Semantic Segmentation of 3D LiDAR Data at Dynamic Urban Scenes

Biao Gao*, Jilin Mei*, Donghao Xu*, Xijun Zhao†, Wen Yao†, Huijing Zhao*
*Peking University, Beijing, China
†China North Vehicle Research Institute, Beijing, China

***Abstract*— This work studies semantic segmentation of 3D LiDAR data at dynamic urban scenes. LiDAR data plays an important role of perception in autonomous driving system. However, most semantic segmentation methods and datasets are designed for camera data nowadays. In this work, we propose a method which can generate semantic segmentation of LiDAR data and we evaluate its performance on a new 3D point cloud dataset collected in dynamic urban scenes by our driving platform. The experiments show that our method can recognize more kinds of labels and achieve an impressive result in dynamic urban scenes.**

I. Introduction

empty

II. Related works

empty

III. Methodology

A. Data Preprocessing

The point cloud data for LiDAR is sparse and unorganized, so it is time-consuming to find neighboring relations between different points. In order to process these unorganized point cloud data with deep convolutional neural network, we convert the point cloud data into 2D range image by cylindrical projection. After that, it will be easier to implement deep convolution neural network on the LiDAR data.

After cylindrical projection, the point cloud data will be encoded as a dense matrix with shape of $[H, W, C]$. H means the number of lines for the specific LiDAR sensor (such as $H = 32$ for Velodyne HDL-32E). W equals to the number of points within each LiDAR scan line. C is the channels' number in the range image. Here, C is set to 3, which represents $[Range, Intensity, Height]$ three channels.

For each point $p_k = \langle x_k, y_k, z_k \rangle$ from the raw point cloud set P, the value of $Range$ in range image R is defined as r_k:

$$r_k = \sqrt{{x_k}^2 + {y_k}^2 + {z_k}^2}, r_k \subset [0, 255] \tag{1}$$

Similarly, the values of $Intensity$ and $Height$ are normalized into $[0, 255]$. These channels are very important properties of LiDAR data, which are enough to describe various objects in dynamic urban scenes.

B. Problem Definition

Let X denotes the range image extracted by cylindrical projection of 3D point cloud data S. In this kind of projection, there is a one-to-one correspondence between a 3D point in one frame point cloud data and a pixel in the range image X. As a result, the semantic segmentation task of 3D point cloud data is equal with giving each pixel x in the range image X a label y. The problem of this work is formulated as learning a semantic segmentation model f_θ which maps each pixel x to a label $y \in \{1, ..., K\}$, and subsequently associate y to the 3D points of S.

$$f_\theta : x \to y \in \{1, ..., K\} \tag{2}$$

The data samples are in the form of range images X. Given a set of supervised data samples $X_l = \{x_i, y_i\}$, where $\{x_i\}$ traverses each pixel of X and $\{y_i\}$ are labels for $\{x_i\}$, annotated manually by human annotators. In order to learning a semantic segmentation model f_θ, we need to find the best parameter set θ^* that minimize a loss function L as below.

$$\theta^* = \arg\max_\theta L(X_l; \theta) \tag{3}$$

C. Network Architecture and Loss Function

We use a FCN (Fully Convolutional Network) architecture for this semantic segmentation task. Compared with common deep convolutional networks, it removes last fully connected layers, and replaces them with the in-network up-sampled or de-convolutional predictions of convolutional layers as predicted feature maps. During training procedure, it generally computes cross-entropy like losses in pixel-wise, between the predicted labels and ground truths.

Our network is trained via end-to-end guided by the designed loss function. Because the number of pixels are imbalanced between different classes and a lot of invalid or unknown-class pixels, the following multi-class weighted cross entropy loss function is designed to regular the weights between imbalanced classes.

For labeled data X_l, we implement some changes on the widely-used definition of cross entropy, and define loss function L_l as below:

$$\Gamma_{i,j} = \begin{cases} \overrightarrow{\varphi_k}, & if \quad [y_{i,j} \neq k \quad and \quad y_{i,j} \neq 0] \\ \overrightarrow{0}, & otherwise \end{cases}$$

$$L_l(X_l, Y_l; \theta) = -\frac{1}{H * W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \sum_{k=0}^{K} \Gamma_{i,j} \omega_k ln(P_\theta^k(x_{i,j})) \tag{4}$$

Where $\Gamma_{i,j}$ is a one-hot vector φ_k of label k, if $y_{i,j} \neq k$ and $y_{i,j} \neq 0$. Label 0 means invalid or unknown pixels, including many fine fragments belong to background or hard to be annotated, so we don't want to evaluate these pixels if they are predicted as non-zero labels. ω_k here is used to balance the sample numbers between different labels and $P_\theta^k(x_{i,j})$ is the probability that pixel $x_{i,j}$ be assigned a label k by our semantic segmentation model with the set of parameters θ.

IV. Experiment

A. Data Set

The performance of the proposed method is evaluated on a dynamic campus data set collected by an instrumented vehicle, which has a GPS/IMU suite and a Velodyne-HDL32, as shown in Fig. 1. The total route contains 1375 LiDAR frames. 880 frames for training, 220 frames for validation and 275 frames for testing.

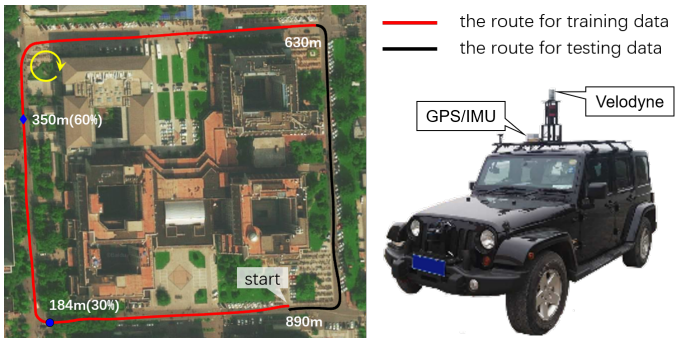


Fig. 1: The routes of data collection and the platform configuration.

High quality pixel-level annotation is necessary for network training. Instead of working on the raw point cloud, human annotators work on the range image where object regions are associated with the ones in adjacent frames. Annotators only need to assign the category of some region in one frame, then a series of associated regions are marked with the same label. Although sometimes the data association brings errors, it largely reduce the annotation time.

The categories distribution in this dataset is shown in TABLE . I. Obviously, the data distribution is imbalanced between categories. So, we apply each label an unique weight based on data distribution to reduce the influence of data imbalance.

B. Setup

Our method is implemented with a FCN (Fully Convolutional Network). The range image size is 1080x32, which width is down-sampled for efficiency. A small batch size for training sets will be better. The network is implemented with TensorFlow in the environment with NVIDIA TITAN X GPU. We use the AdamOptimizer with 1e-5 learning rate.

It's important to aware that our data frames are captured sequentially. In order to avoid data correlation between adjacent frames, shuffling them before training process is necessary.

C. Results Evaluation

We use mPA (mean pixel accuracy) for quantitative evaluation of semantic segmentation performance. Specially, when calculating the mean pixel accuracy, unknown or invalid pixels with label 0 will not be involved in. The mPA is computed as the following formula, while $\dot{Y}_k$ is the predicted pixel set with label k and Y_k is the ground truth pixel set.

$$mPA = \frac{1}{K} * \sum_{k=1}^{K} \frac{\left|\dot{Y}_k \cap Y_k\right|}{\left|\dot{Y}_k\right|} \tag{5}$$

As shown in Table. II , we list the semantic segmentation performance (mPA) of each label, and use the result trained with original cross-entropy as baseline for comparison. Obviously, most categories achieve higher mPA performance after use our proposed method with weighted loss function. Especially, the *pedestrian* label gets 37% accuracy increase, and not only the category with fewer samples gets higher accuracy, but also the category with large amounts of samples gets higher accuracy, such as the *vegetation*. However, some categories like *sign/pole* still don't have ideal pixel accuracy by our method. This is most likely for too few pixel samples of these categories, so the deep learning model can't fit the data well.

The qualitative semantic segmentation results are shown in Fig.2. Dynamic urban scene is very vivid and contains abundant categories objects and details, which is different from simple environment on highway. As we can see in Fig.2(a), the weighted loss leads to better performance of labels with less samples, such as the traffic sign in Box A. Similarly, Box B,C,D in Fig.2(b) also give some improved examples.

Fig.3 shows the confusion matrix of baseline and weighted-loss model. It's easy to find that our model achieves a great improvement of most categories. However, some confusion between categories still exist, such as many *sign/pole* samples are predicted as *vegetation* and some *cyclist* samples are predicted as *pedestrian*. The second situation is mainly caused by the cyclists close to our vehicle, so only the upper part of the body can be seen.

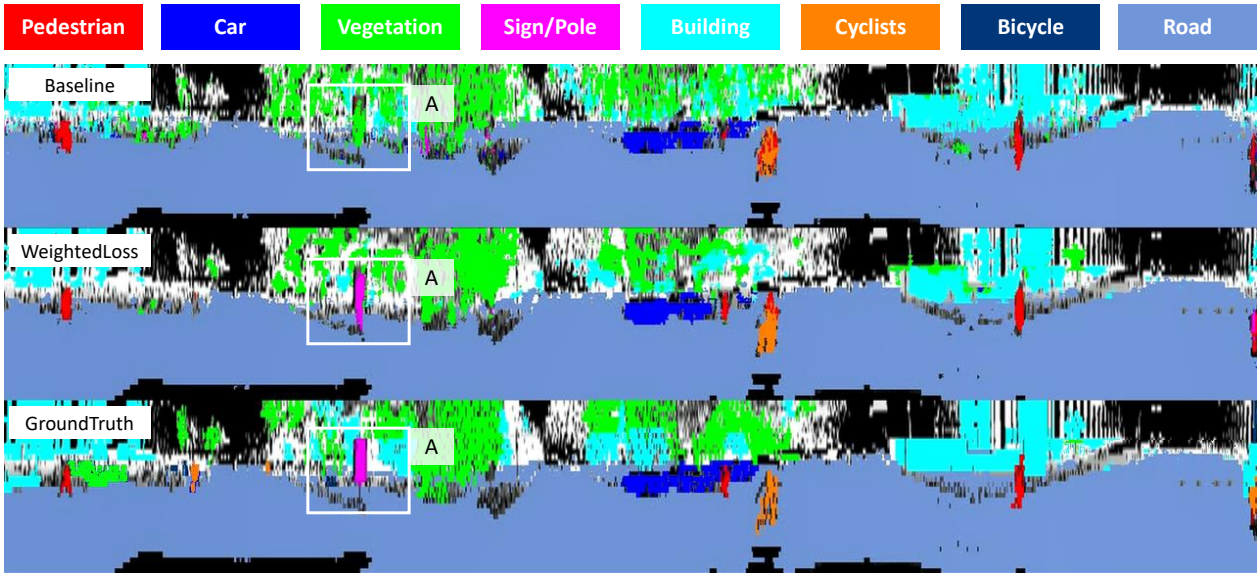
By the way, there is an interesting explanation for low pixel accuracy of label *sign/pole*. As shown in Fig.2(b), we can find that some trees trunks are labeled by *sign/pole*. This is mainly because of the high reflective moth-proofing paint on the trunks, which has high value of intensity and shape as a pole, just

TABLE I: Categories Distribution in Dataset

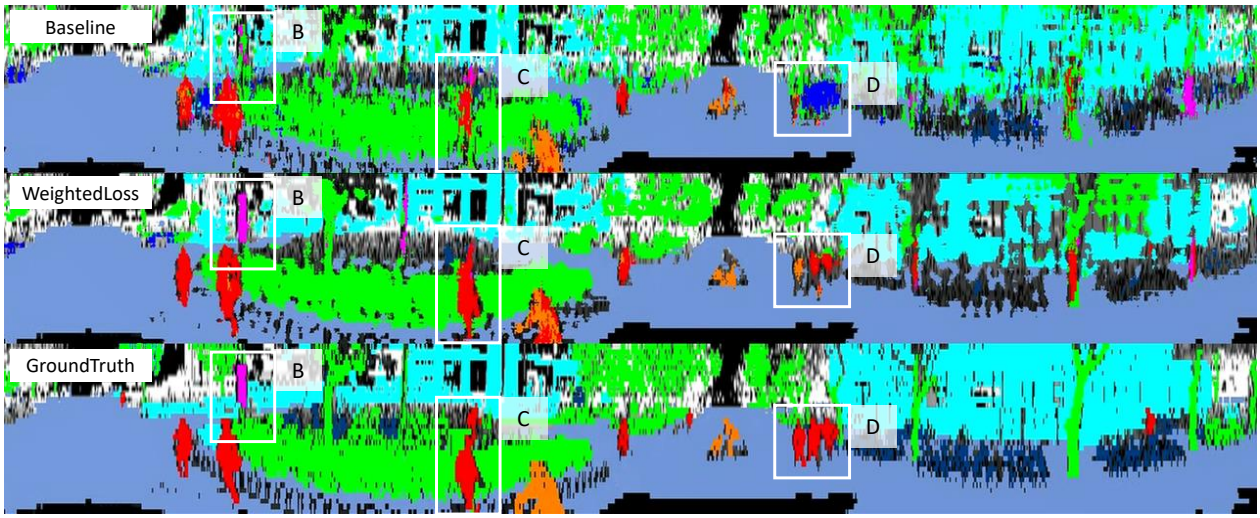
-	Pedestrian	Car	Vegetation	Sign/Pole	Building	Cyclist	Bicycle	Road
Pixels	458,673	257,239	4,661,880	80,384	6,579,360	274,991	1,356,518	16,856,614

TABLE II: semantic segmentation performance (mPA)

-	Pedestrian	Car	Vegetation	Sign/Pole	Building	Cyclist	Bicycle	Road
Baseline Model	0.41	0.71	0.57	0.36	**0.66**	0.49	0.21	**0.99**
Weighted-Loss Model	**0.78**	**0.93**	**0.74**	**0.40**	**0.66**	**0.56**	**0.49**	**0.99**

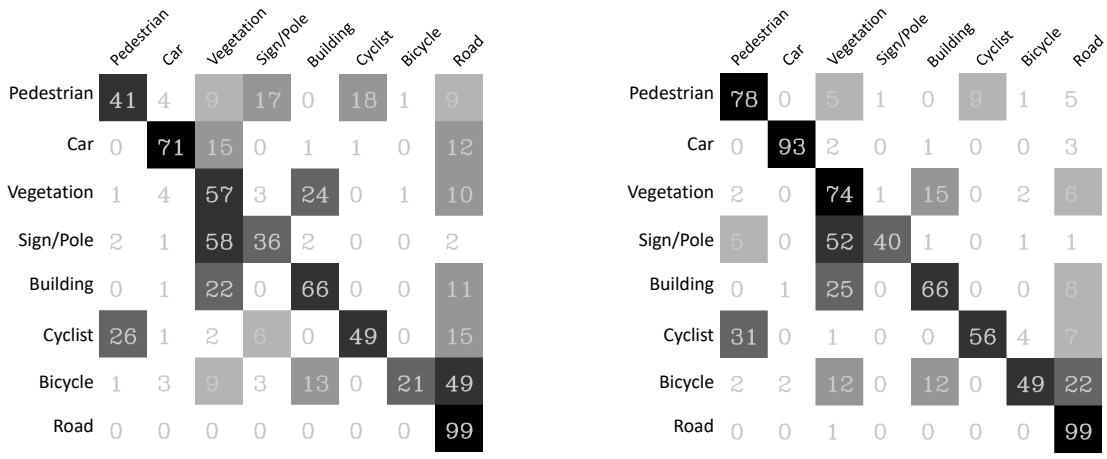


(a) Weighted loss leads to better performance of labels with less samples, such as the traffic sign in Box A.



(b) Weighted loss version segmentation model can achieve more balanced result than the baseline. For example, Box B shows the improvement of a traffic sign's label result, and Box C,D include the results improvement of pedestrain.

Fig. 2: Visualization result of semantic segmentation.



(a) Confusion matrix of baseline model

(b) Confustion matrix of weighted loss model

Fig. 3: The confusion matrix on testing dataset, values in matrices are shown in percentage. Rows means ground truth labels and columns means predicted labels.

like the traffic sign. In order to deal with this situation, maybe object-level information needs to be imported.

V. Conclusion

This paper proposes a semantic segmentation method of 3D LiDAR data at dynamic urban scenes. We convert the LiDAR data into the form of range images and use a fully convolutional network to predict the semantic segmentation result. We use our new dataset for evaluation, and find the weighted loss distinctly improves the pixel accuracy performance.

There are still much work need to be done in the future. In order to further improve the semantic segmentation performance, object-level information, spacial and temporal constraints between samples will be addressed on.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.