3D Semantic Segmentation for Autonomous Driving

Loahit Krishnamurthy [1] **Jinesh Rajasekhar** [1] **Girivaasan Chandrasekaran** [1]
Lalith Athithya Navaneetha Krishnan [1]

Abstract

This project focuses on the task of 3D semantic segmentation, which involves assigning class labels to each point in a point cloud captured by a LiDAR sensor. We propose a rangenet model as the baseline and introduce an additional encoder called Perception-aware Multi-sensor Fusion (PMF) to incorporate RGB camera images into the architecture. An attention-based fusion module is used to fuse image features into the LiDAR stream, allowing the model to learn features from both sources. The proposed model is trained on the SemanticKITTI dataset, which provides 3D point cloud labels for the odometry dataset. The model achieves a mean intersection over union (mIOU) score of 42.1%, outperforming the baseline model in certain classes. The results show that the proposed model performs better in segmenting smaller objects and objects that are far away. Overall, this project demonstrates the importance of semantic segmentation in scene understanding for autonomous driving and proposes a novel architecture that fuses LiDAR and camera data to improve performance.

1. Introduction

The development of autonomous driving technology has generated a growing demand for accurate and reliable perception systems. One critical aspect of these systems is 3D semantic segmentation, a process that assigns class labels to each point in a point cloud captured by LiDAR sensors. This segmentation enables vehicles to better understand their surroundings, ultimately contributing to safer and more effective navigation in complex environments. In this study, we explore the potential of combining LiDAR and RGB camera data to enhance the performance of 3D semantic segmentation for autonomous driving. The integration of these two data sources aims to address the limitations of using LiDAR data alone, particularly with regard to segmenting smaller or distant objects.

1.1. Research contributions

To this end, we propose a novel architecture that fuses LiDAR and camera data, building upon a rangenet model as the baseline and introducing an additional encoder called Perception-aware Multi-sensor Fusion (PMF). Our approach employs an attention-based fusion module to effectively combine image features with the LiDAR stream, allowing the model to learn from both sources of information. We believe that this fusion of LiDAR and camera data has the potential to significantly improve the performance of 3D semantic segmentation in autonomous driving applications, ultimately contributing to more robust and reliable perception systems for self-driving vehicles.

2. Related Work

A considerable body of research has been dedicated to semantic segmentation for autonomous driving applications, focusing primarily on LiDAR and camera data. In this section, we provide an overview of the most relevant studies and techniques in this domain, highlighting their contributions and limitations.

LiDAR-based approaches: Many studies have explored semantic segmentation using LiDAR data alone. For example, Wu et al. (2018)[1] proposed SqueezeSeg, a compact and computationally efficient neural network designed for real-time LiDAR-based semantic segmentation. While such methods offer advantages in terms of processing speed, they often struggle with segmenting smaller or distant objects due to the sparse nature of LiDAR data.

Camera-based approaches: Several works have focused on using RGB camera images for semantic segmentation. Long et al. (2015)[2] introduced the fully convolutional network (FCN) architecture, which revolutionized semantic segmentation using camera images. Although camera-based methods are capable of capturing rich texture and color information, they are sensitive to lighting conditions and occlusions, limiting their applicability in certain sce-

[1]Worcester Polytechnic Institute. Correspondence to: authorname <authorname@wpi.edu>.

CS/DS 541 Deep Learning Final Project, Worcester Polytechnic Institute. Copyright 2023 by the author(s).

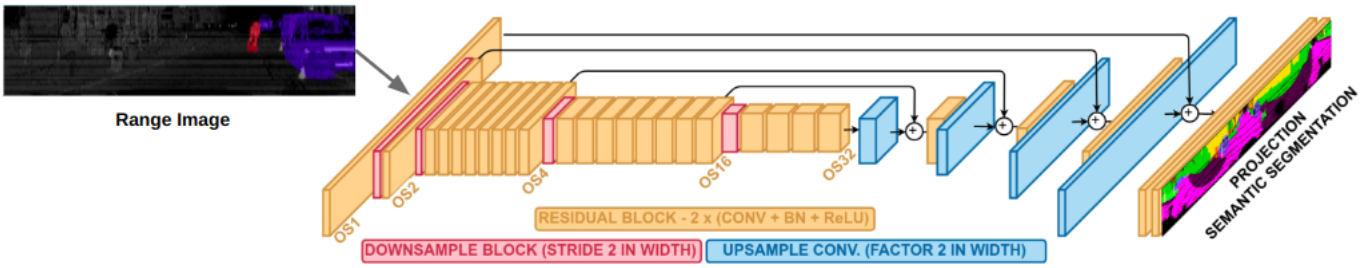

Figure 1. Rangenet architecture

narios.

Multi-sensor fusion: To overcome the limitations of using single-source data, numerous studies have investigated the fusion of LiDAR and camera data for semantic segmentation. Valada et al. (2017)[3] presented the Multi-modal Fusion Network (MFNet), which employs an encoder-decoder architecture to fuse multi-modal features. Another notable work is the MV3D model by Chen et al. (2017), which combines LiDAR and camera data using a 3D region proposal network to enhance object detection and segmentation.

Attention-based fusion: More recently, attention mechanisms have been employed to improve multi-sensor fusion. Xu et al. (2018)[4] proposed the Attention-guided Network (AGNet), which utilizes an attention module to adaptively fuse LiDAR and camera features. This approach enables the model to focus on the most relevant features from each modality, potentially enhancing the performance of semantic segmentation.

While existing methods have made significant strides in improving the accuracy of semantic segmentation for autonomous driving, there is still room for improvement. Our work aims to build upon these previous studies by proposing a novel architecture that effectively fuses LiDAR and camera data, leveraging the strengths of both modalities to enhance the overall performance of 3D semantic segmentation.

The authors in RangeNet [5] propose a novel method for accurate, fast, LiDAR-only semantic segmentation by exploiting range images as an intermediate representation in combination with a Convolutional Neural Network (CNN) exploiting the rotating LiDAR sensor model. Figure 1 shows us this RangeNet architecture. The paper highlights that semantic segmentation assigns a class label to each data point in the input modality, i.e., to a pixel in the case of a camera or to a 3D point obtained by a LiDAR. The proposed method operates on a spherical projection of the input point cloud, which is a 2D image representation, similar to a range image. The paper introduces a post-processing algorithm to deal with problems arising from this intermediate representation such as discretization errors and blurry CNN outputs. The paper demonstrates that the proposed method achieves accurate, fast LiDAR-only semantic segmentation at sensor frame rate. We took this architecture as our baseline model

3. Proposed Method

In our proposed method, we employ RangeNet as our base model and enhance it with a Perception-Aware Multi-Sensor Fusion Encoded Model (PMF) built using ResNet-34 blocks in parallel to the RangeNet encoder. The PMF encoder processes the RGB image, which is then combined with the LiDAR stream through an attention-based fusion module integrated within the intermediate layers of the encoder. Figure 2 illustrates how the fusion module is connected from the Encoder model to the baseline model.

This fusion module takes features from the PMF and RangeNet encoders, concatenates them, and passes them through a convolution layer to obtain feature maps. Figure 3 illustrates the inner workings of the residual-fusion model. Let $\{F_l \in R^{C_l \times H_l \times W_l}\}_{l=1}^{L}$ be a set of image features from the camera stream, where l indicates the layer in which we obtain the features. C_l indicates the number of channels of the l-th layer in the camera stream. H_l and W_l indicate the height and width of the feature maps from the l-th layer, respectively.

Let $\{\tilde{F}_l \in R^{\tilde{C}_l \times H_l \times W_l}\}_{l=1}^{L}$ be a set of image features from the LiDAR stream, where $\tilde{C}_l$ indicates the number of channels of the l-th layer in the LiDAR stream.

The fused features $F_l^{fuse} \in R^{C_l^e \times H_l \times W_l}$ are computed by

$$F_l^{fuse} = f_l([F_l^e; F_l])$$

where $[\cdot;\cdot]$ indicates the concatenation operation. $f_l(\cdot)$ is the convolution operation w.r.t. the l-th fusion module.

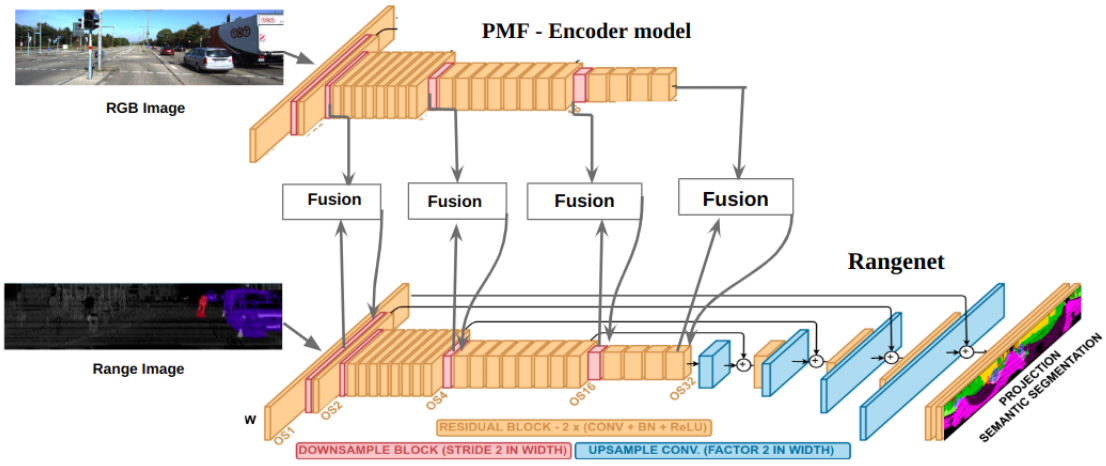

Figure 2. Our Architecture - PMF Encoder model added to Rangenet architecture

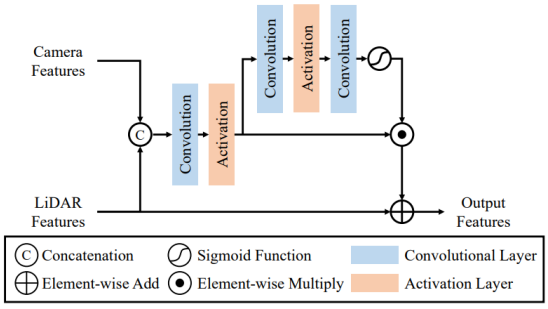

Figure 3. Illustration of residual-fusion model

Attention rates are calculated for each feature map, normalized to ensure their sum equals one, and subsequently used to assign weights to each feature map. This attention mechanism enables the range image branch to identify the most important camera features. Finally, each feature map is multiplied by its corresponding attention weight, and the resulting fused feature map is fed into the main network for further processing.

The output features $F_l^{out} \in R^{\tilde{C}_l \times H_l \times W_l}$ of the fusion module are computed by

$$F_l^{out} = \tilde{F}_l + \sigma(g_l(F_l^{fuse})) \odot F_l^{fuse}$$

where $\sigma(x) = 1/(1+e^{-x})$ indicates the sigmoid function. and $g_l(\cdot)$ indicates the convolution operation in the attention module w.r.t. the l-th fusion module. $\odot$ indicates the elementwise multiplication operation.

This novel architecture allows the model to effectively learn features from both RGB images and range images, improving overall performance.

The last layer during inference is a softmax function over the unbounded logits of the form $\hat{y}_c = \frac{e^{logit_c}}{\sum_c e^{logit_c}}$. This gives a probability distribution per pixel in the range image, where $logit_c$ is the unbounded output in the slice corresponding to class c. During training, this network is optimized end to end using stochastic gradient descent and a weighted cross-entropy loss L:

$$L = -\sum_{c=1}^{C} w_c y_c \log \hat{y}_c \qquad (2)$$

where $w_c = \frac{1}{\log(f_c+\epsilon)}$ penalizes the class c according to the inverse of its frequency f_c. This handles imbalanced data, as is the case for most datasets in semantic segmentation, e.g. the class "road" represents a significantly larger number of points in the dataset than the class "pedestrian".

4. Experiment

Dataset: The model is trained using the SemanticKITTI dataset, which provides 3D point cloud labels. The dataset consists of 43,000 scans, of which over 21,000 from folder name sequences from 0 to 10 are used for training, sequence 8 is allotted for validation, and the remaining 11 to 21 are used as test data.

Hyperparameter selection: The model is trained with a batch size of 2 and a learning rate of 0.001 for 40 epochs with a decay of 0.99 every epoch, using the SGD with momentum as the optimization, and cross-entropy loss for each pixel. The same hyperparameters were used in validation set.

Metric: The model's performance is evaluated using a mean intersection over-union (IoU) metric, mIoU, over all the object classes in the data set which is given by

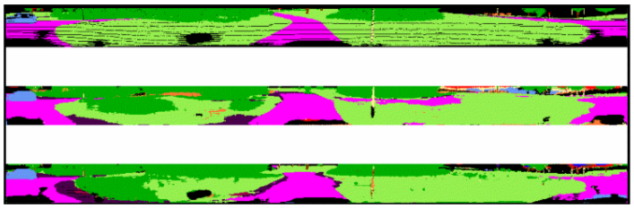

Figure 4. Segmented Image

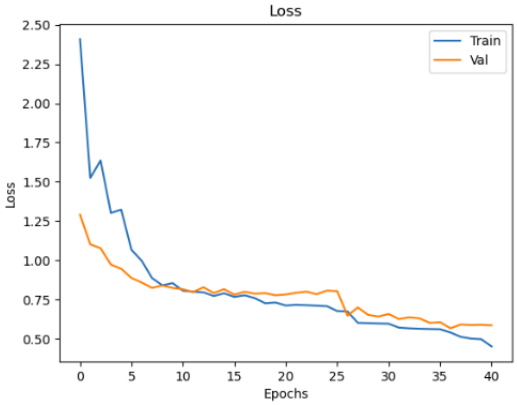

Figure 5. Loss vs Epochs

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{(TP_i + FP_i + FN_i)}$$

where TP_i, FP_i, and FN_i correspond to the number of true positive, false positive, and false negative predictions for class i and N is the number of classes.

5. Results

The results in Table 1 show that the proposed model performs better in segmenting objects that are smaller and far away from LiDAR, such as traffic signs, people, poles, and motorcyclists. The model outperforms the baseline model in terms of the IOU score, achieving a higher score in the validation set. But as excepted, adding an additional branch resulted in an increase in inference time compared to the baseline model. Figure 5 and 6 exhibits training and validation loss and the corresponding mean IoU scores with respect to number of epochs.

From the results in Figure 4, the top image represents the ground truth the middle image represents the semantic image from our model and the bottom image represents the semantic image from the baseline model. Thus it is observed that our model contains more clearly segmented objects.

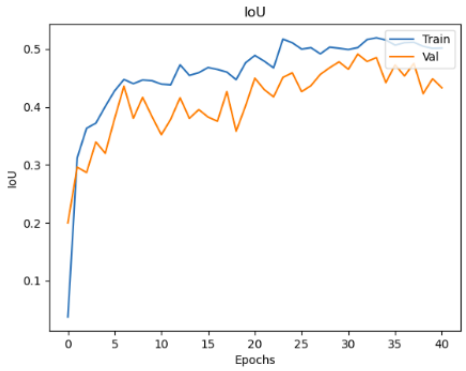

Figure 6. mIoU vs Epochs

6. Discussion

In recent years, the development of autonomous driving systems has gained significant attention from both industry and academia. One of the essential tasks in achieving fully autonomous driving is semantic segmentation of the surrounding environment. We have utilized the RangeNet model as the baseline and introduced an additional encoder called PMF, built using ResNet blocks, to fuse the RGB image with the LiDAR data. The introduction of an attention-based Fusion module in the intermediate layers of the encoder has allowed the model to learn features from both the RGB and LiDAR data effectively. Training the model on the Semantic KITTI dataset has provided reliable results for 3D semantic segmentation, which can aid in scene understanding for autonomous driving applications. This project's success highlights the potential of deep learning techniques to improve autonomous driving technology and further motivates research in this area.

7. Conclusions and Future Work

LiDAR-Camera Fusion improves the segmentation accuracy of smaller objects like Bicycles, Poles, and Traffic signs. Fusion also improves the performance of the model in segmenting far-away objects. RGB image fusion gives rich features even at farther distances. Adding extra parameters in the form of image encoder layers leads to reduced inference speed due to increased model size. More efficient feature Fusion methods like using transformers to do the fusion instead of simple attention. Fusion of Segmented RGB images in the encoder can also improve performance. Improve the existing model by incorporating a decoder and fusion architecture within the image branch's decoder, followed by testing and evaluating its performance. Finding the right tradeoff between inference speed and segmentation accuracy.

Model	Car	Traffic sign	Person	Motorcyclist	Sidewalk	Truck	Pole	Mean IoU	Inference Time (ms)
Baseline	**85.4**	50.0	31.8	4.0	**70.0**	**18.6**	36.0	**47.4**	**20**
Ours	82.7	**52.1**	**33.1**	**5.2**	65.8	16.6	**41.3**	42.4	35

Table 1. Result metrics

References

[1] Wu, Bichen, et al. *"Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud."* 2018 IEEE international conference on robotics and automation (ICRA). IEEE, 2018.

[2] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. *"Fully convolutional networks for semantic segmentation."* Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[3] Valada, Abhinav, Rohit Mohan, and Wolfram Burgard. *"Self-supervised model adaptation for multimodal semantic segmentation."* International Journal of Computer Vision 128.5 (2020): 1239-1285.

[4] Chen, Xiaozhi, et al. *"Multi-view 3d object detection network for autonomous driving."* Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017.

[5] A. Milioto, I. Vizzo, J. Behley and C. Stachniss, *"RangeNet ++: Fast and Accurate LiDAR Semantic Segmentation,"* 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 2019, pp. 4213-4220, doi: 10.1109/IROS40897.2019.8967762.