERIMENTS

In this section, the databases used in our experiment are introduced at first. Then, we conduct the performance comparison of LiSD and other methods on these databases. Finally, the effectiveness of the specially designed modules is verified via the ablation study.

3.1. Datasets

Two large-scale autonomous driving databases equipped with point-wise semantic labels and 3D object bounding box annotations are utilized in our experiment, namely nuScenes dataset [1] and WOD dataset [2].

NuScenes dataset [1]: This database includes 1000 scenarios, each lasting 20 seconds and captured using a 32-beam

Table 1. Segmentation results for each class on the *test* split of the nuScenes dataset. The best and second-best performing results are marked in boldface and underlined.

Model	mIoU	barrier	bicycle	bus	car	construction	motorcycle	pedestrian	traffic cone	trailer	truck	driveable	other flat	sidewalk	terrain	manmade	vegetation
PolarNet [14]	69.8	80.1	19.9	78.6	84.1	53.2	47.9	70.5	66.9	70.0	56.7	96.7	68.7	77.7	72.0	88.5	85.4
PolarStream [15]	73.4	71.4	27.8	78.1	82.0	61.3	77.8	75.1	72.4	79.6	63.7	96.0	66.5	76.9	73.0	88.5	84.8
SPVNAS [16]	77.4	80.0	30.0	91.9	90.8	64.7	79.0	75.6	70.9	81.0	74.6	97.4	69.2	80.0	76.1	89.3	87.1
Cylinder3D++ [3]	77.9	82.8	33.9	84.3	89.4	69.6	79.4	77.3	73.4	84.6	69.4	97.7	70.2	80.3	75.5	90.4	87.6
AF2S3Net [17]	78.3	78.9	52.2	89.9	84.2	77.4	74.3	77.3	72.0	83.9	73.8	97.1	66.5	77.5	74.0	87.7	86.8
SPVCNN++ [16]	81.1	86.4	43.1	91.9	92.2	75.9	75.7	83.4	77.3	86.8	77.4	97.7	71.2	81.1	77.2	91.7	89.0
LidarMultiNet [8]	81.4	80.4	48.4	94.3	90.0	71.5	87.2	85.2	80.4	86.9	74.8	97.8	67.3	80.7	76.5	92.1	89.6
LidarFormer [9]	81.5	84.4	40.8	84.7	92.6	72.7	91.0	84.9	81.7	88.6	73.8	97.9	69.3	81.4	77.4	92.4	89.6
UDeerPep	81.8	85.5	55.5	90.5	91.6	72.2	85.6	81.4	76.3	87.3	74.0	97.7	70.2	81.1	77.4	92.7	90.2
Proposed LiSD	83.3	82.1	67.1	89.8	92.2	80.5	86.9	87.4	79.3	86.6	76.1	97.5	67.2	80.5	77.0	92.3	89.7

lidar sensor at a sampling rate of 20Hz. Keyframes within each scenario are annotated with a 2Hz sampling rate. For the semantic segmentation task, point-wise semantic labels are provided for 16 categories, including 10 foreground classes and 6 background classes, and mean Intersection over Union (mIoU) is employed as the evaluation metric. For the object detection task, bounding box annotations are provided for the same 10 foreground categories as those in the segmentation task, and the evaluation metrics include mean Average Precision (mAP) and NuScenes Detection Score (NDS).

WOD dataset [2]: This dataset contains 2000 scenarios captured by a 64-beam lidar sensor at a sampling rate of 10Hz. Similar to nuScenes, each scenario spans 20 seconds, with detection annotations available for all frames, while point-wise semantic labels are only provided for selected keyframes. For the semantic segmentation task, semantic labels are provided for 23 classes, and the evaluation metric employed is mIoU as well. For the object detection task, bounding box annotations are provided for 3 categories, e.g., vehicles, pedestrians, and cyclists. Average Precision Weighted by Heading (APH) is used as the evaluation metric for detection, and the ground truth objects are categorized as LEVEL_1 (L1) and LEVEL_2 (L2) samples based on the detection difficulty. mAPH L1 is calculated by considering samples labeled as L1, while mAPH L2 is computed by incorporating both L1 and L2 samples.

3.2. Experiment Setup

The AdamW optimizer, coupled with a one-cycle scheduler, is adopted to train the proposed LiSD for 65 epochs, with a max learning rate of 3e-3 and a weight decay of 0.01. The voxel size of cloud range $[-54.0, 54.0]\,m \times [-54.0, 54.0]\,m \times [-5.0, 4.6]\,m$ is set as $[0.075, 0.075, 0.2]$ for nuScenes, and voxel size of cloud range $[-75.2, 75.2]\,m \times [-75.2, 75.2]\,m \times$

$[-2, 4]\,m$ is set as $[0.1, 0.1, 0.15]$ for WOD. Standard data augmentation techniques, including flipping, scaling, rotation, translation, and ground-truth sampling [5] with fade strategy [18] are utilized during training. In our experiment, 6 Nvidia A30 GPUs with a batch size of 18 are employed for NuScenes, while a batch size of 12 is configured for WOD.

Table 2. Comparison results on the *val* split of the nuScenes dataset. Reported by [9]. The best and second-best performing results are marked in boldface and underlined.

	Model	mIoU	mAP	NDS
Segmentation	PolarNet [14]	71.0	-	-
	Cylinder3D [3]	76.1	-	-
	RPVNet [19]	77.6	-	-
Detection	CenterPoint [20]	-	57.4	65.2
	VoxelNeXt [7]	-	60.5	66.6
	TransFusion-L [21]	-	60.0	66.8
Multi-task	LidarMultiNet [8]	82.0	63.8	69.5
	LidarFormer [9]	82.7	66.6	70.8
	Proposed LiSD	83.0	65.0	69.5

3.3. Experiment Results

Segmentation and detection results for both the nuScenes and WOD datasets are presented to substantiate the effectiveness of LiSD.

NuScenes dataset: Performance comparisons of LiSD and other SOTA methodologies are listed in Table 1, and we can observe that LiSD achieves top segmentation performance of 83.3% mIoU on the *test* split of nuScenes. The mIoU of LiSD is 1.5% higher than that of UDeerPep, establishing its superiority over the best-performing lidar-based methods currently positioned atop the leaderboard. In terms of multi-task models, a second stage is regarded as unneces-

sary in LiSD when compared to LidarMultiNet [8]. Besides, LiSD demonstrates competitive performance when contrasted with the complicated cross-space and cross-task transformers featured in LidarFormer [9]. Furthermore, the detection and segmentation performance on the *val* split is illustrated in Table 2. LiSD outperforms models tailored for segmentation tasks, including PolarNet [14], Cylinder3D [3], and RPVNet [19], in terms of mIoU. Concurrently, LiSD surpasses models specifically designed for detection tasks, such as CenterPoint [20], VoxelNeXt [7], and TransFusion-L [21], by achieving higher mAP.

Table 3. Comparison results on the *val* split of the WOD dataset. Reported by [9]. The best and second-best performing results are marked in boldface and underlined.

	Model	mIoU	L2mAPH
Segmentation	PolarNet [14]	61.6	-
	Cylinder3D [3]	66.6	-
Detection	CenterPoint++ [20]	-	71.6
	CenterFormer [22]	-	73.7
	MPPNet [23]	-	74.9
Multi-task	LidarMultiNet [8]	71.9	75.2
	LidarFormer [9]	72.2	**76.2**
	Proposed LiSD	**72.6**	76.1

WOD dataset: Table 3 illustrates the performance comparison of semantic segmentation and object detection on the *val* split of the WOD dataset. The segmentation results of PolarNet [14] and Cylinder3D [3] are reproduced by [9]. As observed in Table 3, the proposed LiSD attains a mIoU of 72.6% for the segmentation task and an L2mAPH of 76.1% for the detection task, outperforming the single-task model. Meanwhile, LiSD exhibits competitive performance in comparison with multi-task models, such as LidarMultiNet [8] and Lidar-Former [9].

The experimental results of LiSD on both datasets illustrate that multi-task learning facilitates the interaction of information across tasks, thereby contributing to the high performance for both tasks.

3.4. Ablation study

The ablation study, depicted in Table 4, systematically validates the effectiveness of each key component within our proposed LiSD. The baseline segmentation model employed in our experiment is based on the standard 3D sparse U-Net architecture, and the baseline detection model is Transfusion-L [21]. The performance improvements brought by specially designed modules, including HIAM, HFCM, and IARM are further verified on the *val* split of the nuScenes dataset.

As we can observe from Table 4, the baseline model achieves a mIoU of 77.1% and a mAP of 60.0% on the validation set. Building upon the baseline model, the incorporation of the HIAM aimed at enlarging the receptive

Table 4. Ablation study for improvement of mIoU and mAP on the *val* split of the nuScenes dataset.

Baseline	HIAM	HFCM	Multi Task	BEV Loss	IARM	TTA	mIoU	mAP
✓							77.1	60.0
✓	✓						77.9	61.7
✓	✓	✓					79.0	62.1
✓	✓	✓	✓				80.4	62.8
✓	✓	✓	✓	✓			81.1	64.1
✓	✓	✓	✓	✓	✓		81.9	64.5
✓	✓	✓	✓	✓	✓	✓	**83.0**	**65.0**

field yields an improvement of 0.8% in mIoU and 1.7% in mAP. Incorporating the HFCM leads to an increase of 1.1% in mIoU and 0.4% in mAP. The combination of segmentation and detection tasks further improves mIoU by 1.4% and mAP by 0.7%. Besides, through the utilization of BEV segmentation loss coupled with uncertainty weighting, the mIoU and mAP increase by 0.7% and 1.3%. The IARM, designed for the collaboration of cross-task information, brings 0.8% mIoU and 0.4% mAP improvement. The visualization results are depicted in Fig. 4. Forming our optimal model on the validation set, the implementation of Test Time Augmentation (TTA) improves the mIoU and mAP to 83.0% and 65.0%, respectively.
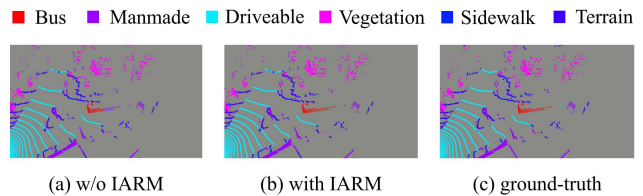

(a) w/o IARM   (b) with IARM   (c) ground-truth

Fig. 4. Semantic segmentation results on the nuScenes dataset. (a) Predicted results of LiSD without IARM, (b) Predicted results of LiSD with IARM, (c) Ground-truth segmentation labels.

4. CONCLUSION

In this paper, we propose an efficient multi-task learning framework named LiSD for lidar segmentation and detection, which are predominantly addressed separately in previous works. Comprehensive experimental results verified the effectiveness of LiSD's design, including key components such as HIAM, HFCM, and IARM. Moreover, LiSD achieves a higher mIoU compared to the top-performing lidar-based method currently positioned on the leaderboard of nuScenes lidar segmentation task. We hope that our proposed LiSD can serve as an inspiration for future endeavors in the development of multi-modal multi-task learning frameworks.

5. REFERENCES

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11621–11631.

[2] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[3] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9939–9948.

[4] Di Feng, Yiyang Zhou, Chenfeng Xu, Masayoshi Tomizuka, and Wei Zhan, "A simple and efficient multi-task network for 3d object detection and road understanding," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7067–7074.

[5] Yan Yan, Yuxing Mao, and Bo Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, pp. 3337, 2018.

[6] Benjamin Graham and Laurens Van der Maaten, "Submanifold sparse convolutional networks," *arXiv preprint arXiv:1706.01307*, 2017.

[7] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia, "Voxelnext: Fully sparse voxelnet for 3d object detection and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21674–21683.

[8] Dongqiangzi Ye, Zixiang Zhou, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh, "Lidarmultinet: Towards a unified multi-task network for lidar perception," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 3231–3240.

[9] Zixiang Zhou, Dongqiangzi Ye, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh, "Lidarformer: A unified transformer-based multi-task network for lidar perception," *arXiv preprint arXiv:2303.12194*, 2023.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[11] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4413–4421.

[12] Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang, "Iou loss for 2d/3d object detection," in *2019 international conference on 3D vision (3DV)*. IEEE, 2019, pp. 85–94.

[13] Alex Kendall, Yarin Gal, and Roberto Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[14] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh, "Polarnet: An improved grid representation for online lidar point clouds semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9601–9610.

[15] Qi Chen, Sourabh Vora, and Oscar Beijbom, "Polarstream: Streaming object detection and segmentation with polar pillars," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26871–26883, 2021.

[16] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han, "Searching efficient 3d architectures with sparse point-voxel convolution," in *European conference on computer vision*. Springer, 2020, pp. 685–702.

[17] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu, "2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12547–12556.

[18] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11794–11803.

[19] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu, "Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16024–16033.

[20] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11784–11793.

[21] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.

[22] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh, "Centerformer: Center-based transformer for 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 496–513.

[23] Xuesong Chen, Shaoshuai Shi, Benjin Zhu, Ka Chun Cheung, Hang Xu, and Hongsheng Li, "Mppnet: Multi-frame feature intertwining with proxy points for 3d temporal object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 680–697.