Top-performing Multi-Modal Solution for 3D Semantic Segmentation of Waymo Open Dataset Challenge

Jiale Li[1] Hang Dai[2] Yong Ding[1]

[1]Zhejiang University

[2]Mohamed bin Zayed University of Artificial Intelligence

jialeli@zju.edu.cn hang.dai@mbzuai.ac.ae dingy@vlsi.zju.edu.cn

Abstract

This technical report presents our top-performing multi-modal solution for the 3D semantic segmentation track in Waymo Open Dataset Challenge at CVPR 2022. Almost all the existing LiDAR 3D semantic segmentation methods take only sparse laser points as input, suffering from inaccurate distinction on small objects. We propose a winning multi-modal method that uses multi-camera images to complement the point cloud. However, multi-modal data also introduces issues on modal heterogeneity and multi-modal data augmentation. To address the former, we propose to perform specific intra-modal feature extraction and inter-modal fusion in a jointly optimized model. The later limitation of multi-modal data augmentation is decoupled as the asymmetric transformations on the point cloud and the images. Besides, the final segmentation performance also benefits from using historical multi-frame point clouds as input, Test-Time Augmentation (TTA), and model ensemble. We achieve the 2^{nd} place with 70.48 mIoU on the official leaderboard.

1. Introduction

The competitive Waymo Open Dataset Challenges at CVPR 2022 ends on May 23 2022, which provides the largest autonomous driving dataset currently publicly available to enable the exciting research [13]. We mainly focus on the 3D semantic segmentation track, which requires to predict one of the 23 semantic categories for each point in the 3D point cloud surrounding the ego-vehicle in the real self-driving scenarios. The subset for 3D semantic segmentation includes 23,691 training samples, 5,976 validation samples and 2,982 testing samples. Given a point cloud sample of N points, the multi-camera images are temporally aligned to this point cloud, and each point with 3D coordinates (x, y, z) is spatially aligned into the image plane with the projected coordinates (u, v).

The 3D semantic segmentation track allows the usage of unlimited historical frame data prior to the current frame, including LiDAR point clouds and multi-camera images. With careful and effective design, our proposed multi-modal 3D segmentation method achieves the 2^{nd} place among all the competing methods.

2. Method

The point cloud data accurately measures the driving scenario using sparse laser points, and the artificial neural network is also powerful enough to learn the structural information of objects from it [2, 9, 14, 20]. But the fine-grained segmentation of some small objects is still challenging for point clouds with sparse laser points. We thus complement point clouds with the color, texture and other information from the corresponding high-resolution multi-camera images for building a top-performing multi-modal 3D semantic segmentation model. However, there is few multi-modal 3D semantic segmentation methods for autonomous driving. We rationally reason about these two factors: **i**) **Heterogeneity between modalities**. Multi-modal models need to be carefully designed with effective intra-modal feature extraction and inter-modal fusion mechanisms to nourish from such heterogeneous data of sparse points and dense pixels. **ii**) **Multi-modal data augmentation**. Compared with the single modality, the multi-modal data augmentation is of more challenges to achieve high quality of inter-modal spatial alignment, while the alignment quality is critical for effective multi-modal association [10]. Some transformations like the rotation of the point cloud cannot be simultaneously applied to the corresponding image, which results in that such useful augmentation transformations are inapplicable.

This section presents our effective and novel solutions to the above issues for multi-modal 3D semantic segmentation of the Waymo Open Dataset challenge. As shown in Fig. 1, we perform feature-level multi-modal fusion through specific intra-modal feature extraction (Sec. 2.1)

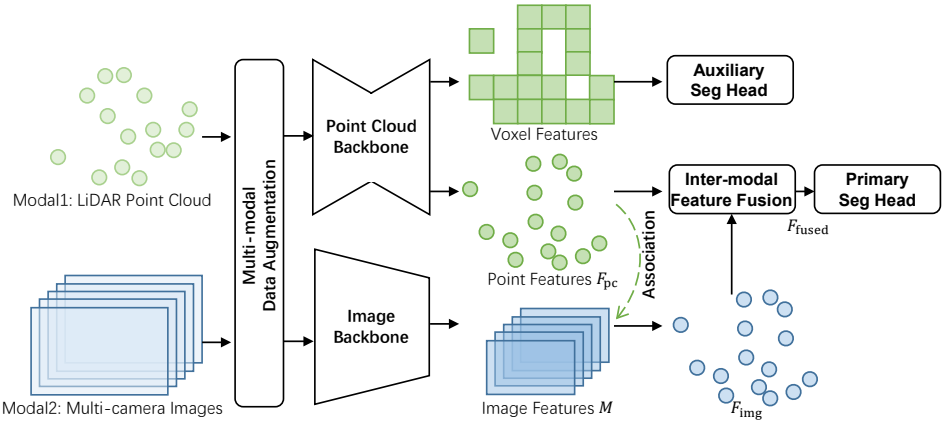

Figure 1. Overview of our multi-modal 3D semantic segmentation ("Seg") model.

and inter-modal feature fusion (Sec. 2.2), then predict the point-wise segmentation using an Multi-Layer Perceptron (MLP) segmentation head based on the point-wise fused feature. Sec. 2.3 introduces the proposed multi-modal data augmentation, followed by the loss function in Sec. 2.4. An optional extension to multi-frame point clouds input is discussed in Sec. 2.5.

2.1. Intra-modal Feature Extraction

Point Cloud Feature Extraction. For effective and efficient point cloud feature learning in a large-scale autonomous driving scenario, we use the voxel-based point cloud backbone as shown in Fig.2.

The input point cloud is divided by a quantization step d and rearranged as non-empty voxels by the Voxel Feature Encoder (VFE), which averages the point-wise initial features (3D coordinates, reflectance and elongation) of the local points inside a voxel. These non-empty voxels are then token as input to a 3D sparse U-Net for 3D convolutional processing. Inspired by the 3D object detection work Part-A^2 [11], we build the 3D sparse U-Net backbone by stacking four down-sampling blocks to make the voxel-wise features more informative with increasing receptive fields, and stack four up-sampling blocks for resolution restoration and feature refinement.

Fig.2 demonstrates the details of the point cloud backbone, which is implemented with the 3D sparse convolution (SparseConv3D), 3D inverse sparse convolution (InverseSparseConv3D) [18], and 3D submanifold convolution (SubMConv3D) [7]. We decompose a pair of down-sampling block D_i and up-sampling block U_i for a clear illustration. More details are in the OpenPCDet [4] toolbox.

For point segmentation, we devoxelize the voxel-wise features into point-wise features F_{pc}. For each point, we interpolate the point feature from its K nearest neighboring voxels [9]. We set the K to 3 for computation efficiency.

Image Feature Extraction. A naive way to enhance the point cloud with images is painting the input points with the RGB values or image segmentation results inferred by a well-trained image segmentation model [15, 17]. However, there are three main limitations: **i**) The image inherently has structural information in dense pixels, and point cloud networks are not designed for image data, which leads to insufficient image feature extraction. **ii**) High-quality image segmentation results require a large number of image annotations for model training, but they are unavailable on Waymo Open Dataset. Models trained on other segmentation datasets will always have distribution gaps and also are not allowed by the challenge rules. **iii**) Inferring the image segmentation results in advance disable the joint optimization of the image network and the point cloud network, which wastes the powerful learning ability of the artificial neural networks.

Instead, we propose to employ an image backbone to provide the hidden feature maps and jointly optimize it in the overall 3D segmentation model. With the clear modular design of our method, our image backbone can be flexibly selected from a large number of mature and off-the-shelf image backbone networks. In practice, we adopt the widely used HRNet-w48 [16], which provides $\times 4$ down-sampled feature map with large receptive fields and rich details. A 1×1 2D convolution is further employed to compress the channels for saving computation, resulting in a feature map $M \in \mathbb{R}^{C_{\text{img}} \times H_o \times W_o}$.

2.2. Inter-modal Feature Fusion

The point cloud features F_{pc} and image features M are specifically extracted with expressive semantic representations, which guarantees the effectiveness of inter-modal feature fusion. Technically, there are two steps to achieve inter-modal feature fusion: association and fusion.

Association. For the i-th point (x_i, y_i, z_i) with point cloud feature $f_{\text{pc},i}$ in F_{pc}, we use the provided point-wise image coordinates $(\frac{u_i}{4}, \frac{v_i}{4})$ to perform bilinear interpolation on the $\times 4$ down-sampled image features M within the corresponding local camera. Zeros are padded to the

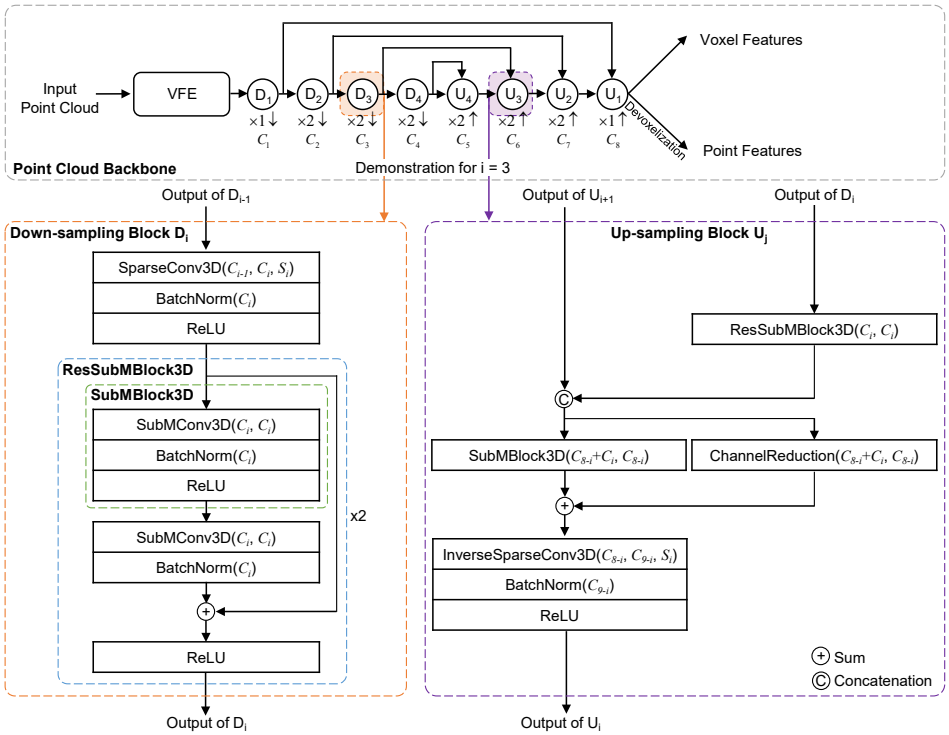

Figure 2. Network architecture details of our point cloud backbone. Two segmentation heads in Fig. 1 can be built on the point cloud backbone output, which serves as the point cloud based single-modal model in Tab 1.

overflow points, which allows each point to be decorated with the feature $f_{\text{img},i}$ of the same dimension.

Fusion. To balance the two feature sources of $f_{\text{pc},i}$ and $f_{\text{img},i}$, we first project them into a C_{int}-dimensional intermediate space using fully connected layers $\mathcal{F}_{\text{pc}}$ and $\mathcal{F}_{\text{img}}$, and then concatenate ("©") them together for subsequent learnable information aggregation using another MLP with output channels of C_{fused} as

$$f_{\text{fused},i} = MLP(\mathcal{F}_{\text{pc}}(f_{\text{pc},i}) © \mathcal{F}_{\text{img}}(f_{\text{img},i})), \quad (1)$$

$$F_{\text{fused}} = [f_{\text{fused},1}, \cdots, f_{\text{fused},N}] \in \mathbb{R}^{C_{\text{fused}} \times N}. \quad (2)$$

The simple yet effective fusion mechanism achieves sufficiently significant performance improvements through unbiased considerations on both modalities. More elaborate fusion mechanisms will be explored in our future work.

2.3. Multi-modal Data Augmentation

The effect of multi-modal fusion depends heavily on the alignment quality between modalities in the association process, so that the multi-modal data augmentation is necessary to be addressed. Our solution does not consider performing the symmetric augmentation transformations on both the point cloud and images.

Both the segmentation output and our multi-modal fusion mechanism are point-centric. The point-wise feature association is applicable by using (u, v), as long as the order of the coordinate pairs of the points in the point cloud and the camera coordinate systems is kept in synchronization. Bare this in mind, we now have the flexibility to decouple multi-modal data augmentation: **i**) We can apply arbitrary transformations to the point cloud while keeping the images and the coordinates (u, v) unchanged. **ii**) We can even apply different and asymmetric transformations to the point cloud and the images if we ensure that the coordinates (u, v) are synchronously transformed with the images.

In such a manner, we are not only able to keep all the point cloud augmentation transformations, but also introduce more image transformations. The permutations between the asymmetric augmentation transformations significantly diversify the multi-modal data augmentation and improve the model robustness.

2.4. Loss Function

As shown in Fig. 1, there are primary and auxiliary segmentation heads built on the final point-wise fused features and the voxel-wise features of the 3D sparse U-Net output, respectively. For each segmentation head, we employ a combination of the cross-entropy loss $\mathcal{L}_{\text{ce}}$ and lovasz-softmax loss $\mathcal{L}_{\text{lovasz}}$ [1]. Let $\alpha \in \{\text{V}, \text{P}\}$ denote the voxel-wise and point-wise, respectively. The total loss term can be formulated as $\mathcal{L} = \sum_{\alpha} (\mathcal{L}_{\text{ce},\alpha} + \mathcal{L}_{\text{lovasz},\alpha})$. Note that the voxel-wise prediction is for auxiliary supervision only, while the point-wise prediction is set as the final segmentation results. To avoid ambiguity in voxel-wise supervision, we ignore the voxels containing points of multiple classes in the auxiliary voxel-wise loss.

P-Frame	Image	P-DA	I-DA	TTA	**mIoU**	Car	Truck	Bus	Other Vehicle	Motorcyclist	Bicyclist	Pedestrian	Sign	Traffic Light	Pole	Cons. Cone	Bicycle	Motorcycle	Building	Vegetation	Tree Trunk	Curb	Road	Lane Marker	Other Ground	Walkable	Sidewalk
1	×	√	×	×	61.92	93.93	56.28	68.86	8.60	0.00	64.08	86.10	68.37	20.65	69.08	64.32	54.30	56.98	94.15	89.93	63.99	67.34	92.16	50.32	41.22	78.05	73.54
1	×	√	×	√	63.70	94.42	57.07	70.88	6.02	0.00	66.76	87.62	70.73	24.77	71.02	67.56	61.96	68.57	94.71	90.66	64.92	69.36	92.45	48.87	40.01	78.82	74.25
1	√	×	×	×	59.88	94.09	52.88	57.96	14.30	0.00	52.85	85.21	67.31	24.48	69.57	62.45	40.64	52.98	94.06	89.50	63.17	66.76	91.78	47.75	39.96	76.96	72.76
1	√	√	×	×	63.24	94.69	57.68	68.66	18.47	0.00	64.15	86.36	69.51	25.55	72.21	65.96	51.28	56.8	94.52	90.17	66.24	67.81	92.11	51.84	45.89	77.95	73.41
1	√	√	√	×	64.18	94.67	58.90	73.37	18.75	0.00	66.96	87.34	70.30	28.09	73.79	67.30	50.99	55.66	94.70	90.37	66.70	68.48	92.53	52.84	46.31	78.65	75.28
1	√	√	√	√	65.45	94.89	59.64	75.26	19.61	0.00	67.08	88.30	71.51	28.61	75.18	69.14	58.76	62.98	95.07	90.87	67.60	69.99	92.68	52.42	45.58	79.09	75.61
9	×	√	×	×	63.83	94.67	55.76	76.94	13.17	0.00	63.41	86.75	67.31	26.48	71.24	67.11	55.99	65.81	94.72	90.32	65.41	68.74	92.39	52.66	44.00	78.28	73.12
9	×	√	×	√	65.56	94.89	56.98	76.67	12.46	0.00	67.77	88.05	69.53	28.99	73.24	69.22	63.89	74.70	95.23	91.02	66.78	70.31	92.74	53.16	43.85	79.08	73.67
9	√	√	√	×	65.66	95.30	60.15	77.32	19.66	0.00	66.74	88.15	71.29	31.62	74.41	68.32	60.90	60.25	95.20	90.80	66.93	69.72	92.65	53.34	46.53	79.18	76.07
9	√	√	√	√	66.91	95.44	60.79	78.57	20.12	0.00	68.52	88.96	72.32	31.77	75.45	69.64	67.93	68.09	95.50	91.25	67.70	71.26	92.82	53.27	46.64	79.52	76.43
1/5/9	√	√	√	√	70.48	95.73	69.03	79.74	37.00	0.00	88.77	92.66	71.82	30.02	80.85	65.97	69.53	76.97	97.15	88.18	72.76	76.40	93.27	49.49	52.61	75.40	87.25

Table 1. 3D segmentation performance of the variants of our method evaluated on the validation set (rows 1 - 10) and testing set (the last raw). For quick validation, a uniformly sampled **1/3** training set with 7,897 samples are used for model training in rows 1 - 10. "P-Frame" and "TTA" denote number of used point cloud frames and test-time augmentation, respectively. The multi-modal data augmentation is decomposed as "P-DA" and "I-DA" for the point cloud and image data augmentations, respectively.

2.5. Extension to Multi-frame Point Clouds

Collapsing the adjacent point cloud frames can improve the sparsity of the point cloud in the current frame [8]. We provide an optional extension to boost performance using multi-frame point clouds as input. We follow Yin *et al.* [19] to align and collapse the multiple historical frames to the current one by the provided ego-vehicle motion information. The relative timestamp is used as additional point-wise initial feature.

3. Experiment

3.1. Implementation Details

Network Architecture. The voxelization step d is set as $(0.1, 0.1, 0.15)$ meters to voxelize point cloud within the range of $[-75.2, +75.2]$, $[-75.2, +75.2]$, $[-4.0, +2.0]$ meters along the X, Y, Z axis. We configure the 3D sparse U-Net with the feature dimensions C_1 - C_8 of 32, 64, 128, 128, 128, 128, 64, 32, respectively. For the image backbone, we adopt the HRNet-w48 [16] with the parameters pre-trained on ImageNet [6], which is publicly available at mmSegmentation toolbox [3].

Training. All our models are trained by the same schedules with the Adam optimizer and one-cycle policy [12] with LR 0.01, division factor 10, momentum ranges from 0.95 to 0.85, weight decay 0.01. A batch of 32 random samples is trained on 16 Tesla V100 GPUs with 12 epochs. Due to the limited memories, we only optimize the last stage of HRNet-w48 based on the frozen first three stages.

We employ the multi-modal data augmentation mentioned in Sec. 2.3. The point cloud augmentation transformations include global scaling with a random scaling factor in $[0.95, 1.05]$, random flipping along the X, Y axis, global rotation around the Z axis with a random angle in $\left[-\frac{\pi}{4}, +\frac{\pi}{4}\right]$, global translation with a random vector $(\Delta x, \Delta y, \Delta z)$ sampled from a Gaussian distribution with mean zero and the standard deviation 0.5. For image, we first resize it as 640×960, then perform image transformations of random horizontal flipping with 0.5 probabilities, scaling with a random factor in [1.0, 1.5], random cropping with a size of 640×960.

Inference. **TTA** is used for performance boost in inference stage by averaging the predictions from 8 augmented input variants. We perform our multi-modal data augmentation again with some slightly modified parameters to generate the random multi-modal input variants. All the point cloud transformations are kept with default parameters. For images, we only perform random horizontal flipping with 0.5 probabilities for simplicity. Besides, We use **model ensemble** of the TTA results of 6 candidate models for the final submission: multi-modal and point cloud based single-modal segmentation models, configured and trained with 1, 5, 9 point cloud frames, respectively.

3.2. Ablation Study and Results

Tab. 1 reports the ablation study and our final results. Rows 1 and 3-5 show that multi-modal data augmentation is the key to effectively train multi-modal models with improved performance, and excluding the data augmentation on either modality degrades the performance a lot. With fairly applying data augmentations, the multi-modal model in row 5 significantly improves the single-modal model in the first row, especially for small objects such as traffic light, pole, cone, bicycle, and so on. TTA stably improves the segmentation robustness of inference. Moreover, the rows 7-10 also representatively demonstrate that the multi-frame point clouds can further improve the segmentation performance. By employing all the components with the full training data, our final results on the testing set are shown in the last raw and can be retrieved on the official leaderboard [5] by the submission entry of "SegNet3D".

References

[1] 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 *CVPR*, pages 4413–4421, 2018. 3

[2] Christopher B. Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In *CVPR*, pages 3075–3084, 2019. 1

[3] MMSegmentation Contributors. MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2022. 4

[4] OpenPCDet Contributors. OpenPCDet: An open-source toolbox for 3D object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2022. 2

[5] Waymo Open Dataset. Waymo Open Dataset leaderboard of 3D semantic segmentation challenge. https://waymo.com/open/challenges/2022/3d-semantic-segmentation/, 2022. 4

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4

[7] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232, 2018. 2

[8] Peiyun Hu, Jason Ziglar, David Held, and Deva Ramanan. What you see is what you get: Exploiting visibility for 3D object detection. In *CVPR*, pages 10998–11006, 2020. 4

[9] Jiale Li, Hang Dai, Ling Shao, and Yong Ding. From voxel to point: IoU-guided 3D object detection for point cloud with voxel-to-point decoder. In *ACM MM*, 2021. 1, 2

[10] Yingwei Li, Adams Wei Yu, Tianjian Meng, Benjamin Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Bo Wu, Yifeng Lu, Denny Zhou, Quoc V. Le, Alan L. Yuille, and Mingxing Tan. Deepfusion: Lidar-camera deep fusion for multi-modal 3D object detection. In *CVPR*, 2022. 1

[11] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network. *IEEE TPAMI*, 43(8):2647–2664, 2021. 2

[12] Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612, 2019. 4

[13] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo Open Dataset. In *CVPR*, pages 2443–2451, 2020. 1

[14] 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 *ECCV*, pages 685–702, 2020. 1

[15] Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3D object detection. In *CVPR*, pages 4603–4611, 2020. 2

[16] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10):3349–3364, 2021. 2, 4

[17] Shaoqing Xu, Dingfu Zhou, Jin Fang, Junbo Yin, Bin Zhou, and Liangjun Zhang. Fusionpainting: Multimodal fusion with adaptive attention for 3D object detection. In *ITSC*, pages 3047–3054, 2021. 2

[18] Yan Yan, Yuxing Mao, and Bo Li. SECOND: sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 2

[19] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3D object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 4

[20] 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 *CVPR*, pages 9939–9948, 2021. 1