SalLiDAR: Saliency Knowledge Transfer Learning for 3D Point Cloud Understanding

Guanqun Ding*[1,2]
guanqun.ding@aist.go.jp
Nevrez İmamoğlu*[2]
nevrez.imamoglu@aist.go.jp
Ali Caglayan[2]
ali.caglayan@aist.go.jp
Masahiro Murakawa[1,2]
m.murakawa@aist.go.jp
Ryosuke Nakamura[2]
r.nakamura@aist.go.jp

[1] Graduate School of Science and Technology,
University of Tsukuba,
Ibaraki, Japan

[2] National Institute of Advanced Industrial Science and Technology,
Tokyo, Japan

Abstract

Saliency prediction has achieved significant progress in color images owing to deep neural networks trained on annotated human eye-fixation data or ground truth saliency maps. Unlike in image/video domain, only a few works have addressed saliency information to further guide 3D point cloud understanding due to the lack of annotated training data. Moreover, it is certainly difficult and not feasible for subjects to annotate eye-fixation or density saliency map groundtruth for point clouds due to the irregular, unordered, and sparse characteristics of 3D point cloud data. To alleviate this issue, we present a universal framework to transfer saliency distribution knowledge from color images to point clouds. We first apply pre-trained RGB saliency models to predict saliency maps for images. We then assign saliency value of each point on 3D point cloud registered to respective 2D multi-view color images by using the RGB saliency prediction. Based on that, we construct a pseudo-saliency dataset (i.e. FordSaliency) that presents 2D to 3D transferred saliency information for point clouds. Furthermore, we adopt existing point cloud-based models to learn saliency distribution from pseudo-saliency labels. Experimental results on our FordSaliency dataset verify that the point cloud-based models can learn saliency distributions from point cloud pseudo-labels. Finally, we demonstrate an application of point cloud saliency predictions on 3D semantic segmentation. Specifically, we propose an attention guided learning model by combining learned saliency knowledge and semantic features for large-scale point cloud segmentation. Extensive experiments of the proposed attention guided learning model on SemanticKITTI [1] dataset show that the learned saliency knowledge effectively improves the performance of the 3D semantic segmentation task.

*Corresponding authors.

1 Introduction

3D point cloud understanding has gained increasing attention with the wider application of robotics technologies such as autonomous vehicles and augmented/virtual/mixed reality. Concretely, large-scale data-based applications, such as 3D object detection [42], LiDAR semantic segmentation [9, 44], and odometry estimation [31] empower such robotics technologies. On the other hand, utilizing saliency information in various 2D computer vision tasks including image translation [12], object tracking [18, 43], key-point selection [28, 30], and person re-identification [14, 25, 40] has moved forward the-state-of-the-art results thanks to its capacity to cover pre-dominant information in a scene. However, the unstructured, unordered, and density-varied properties of point clouds make it difficult for conventional point-cloud-based methods to effectively and rapidly process informative visual features in large-scale scenes. Due to its importance in real-time autonomous vehicles, several works [5, 26, 41] have attempted to apply saliency detection algorithms to point cloud data-based tasks. As these works also verify that performance of 3D point cloud understanding tasks can be improved further with efficient saliency knowledge integration.

Although several attempts have been made to find effective solutions for saliency detection on point clouds [5, 26, 28, 41], most challenges are yet to be explored further. First, previous saliency methods such as [5, 41] have operated on relative small-scale point cloud data of indoor scenes or dense mesh data of 3D objects, where scenes are less complicated with only a few background points. These methods [5, 26] have not been developed for large-scale outdoor driving scenes, *e.g.* a dataset of the similar scale as the SemanticKITTI [1] dataset. Second, due to the lack of human-annotated training datasets, it is unlikely to employ supervised learning scheme or convolutional networks with powerful representation capabilities for point cloud saliency detection. Existing saliency approaches on point clouds [26, 28] mainly utilize traditional computation models to calculate saliency/importance value of each point. Third, there are numerous different rotation angles and scaling/zoom sizes for the same point cloud scan, and the saliency distribution could be varied when the viewing angle or observing scale is changed. Thus, it is not feasible for subjects/human to annotate eye-fixation ground-truth for disordered point clouds. In a word, the mentioned typical issues make saliency prediction on point cloud challenging. Therefore, it is highly desired to develop a practical pipeline based on deep learning for saliency prediction on large-scale point clouds.

Currently, many previous studies [22, 24] have suggested that it is possible to establish correspondences between a 3D laser scanner and an optical camera system. By means of the camera parameters, these LiDAR-based methods [22, 24] are able to take full advantage of visual information from mature 2D computer vision models. In this work, we present a common framework (see Figure 1) to transfer saliency distribution knowledge from color images to point clouds. And then, we explore the use case of this learned saliency information by integrating it to point cloud segmentation task (see Figure 2).

In brief, the key contributions of this work can be summarized as follows:

- We design a universal framework to transfer saliency distribution knowledge for point clouds. Based on this pipeline, we build a point cloud saliency dataset (FordSaliency) based on the FordCampus dataset [21] for training LiDAR-based saliency models.
- We adopt existing LiDAR-based models to learn saliency distribution from pseudo-labels, and we compare the performance of these point cloud saliency models on our FordSaliency dataset as an initial evaluation benchmark.

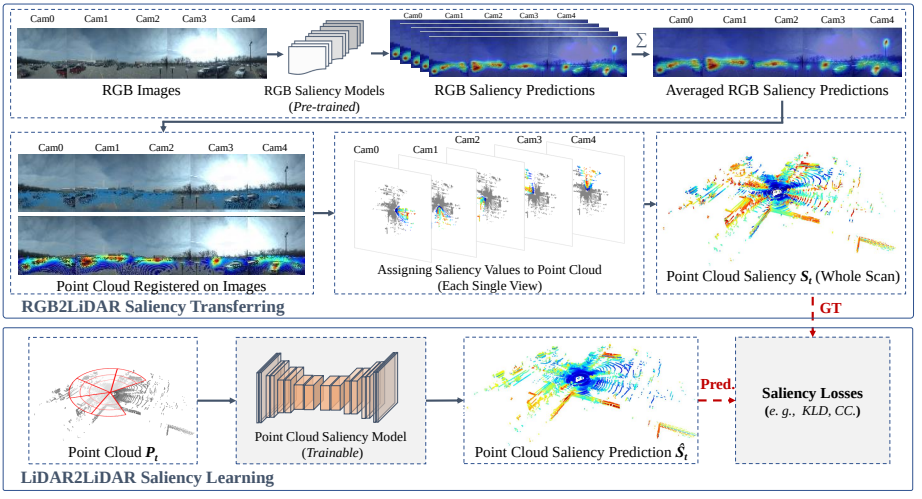

Figure 1: Framework overview of LiDAR saliency annotation and model training.

- We take advantage of these learned point cloud saliency prediction networks in semantic segmentation task to enhance accuracy. To do this, we propose attention guided point cloud semantic segmentation learning with a two-stream network. First stream is the pre-trained saliency prediction network to guide the segmentation task. And second stream is semantic segmentation module fine-tuned with the saliency information.

Extensive experimental results on our FordSaliency dataset show that the LiDAR-based methods could learn saliency knowledge distribution from pseudo-saliency annotations on point clouds. Furthermore, the experiments on SemanticKITTI [1] dataset suggest that our two-stream segmentation model with saliency distribution knowledge significantly improves the performance of semantic segmentation on point cloud of large-scale driving scenes.

2 Related Works

Saliency detection: Saliency detection aims to find the most eye-attracting locations in a visual scene, which can be traced back to the pioneering work of Itti *et al.* [11]. With rapidly emerging advances and applications of deep learning techniques, saliency detection on color images/videos [6, 17] has made great progress in recent years. There are also several works [5, 26, 28, 30, 41] for saliency computation on point clouds. However, saliency methods focusing on 3D meshes or indoor scenes are limited in their ability to process large-scale 3D point clouds such as 3D driving data. Also, saliency models extracting handcrafted descriptors may ignore informative representations for point clouds with varying density and complex background in outdoor scenarios.

LiDAR semantic segmentation: LiDAR semantic segmentation [9, 10, 19, 23, 27, 36, 44] is a crucial 3D computer vision task for autonomous driving, which aims to predict the semantic class of each point on a LiDAR scan. With different representation of the input point cloud, 3D semantic segmentation can be categorized into point set-based model [10, 23], range

image-based model [19], and voxel-based [9, 44] model. As a pioneering point set-based method, PointNet [23] uses Multiple Layer Perceptrons (MLPs) to learn point-wise features for classification and segmentation. RandLA-Net [10] presents randomly sampling the input point cloud, and employs a local feature aggregation module to compensate information loss introduced by the random sampling. It first predicts semantics on a subset of point cloud, then projects the predictions to the full LiDAR scan [10]. RangeNet++ [19] first projects 3D point clouds onto 2D range images, then utilizes 2D convolutional networks to extract features for semantic segmentation. Considering the range property of LiDAR point cloud, Cylinder3D [44] proposes a solution to leverage cylinder partition for 3D semantic segmentation. It also brings an asymmetrical model to encoder-decoder voxel-based features by 3D sparse convolutional networks.

3 Proposed Framework

3.1 Problem Formulation

Given an input point cloud $P=\{p_i|\ i=1,...,N,\ p_i \in \mathbb{R}^d\}$ with a set of disordered points, where N represents the point number of LiDAR frame and each point p_i could contain d dimensional features, such as point coordinates (x,y,z), colors (r,g,b), reflectivity, and normal feature. The objective of saliency detection model on point cloud is to predict the saliency score map $S=\{s_i|\ i=1,...,N,\ s_i \in [0,1]\}$, where s_i denotes the saliency score of point p_i. After normalizing the saliency prediction, the closer the saliency score s_i to 1, the more attentive the point p_i. In 3D semantic segmentation task, its goal is to predict the semantic class map $C=\{c_i|\ i=1,...,N,\ c_i \in \mathbb{R}\}$, where c_i indicates the semantic category of point p_i.

3.2 Saliency Knowledge Transfer

The unstructured and density-varying properties of point cloud make it difficult to annotate eye-fixation ground truth by human subjects. Thus, we propose a common pipeline to transfer saliency knowledge from color images to point clouds, as shown in Figure 1. Inspired by the works on transfer learning [13, 37] and knowledge distillation [4, 8, 32], our main idea of this work is based on a hypothesis that, a pre-trained saliency model on RGB images or point cloud has learned abundant saliency distribution knowledge from training data. And it could be a good transmitter to transfer the learned knowledge if the generalization capability of the saliency model is good enough. To verify our assumption, we transfer the saliency knowledge from RGB images to point clouds in this work.

In Figure 1, we show the procedure of point cloud saliency projection. We first leverage existing pre-trained RGB saliency models as saliency pseudo-annotators to *label* the saliency distributions of color images from LiDAR-based FordCampus [21] dataset. In order to avoid bias introduced by a single saliency method, we average the predictions of M pre-trained models to generate final saliency distribution of multi-view color images. The averaged annotation can be represented by $S_{avg} = \sigma(\frac{1}{M}\sum_i^M S_i)$, where S_i is the saliency prediction of i-th RGB deep learning model; σ denotes normalization function. We regard these predicted saliency distributions as pseudo-saliency ground-truth annotations of point clouds. Next, we assign the values on pseudo-saliency maps to point cloud. By rigid-body transformations and sensor/camera parameters provided in FordCampus [21] dataset, we can easily project any points from one coordinate frame to the other, more details about the transformation can

be referred to the study of [21]. After we obtain the saliency distribution of images from different camera views, we assign saliency values to the corresponding points registered on multi-view images. Finally, we assign the saliency values of points from the camera image coordinate system to the corresponding point cloud domain. In this way, we build a point cloud pseudo-saliency dataset (FordSaliency) based on FordCampus [21] dataset. Although the annotations of FordSaliency are pseudo-labels, we believe that these pseudo-labels could contain saliency distribution knowledge, which can be utilized to train LiDAR-based models for saliency detection on point clouds.

3.3 Learning Point Cloud Saliency

In order to learn point cloud saliency representations, we adopt existing LiDAR-based semantic segmentation models [10, 23, 44] as backbones of the feature extractor. As shown in Figure 1, given a 3D LiDAR point cloud with coordinates and corresponding point-wise features, we first feed it into the feature extractor to obtain the representations of each point. Next, these learned features are passed by a saliency prediction layer to output the saliency score map of the input point cloud. We considered two types of model to learn saliency distribution on point clouds: i) classification based saliency prediction and ii) commonly used saliency regression.

In saliency classification modelling, our key motivation is that saliency score and probability can be split into different levels for high/low attentive points by quantization of the ground truth saliency values. Considering the fact that saliency map of color image is a 8-bit gray image, we experimentally set the number of class to K based on the specific bin size/width. We convert 8-bit saliency label s to class k by the following formula:

$$s_{cls} = k, \quad \text{if } s \in [k \times bin, (k+1) \times bin]. \tag{1}$$

where $bin = 2^8/K$ and $k = 0, ..., K-1$. To obtain the probabilities of each class, we apply softmax activation function to the embedding features F_{cls} from the last prediction layer of saliency classifier, which can be formulated as follows:

$$\mathbb{P} = softmax(F_{cls}) \tag{2}$$

where $\mathbb{P}$ includes K probabilities of saliency classes for each point p_i. Then we calculate the predicted saliency value $\hat{s}_i$ for each point p_i by using the probabilities:

$$\hat{s}_i = \sum_{j=1}^{K} \mathbb{P}(i,j) * j \tag{3}$$

where i and j denote the index of point and class, respectively. Thus, the saliency prediction map of input point cloud can be represented as $\hat{S}_{cls} = \{\hat{s}_i | i = 1, ..., N, \hat{s}_i \in [0,1]\}$. Here, we leverage the cross-entropy loss for optimizing the saliency classifier on point clouds.

In regression based saliency learning, network is modeled to map input data to output saliency values learned from the distribution of ground truth saliency maps. Like the saliency detection models [4] on color images, the predicted saliency score map $\hat{S}_{reg}$ of point cloud is scaled to the range of [0, 1]. Following the previous RGB saliency studies [4, 6], we adopt saliency distribution-based loss in Eq. 4 to optimize the model parameters.

$$\mathcal{L}_{reg} = \gamma \mathcal{L}_{KLD}(\hat{S}_{reg}, S) + \delta \mathcal{L}_{CC}(\hat{S}_{reg}, S) + \eta \mathcal{L}_{MSE}(\hat{S}_{reg}, S) \tag{4}$$

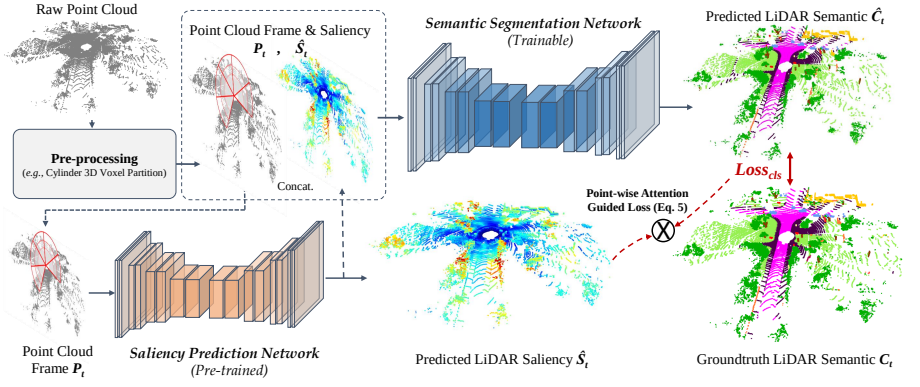

Figure 2: Framework of proposed two-stream semantic segmentation model. The saliency prediction network is pre-trained on our FordSaliency dataset.

where $\mathcal{L}_{KLD}, \mathcal{L}_{CC}, \mathcal{L}_{MSE}$ denote Kullback-Leibler Divergence (KLD) loss, Correlation Coefficient (CC) loss, and Mean Square Error (MSE) loss, respectively; γ, δ, η are the weighting constants of the three losses, respectively; S indicates the pseudo-saliency-label of point clouds. Once the model training is completed, we believe these point cloud saliency models have the saliency knowledge transferred from color images. We then utilize them to improve the performance of 3D semantic segmentation task.

3.4 Two-Stream Segmentation Model

As depicted in Figure 2, we develop a two-stream semantic segmentation model on point cloud by combining features from saliency module and semantic module. We feed an input point cloud into the saliency branch to predict saliency distribution of the whole scene. Meanwhile, the point cloud is also fed into the semantic branch to extract point features and output the predictions of the semantic class. To validate the effectiveness of the learned point cloud saliency distribution knowledge, we initialize and freeze the parameters of the saliency branch with the weights pre-trained on FordSaliency dataset. We consider three different integration ways to combine point saliency embedding/distribution and point semantic information:

1) SalLiDAR-I: Attention guided loss for semantic segmentation. Since the predicted saliency distribution represents the attention level of each point, we can apply it to guide the parameter optimization of the semantic segmentation model. In Figure 2, SalLiDAR-I is the model without saliency concatenation module. The key motivation of this idea is guiding the semantic module by using a weighted attentive loss based on saliency distribution that pays more attention to class prediction of certain points. We utilize normalized saliency distribution $\hat{S}$ of the whole point cloud to weight semantic loss by:

$$\hat{\mathcal{L}}^{sem} = \frac{1}{N}\sum_{i=1}^{N} l_i^{sem} * \exp(\hat{s}_i) \tag{5}$$

where i is the index of point; $\hat{\mathcal{L}}^{sem}$ is saliency weighted loss for segmentation; l_i^{sem} denotes the semantic loss of point p_i. This equation shows that, if saliency score of point p_i is closer

Table 1: Results of SalLiDAR models with different backbones on FordSaliency dataset.

Model	Regression			Classification		
	CC ↑	SIM↑	KLD↓	CC↑	SIM↑	KLD↓
SalLiDAR w/ PointNet	0.3465	0.6655	0.4263	0.3636	0.6584	0.4892
SalLiDAR w/ RandLA-Net	0.6368	0.7784	0.1688	0.6381	0.7377	0.2282
SalLiDAR w/ Cylinder3D	**0.6760**	**0.7854**	**0.1574**	**0.6790**	**0.7834**	**0.1606**

to 1, the semantic loss of p_i would be weighted more; otherwise, the semantic loss would be kept without attentive weighting. It should be noted that attention guided optimization process is applied only at training stage; therefore, in this case, saliency prediction is not needed during inference.

2) SalLiDAR-II: Saliency distribution as a descriptor for semantic module. Several previous RGB saliency approaches [33, 34] have shown that the prediction of saliency can be adopted as an input descriptor to improve the model performance. Similarly, it could be regarded as a descriptor of the scene since it represents the prior knowledge of saliency distribution for the whole LiDAR scan. Inspired by the works in [33, 34], we take the predicted saliency values as an extra input descriptor for the semantic segmentation branch. In Figure 2, SalLiDAR-II is the model without the module of point-wise attention guided loss. In this model, the semantic module could learn saliency features for each point, which would be helpful to improve the performance of 3D semantic segmentation.

3) SalLiDAR-III: Saliency distribution as a descriptor and attentive loss guiding for semantic segmentation. In this model, we do not only combine the saliency distribution to the input of the semantic module, but also apply the point-wise attention guided loss to optimize the semantic segmentation stream. Combining both attention guided loss (SalLiDAR-I) and using saliency as descriptor (SalLiDAR-II), SalLiDAR-III is given in Figure 2.

4 Experimental Analysis

4.1 Experimental Setup

Implementation Details: We employ PointNet [23], RandLA-Net [10], and Cylinder3D [44] models as feature extractors. For the point cloud saliency models with classification modelling, the number of class K is 16. Experiments of K=$\{8, 16, 32, 64, 128, 256\}$ are discussed in ablation study. For point cloud saliency loss of equation 4, we use γ=1, δ=0.1, η=0.025 by following the studies [4, 6].

LiDAR FordSaliency Dataset: The large-scale FordCampus [21] dataset is composed of dataset1 (3817 scans) and dataset2 (6103 scans) of outdoor driving scenes. All the Velodyne LiDAR scans have point coordinates (x, y, z) and corresponding images of 5 omnidirectional camera views. Based on the data of FordCampus [21] dataset, we build a point cloud saliency dataset (namely FordSaliency) for the training of LiDAR-based saliency models. We select five state-of-the-art RGB saliency models as *pseudo-annotators* from MIT Saliency Benchmark[1]. These models include SAM [2], SalFBNet [4], MSI-Net [16], DeepGazeIIE [17], and SalGAN [20], which were pre-trained on RGB saliency datasets. We create saliency annotations of all LiDAR scans as described in Section 3.2. We utilize dataset1 and dataset2 of FordSaliency as validation set and training set, respectively.

SemanticKITTI Dataset: SemanticKITTI [1] dataset is a well-known large-scale dataset for point cloud semantic segmentation. This dataset consist of 22 Velodyne driving-scene

[1] https://saliency.tuebingen.ai/results.html

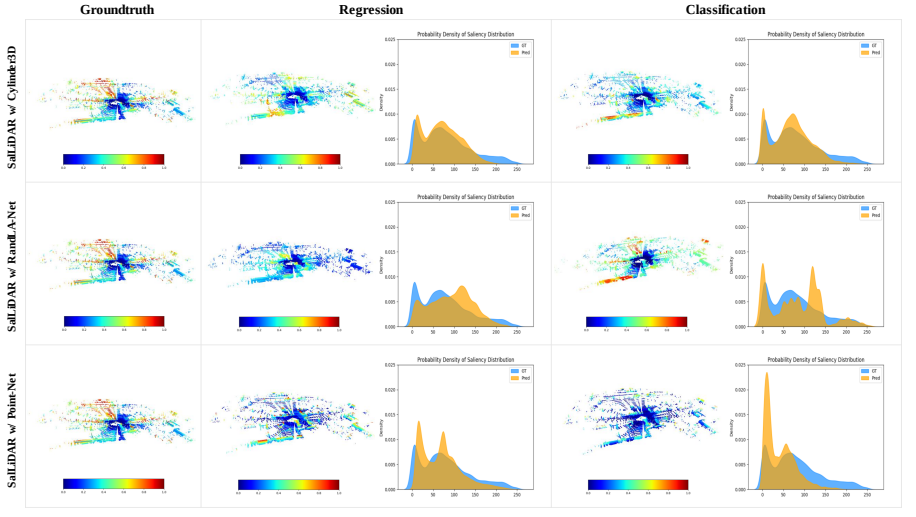

Figure 3: Point cloud saliency prediction results of SalLiDAR model with different backbones on FordSaliency dataset.

sequences, which is split into training set (sequences 00-07 and 09-10), validation set (sequence 08), and testing set (sequences 11-21). In total, there are 43,552 LiDAR scans and 19 semantic categories (*e.g. car*, *road*, *building*, etc.) [1].

Evaluation Metrics: We use popular saliency metrics[2] including Correlation Coefficient (CC), Similarity (SIM), and Kullback-Leibler Divergence (KLD) to evaluate the performance of point cloud saliency model. For LiDAR semantic segmentation, we adopt mean Intersection-over-Union (mIoU) as evaluation metric following the previous studies [10, 44].

4.2 Results on FordSaliency Dataset

We compare the performance of LiDAR-based saliency models with different feature extractors on our FordSaliency dataset. In Figure 3, We show the visualization results of SalLiDAR models with different backbones on FordSaliency validation set. In Table 1, we report the quantitative performance of these models on FordSaliency validation set. From Figure 3 and Table 1, we can observe that although the saliency annotations are pseudo-labels, all these LiDAR-based models are able to learn the discriminative point cloud saliency representations for saliency distribution prediction. Additionally, the models with classification and regression modelling manners can achieve competitive performance of saliency prediction, and the performance difference is small. It shows that these two modelling ways can be adopted to predict the point cloud saliency distribution. On the other hand, the model with Cylinder3D backbone can predict better saliency distribution than the model with other backbones. The models with RandLA-Net backbone and PointNet backbone can learn the correlation and similarity features from point cloud saliency annotations, as evidenced by the CC, SIM and KLD values in Table 1. However, the prediction of the model with Cylinder3D backbone can achieve higher CC, SIM, and lower KLD performance. It suggests that

[2] https://saliency.tuebingen.ai/evaluation.html

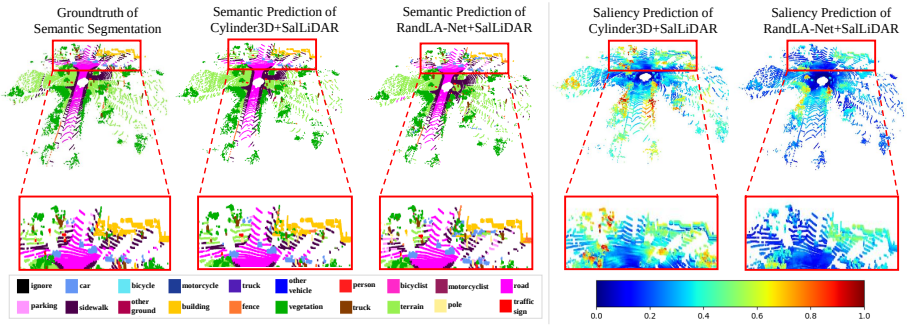

Figure 4: Visualization results of proposed segmentation models on SemanticKITTI [1].

Table 2: Performance comparison of proposed models and existing LiDAR segmentation methods on SemanticKITTI [1] test set. Results are obtained from leaderboard and literature.

Methods	mIoU	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
Darknet53 [8]	49.9	86.4	24.5	32.7	25.5	22.6	36.2	33.6	4.7	91.8	64.8	74.6	27.9	84.1	55.0	78.3	50.1	64.0	38.9	52.2
RangeNet++ [20]	52.2	91.4	25.7	34.4	25.7	23.0	38.3	38.8	4.8	91.8	65.0	75.2	27.8	87.4	58.6	80.5	55.1	64.6	47.9	55.9
RandLA-Net [20]	53.9	94.2	26.0	25.8	40.1	38.9	49.2	48.2	7.2	90.7	60.3	73.7	20.4	86.9	56.3	81.4	61.3	66.8	49.2	47.7
PolarNet [20]	54.3	93.8	40.3	30.1	22.9	28.5	43.2	40.2	5.6	90.8	61.7	74.4	21.7	90.0	61.3	84.0	65.5	67.8	51.8	57.5
SqueezeSegv3 [20]	55.9	92.5	38.7	36.5	29.6	33.0	45.6	46.2	20.1	91.7	63.4	74.8	26.4	89.0	59.4	82.0	58.7	65.4	49.6	58.9
Salsanext [8]	59.5	91.9	48.3	38.6	38.9	31.9	60.2	59.0	19.4	91.7	63.7	75.8	29.1	90.2	64.2	81.8	63.6	66.5	54.3	62.1
KPConv [20]	58.8	96.0	32.0	42.5	33.4	44.3	61.5	61.6	11.8	88.8	61.3	72.7	31.6	95.0	64.2	84.8	69.2	69.1	56.4	47.4
FusionNet [20]	61.3	95.3	47.5	37.7	41.8	34.5	59.5	56.8	11.9	91.8	68.8	77.1	30.8	92.5	69.4	84.5	69.8	68.5	60.4	66.5
KPRNet [20]	63.1	95.5	54.1	47.9	23.6	42.6	65.9	65.0	16.5	93.2	73.9	80.6	30.2	91.7	68.4	85.7	69.8	71.2	58.7	64.1
TORANDONet [8]	63.1	94.2	55.7	48.1	40.0	38.2	63.6	60.1	34.9	89.7	66.3	74.5	28.7	91.3	65.6	85.6	67.0	71.5	58.0	65.9
SPVNAS [20]	66.4	97.3	51.5	50.8	59.8	58.8	65.7	65.2	43.7	90.2	67.6	75.2	16.9	91.3	65.9	86.1	73.4	71.0	64.2	66.9
Cylinder3D [20]	67.8	97.1	67.6	64.0	59.0	58.6	73.9	67.9	36.0	91.4	65.1	75.5	32.3	91.0	66.5	85.4	71.8	68.5	62.6	65.6
PVKD [8]	71.2	97.0	67.9	69.3	53.5	60.2	75.1	73.5	50.5	91.8	70.9	77.5	41.0	92.4	69.4	86.5	73.8	71.9	64.9	65.8
* RandLA-Net (baseline)	52.5	93.8	27.0	22.0	36.1	38.1	49.9	42.5	6.4	90.7	58.8	74.1	11.5	88.9	57.4	79.8	61.2	65.5	49.9	46.0
RandLA-Net+SalLiDAR-I	**54.0**	**94.1**	**28.2**	**24.4**	**45.4**	37.2	48.3	**48.1**	5.9	89.1	**59.7**	72.4	**21.9**	87.5	56.2	**81.7**	**61.6**	**68.6**	49.7	**46.5**
RandLA-Net+SalLiDAR-II	**53.4**	93.8	**30.2**	**24.3**	**37.9**	37.5	**50.1**	**45.5**	**9.5**	89.9	**60.1**	73.9	**13.8**	87.3	56.6	**81.3**	60.7	**67.2**	48.0	**47.8**
RandLA-Net+SalLiDAR-III	**53.8**	**94.4**	**28.9**	**26.6**	35.5	**39.7**	47.0	**47.2**	**11.3**	90.0	**60.5**	73.7	**16.2**	88.3	56.8	**81.3**	60.9	**67.8**	**50.7**	45.8
‡ Cylinder3D (baseline)	71.7	97.1	69.6	72.0	55.8	62.4	76.2	77.8	46.7	91.2	69.8	76.2	40.9	92.6	70.2	86.7	73.8	71.6	65.2	66.3
Cylinder3D+SalLiDAR-I	**72.4**	**97.2**	**70.0**	**73.1**	**59.7**	**63.0**	**77.7**	**78.4**	**50.4**	**91.3**	**70.5**	**76.3**	**41.3**	**92.6**	69.9	86.5	73.4	70.8	**66.2**	**66.4**
Cylinder3D+SalLiDAR-II	**72.0**	**97.2**	69.0	**72.0**	**59.8**	**62.8**	**76.6**	77.3	**48.1**	**91.7**	**70.9**	**77.2**	**41.6**	92.5	69.4	86.4	73.4	71.2	**65.5**	65.3
Cylinder3D+SalLiDAR-III	**72.1**	**97.2**	69.2	**72.1**	**60.6**	**63.0**	**77.5**	77.7	**47.8**	**91.8**	**70.7**	**77.4**	**41.1**	**92.6**	69.5	86.4	73.6	71.1	**65.6**	65.4

* PyTorch implementation of RandLA-Net [20], which is available at: https://github.com/tsunghan-wu/RandLA-Net-pytorch.
‡ The results are obtained from the released version of Cylinder3D model [20] from the work in [8]: https://github.com/cardwing/Codes-for-PVKD.
Best performance results are shown in red color (publications before July 2022). Improved performance results of proposed model against the baseline are shown in **bold**.

the model with voxel-based partition (*e.g.* 3D Cylinder) could learn more powerful saliency representations than point-based models.

4.3 Results on SemanticKITTI Dataset

We report the LiDAR semantic segmentation performance on SemanticKITTI test set in Table 2. Note that all the testing performance results of Table 2 are taken from the literature and the benchmark leaderboard[3] of SemanticKITTI [1] dataset. As shown in Table 1, the SalLiDAR of regression models achieve better results than classification models. Thus we leverage regression based SalLiDAR models in Table 2. Comparing to the baselines, all the models with SalLiDAR obtain better mIoU results. The proposed method also improves the segmentation performance on specific classes, since the combination of our predicted saliency distribution makes the model attentive to these categories, such as *car*, *truck*, and *parking*. Furthermore, the Cylinder3D model with SalLiDAR achieves better segmenta-

[3] http://www.semantic-kitti.org/tasks.html#semseg

tion results than the RandLA-Net with SalLiDAR. It shows that the semantic segmentation model with better saliency prediction could provide more attentive information or features to improve the model performance. Especially, these experimental results demonstrate that the performance of LiDAR semantic segmentation models can be improved by proposed saliency distribution integration and point-wise attention guided loss. These comparison results validate the effectiveness of the pre-trained point cloud saliency models, although they are trained on FordSaliency dataset with pseudo-annotations.

4.4 Complexity Analysis

Regarding SalLiDAR-II and III, complexity is doubled compared to its backbone model because use of saliency as input or feature descriptor requires saliency prediction network to be processed. But it is the case if only saliency is used as input descriptor. On the other hand, it should be noted that cost of SalLiDAR-I is doubled only during the training phase while the inference complexity is the same as its backbone/original model, as it is only used for optimizing the backbone model with a saliency guided loss. In addition, saliency concept regardless of in 2D or 3D vision, brings extra cost similar to other image processing techniques (e.g. normal estimation). But, the potential benefits are also very promising. As our results (see Table 2) demonstrates, the proposed method improves the accuracy of many categories (e.g. truck, person) compared with the baseline model.

5 Conclusion

In this paper, we propose a practical solution for point cloud saliency prediction. We first build a point cloud saliency dataset (FordSaliency) for the training of LiDAR-based saliency models. We then employ existing LiDAR-based models as backbones of saliency feature extraction to learn the saliency distribution on point clouds. After obtaining the learned saliency embedding features, we calculate saliency score map by utilizing two different type of predictors including classification and regression. To demonstrate the effectiveness of the point cloud saliency models trained on FordSaliency dataset, we design a two-stream semantic segmentation model by combing saliency representations and semantic features. Extensive experimental results on FordSaliency dataset show that saliency distribution could be transferred from color images to point clouds effectively. Experiments on SemanticKITTI dataset also confirm that utilizing point cloud saliency information has a high potential to improve the 3D scene understanding tasks with the 3D segmentation performance gain.

Acknowledgement

This paper is in part based on the results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan. This work was supported by JST SPRING, Grant Number JPMJSP2124. Computational resource of *AI Bridging Cloud Infrastructure (ABCI)*[4] provided by National Institute of Advanced Industrial Science and Technology (AIST) was used for training and testing the models during our experiments.

[4] https://abci.ai/

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9297–9307, 2019.

[2] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

[3] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *International Symposium on Visual Computing*, pages 207–222. Springer, 2020.

[4] Guanqun Ding, Nevrez İmamoğlu, Ali Caglayan, Masahiro Murakawa, and Ryosuke Nakamura. Salfbnet: Learning pseudo-saliency distribution via feedback convolutional networks. *Image and Vision Computing*, 120:104395, 2022.

[5] Xiaoying Ding, Weisi Lin, Zhenzhong Chen, and Xinfeng Zhang. Point cloud saliency detection by local and global feature fusion. *IEEE Transactions on Image Processing*, 28(11):5379–5393, 2019.

[6] Richard Droste, Jianbo Jiao, and J Alison Noble. Unified image and video saliency modeling. In *European Conference on Computer Vision*, pages 419–435. Springer, 2020.

[7] Martin Gerdzhev, Ryan Razani, Ehsan Taghavi, and Liu Bingbing. Tornado-net: multiview total variation semantic segmentation with diamond inception module. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9543–9549. IEEE, 2021.

[8] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[9] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8479–8488, 2022.

[10] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020.

[11] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[12] Lai Jiang, Mai Xu, Xiaofei Wang, and Leonid Sigal. Saliency-guided image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16509–16518, 2021.

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[14] Hanjae Kim, Sunghun Joung, Ig-Jae Kim, and Kwanghoon Sohn. Prototype-guided saliency feature learning for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4865–4874, 2021.

[15] Deyvid Kochanov, Fatemeh Karimi Nejadasl, and Olaf Booij. Kprnet: Improving projection-based lidar semantic segmentation. *arXiv preprint arXiv:2007.12668*, 2020.

[16] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020.

[17] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12919–12928, 2021.

[18] Chunlei Liu, Wenrui Ding, Jinyu Yang, Vittorio Murino, Baochang Zhang, Jungong Han, and Guodong Guo. Aggregation signature for small object tracking. *IEEE Transactions on Image Processing*, 29:1738–1747, 2019.

[19] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220. IEEE, 2019.

[20] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[21] Gaurav Pandey, James R McBride, and Ryan M Eustice. Ford campus vision and lidar data set. *The International Journal of Robotics Research*, 30(13):1543–1552, 2011.

[22] Gaurav Pandey, James R McBride, Silvio Savarese, and Ryan M Eustice. Automatic targetless extrinsic calibration of a 3d lidar and camera by maximizing mutual information. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[23] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[24] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2018.

[25] Xuena Ren, Dongming Zhang, Xiuguo Bao, and Yongdong Zhang. S^2-net: Semantic and salient attention network for person re-identification. *IEEE Transactions on Multimedia*, 2022.

[26] Elizabeth Shtrom, George Leifman, and Ayellet Tal. Saliency detection in large point sets. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3591–3598, 2013.

[27] 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*, pages 685–702. Springer, 2020.

[28] Flora Ponjou Tasse, Jiri Kosinka, and Neil Dodgson. Cluster-based point set saliency. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 163–171, 2015.

[29] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6411–6420, 2019.

[30] Georgi Tinchev, Adrian Penate-Sanchez, and Maurice Fallon. Skd: Keypoint detection for point clouds using saliency estimation. *IEEE Robotics and Automation Letters*, 6 (2):3785–3792, 2021.

[31] Guangming Wang, Xinrui Wu, Zhe Liu, and Hesheng Wang. Pwclo-net: Deep lidar odometry in 3d point clouds using hierarchical embedding mask optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15910–15919, 2021.

[32] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[33] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *European Conference on Computer Vision*, pages 825–841. Springer, 2016.

[34] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017.

[35] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2020.

[36] 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*, pages 16024–16033, 2021.

[37] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7693–7702, 2022.

[38] Feihu Zhang, Jin Fang, Benjamin Wah, and Philip Torr. Deep fusionnet for point cloud semantic segmentation. In *European Conference on Computer Vision*, pages 644–663. Springer, 2020.

[39] 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*, pages 9601–9610, 2020.

[40] Rui Zhao, Wanli Oyang, and Xiaogang Wang. Person re-identification by saliency learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2): 356–370, 2016.

[41] Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, and Kui Ren. Pointcloud saliency maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1598–1606, 2019.

[42] Wu Zheng, Weiliang Tang, Li Jiang, and Chi-Wing Fu. Se-ssd: Self-ensembling single-stage object detector from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14494–14503, 2021.

[43] Zikun Zhou, Wenjie Pei, Xin Li, Hongpeng Wang, Feng Zheng, and Zhenyu He. Saliency-associated object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9866–9875, 2021.

[44] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Wei Li, Yuexin Ma, Hongsheng Li, Ruigang Yang, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.