Fine-Grained Off-Road Semantic Segmentation and Mapping via Contrastive Learning

Biao Gao[1], Shaochi Hu[1], Xijun Zhao[2], Huijing Zhao[1]

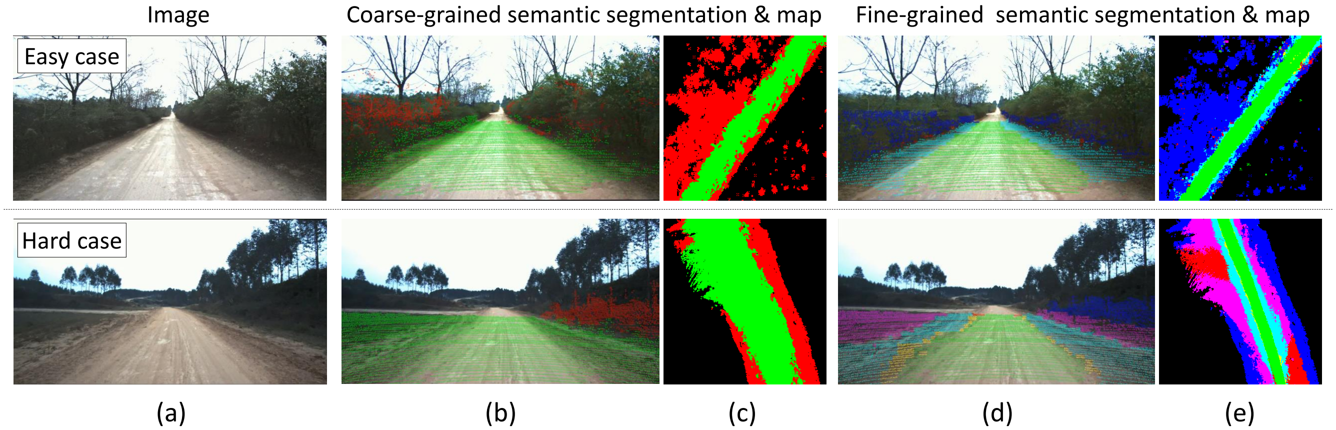


Image
Coarse-grained semantic segmentation & map
Fine-grained semantic segmentation & map
Easy case
Hard case
(a)
(b)
(c)
(d)
(e)

Fig. 1: The significance of fine-grained semantic segmentation and mapping in off-road environment, where coarse-grained results can hardly adapt diverse scenes with unified threshold. (a) scene image. (b) coarse-grained semantic segmentation (binary classification). (c) coarse-grained bird's-eye-view semantic map. (d) fine-grained semantic segmentation. (e) fine-grained bird's-eye-view semantic map.

***Abstract*— empty.**

I. Introduction

Recent years, a considerable development has grown up around the theme of intelligent vehicles. Driving scene understanding plays a critical role in the preconditions of decision planning and self-driving. For the moment, a considerable amount of studies for urban scenes has been published. However, the off-road environment mainly consists of natural objects and lacks structured features, leading to ambiguous definition of this problem. Different self-driving platforms have diverse trafficability, while the fine-grained semantic understanding could help to distinguish driving areas with diverse traversability cost. When facing off-road scenarios like Fig. 1(a), how to extract appropriate drivable area for self-driving platforms? How to evaluate traversability cost of different regions? To answer these questions, fine-grained semantic understanding is in great request. However, the particularity of off-road environment results in lack of clear definition and widely recognized standards.

Existing researches can be classified as traditional methods and deep learning methods. Traditional methods mainly make use of geometrical and visual features to extract semantic meanings, but usually depend on manual defined features and empirical parameters, limiting its performance. As shown in Fig. 1(b-c), traditional methods may be in trouble when facing diverse scenarios. With uniform parameters, the drivable area may be too wide to provide human preferred area at flat scene, while cannot extract consecutive drivable area at rugged scene. Expected fine-grained semantic segmentation as shown in Fig. 1(d-e) could extract different semantic regions in different scenes, which improves the traversability analysis ability for off-road driving. Deep learning methods depend on deep networks to obtain better feature representations, but rely heavily on massive annotated data. In off-road environments, the ambiguous problem definition leads to difficulty and scarcity of human annotations. Several attempts about contrastive learning have been made in computer vision researches, which can learn effective feature representations through self-supervised pipeline. It has been proved to support fine-grained image classification tasks in large-scale datasets like ImageNet [1].

To solve the challenges in fine-grained off-road semantic segmentation, i.e. ambiguous problem definition and the scarcity of elaborate annotated datasets, this work proposed a fine-grained off-road semantic segmentation method based on contrastive learning techniques. Compared with traditional methods, the proposed framework can automatically learn feature representations through contrastive learning. Meanwhile, only a small number of anchor annotations are required to get fine-grained semantic segmentation results,

*This work is partially supported by ***.

[1]B. Gao, S. Hu and H. Zhao are with the Key Lab of Machine Perception (MOE), Peking University, Beijing, China. [2]X. Zhao is with China North Vehicle Research Institute, Beijing, China.

Correspondence: H. Zhao, zhaohj@cis.pku.edu.cn.

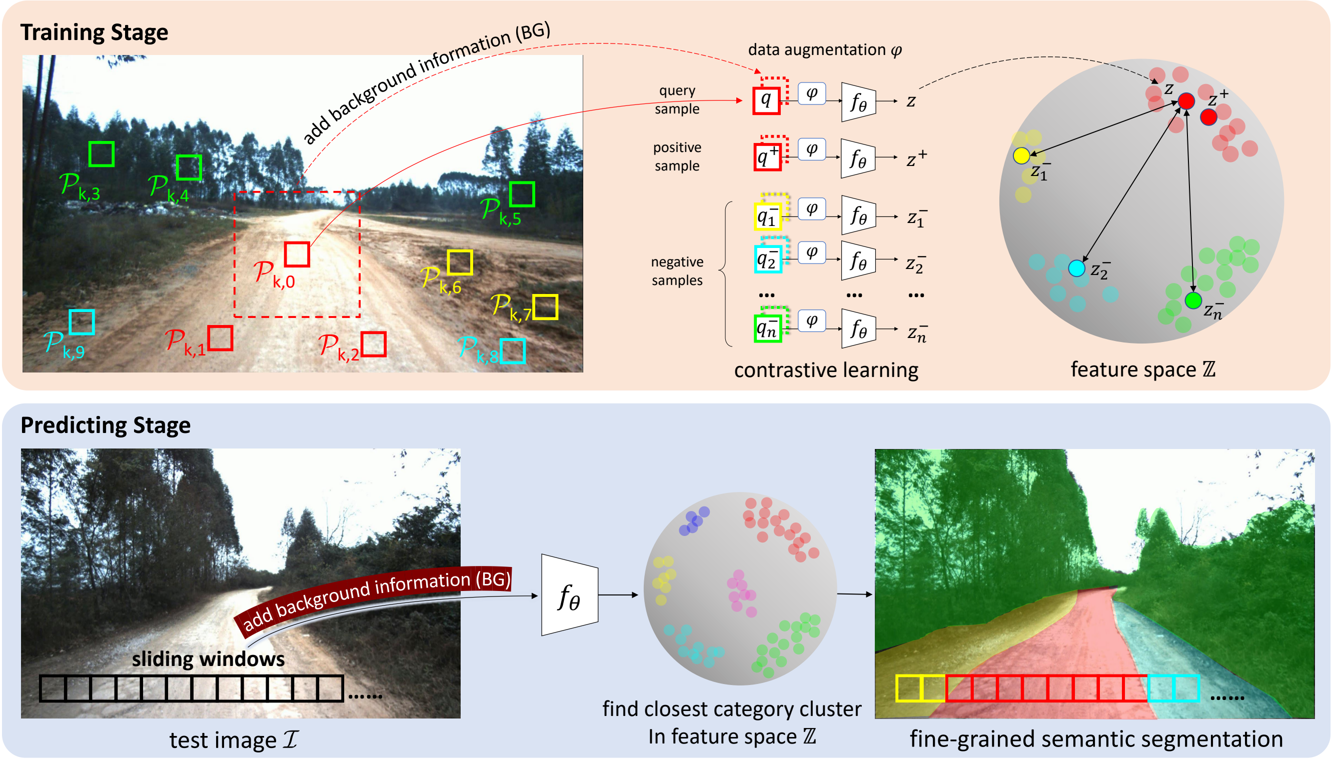


Fig. 2: The proposed pipeline for fine-grained off-road semantic segmentation via contrastive learning.

which significantly reduce the deep network's demand of laborious human annotations. The experimental results prove the validity of our proposed method, and show its potential in applications like semantic mapping.

This paper is organized as follows. First, the related works are introduced in Section II. Section III presents the proposed methodology in detail. Section IV shows experimental results. Finally, we draw conclusions in Section V.

II. Related Works

A. Traditional Methods

Segmentation-based: AdaBoost, SVM, GMM

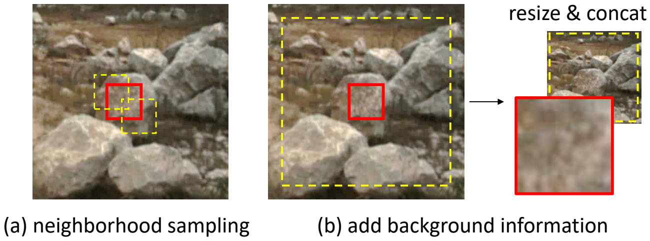
B. Deep Learning Methods

C. Contrastive Learning

CMC, MoCo, SimCLR

III. Methodology

A. Problem Formulation



(a) neighborhood sampling

(b) add background information

Fig. 3: Illustration of (a) neighborhood sampling strategy, and (b) how to add background information with the origin image patch.

A training image I_k has a number of anchor patches $A_k = \{\mathcal{P}_{k,i} =< p_{k,i}, a_{k,i} >\}$, where an anchor patch $\mathcal{P}_{k,i}$ is a pair of an image patch $p_{k,i}$ and a label $a_{k,i}$. Here, $a_{k,i}$ has no semantic meaning, but is an identifier of the image patches with similar or different semantic properties. Let $z = f_\theta(p)$ be an encoder converting a high-dimensional image patch p to a low-dimensional feature vector $z \in \mathbb{Z}^D$. We use exponential cosine distance $d(p_i, p_j) = exp(z_i^T \cdot z_j)$ to measure the similarity of two image patches via their low-dimensional feature vectors. Therefore, given an image patch $\mathcal{P}_{k,i}$, its distance to another image patch $\mathcal{P}_{k,j}$, i.e. $d(p_{k,i}, p_{k,j})$, should be smaller if they share the same label $a_{k,i} = a_{k,j}$, whereas larger if the labels are different $a_{k,i} \neq a_{k,j}$. In order to make the annotation operational easy, in this research, the labels of the anchor patches are comparable only if they belong to the same image.

Given a set of training images $\mathcal{I} = \{I_k\}$ with anchor patches $\mathcal{A} = \{A_k\}$ on each of them, this research is to find a representation f_θ that encodes image patch p to z, where at the low-dimensional feature space $\mathbb{Z}^D$, the zs of similar semantic meaning distribute closely. This research finds f_θ through contrastive learning, which is further used in an application of fine-grained semantic segmentation for off-road traversability analysis.

B. Feature Representation through Contrastive Learning

1) Sampling strategy: In each training step, an anchor patch $\mathcal{P}_{k,i}$ is selected to compose a query sample q, and a positive sample q^+ and n negative samples $\{q_i^-|i=1,..,n\}$ are subsequently composed on the anchor patches of the same image I_k.

Based on the label $a_{k,i}$ of $\mathcal{P}_{k,i}$, the anchor patches of the same image I_k are divided into two sets, where $\{\mathcal{P}_{k,i}^+\}$ denotes those sharing the same label $a_{k,i}$, whereas $\{\mathcal{P}_{k,i}^-\}$ for the rest. Assuming that an off-road scene is spatially continuous, i.e. nearby regions could be semantically similar. An anchor patch is first selected randomly from $\{\mathcal{P}_{k,i}^+\}$, where an image patch is randomly clipped from its neighborhood to compose a positive sample q^+. As illustrated in Fig. 3(a), the neighborhood is defined to have the center point of the randomly clipped image patch within the original one. Similarly, n negative samples $\{q_i^-\}$ are composed on $\{\mathcal{P}_{k,i}^-\}$.

2) Composing sample data: With an image patch p, a sample data is composed in the same way for query, positive or negative samples. As shown in Fig. 3(b), a sample data contains foreground and background image patches to describe both local and global features. Foreground is image patch p, while background is centered at p but with a larger region to provide global scene context. The foreground and background patches are firstly resized to the same scale, then the two patches are concatenated along the channel dimension to compose a 6-channel tensor.

In order to improve robustness in diverse scenes, data augmentation (denoted by ϕ in Fig. 2) is conducted on the 6-channel tensor of each sample data before forwarding it to the network of f_θ. In this research, data augmentation includes random flip, random grey scale and color jitter, which randomly changes the brightness, contrast and saturation of an image.

3) Network Design and Loss Function: A CNN backbone network in practical terms, e.g. AlexNet [2] is used to model f_θ, which converts the 6-channel tensor of a query, positive or negative sample to a low-dimensional feature vector $z \in \mathbb{Z}^D$. Contrastive learning is used to find a θ of f_θ, with which the exponential cosine distance of the zs are close if they share the same labels, whereas far for those different. Following the principle of previous contrastive learning studies, a contrastive loss function InfoNCE [4] is implemented:

$$L = -\log \frac{\exp(z^T \cdot z^+/\tau)}{\exp(z^T \cdot z^+/\tau) + \sum_{i=0}^{n} \exp(z^T \cdot z_i^-/\tau)} \quad (1)$$

where τ denotes a temperature hyper-parameter.

In this work, since the positive and negative samples are comparable only in the same image, the limited quantity makes it possible to get the feature representations with reasonable memory consumption. In practice, unlike the typical contrastive learning studies [3] using memory bank to store feature vectors for each training sample, we randomly select positive/negative samples and calculate their features at each training step.

C. Off-road Semantic Segmentation and Mapping

As illustrated in Fig. 2, the work flow contains off-line learning and on-line prediction, while the latter is composed of further two steps: semantic segmentation of single images and semantic mapping using multiple images.

1) Off-line learning: Given a set of training images $\mathcal{I} = \{I_k\}$ with anchor patches $\mathcal{A} = \{A_k\}$ on each of them, a feature encoder f_θ is thus learnt to convert each image patch to a low-dimensional vector $z \in \mathbb{Z}^D$, where in the space of $\mathbb{Z}^D$, the anchor patches with the same labels are projected close on the exponential cosine distance $exp(z_i^T \cdot z_j)$, whereas far for the others.

The zs of the anchor patches are then clustered by K-means method, where a set of mean points $\mathcal{C} = \{\tilde{z}_c\}$ are extracted, representing the features of $\mathcal{K}$ dominant semantic clusters. Here, $\mathcal{K}$ is a hyper-parameter, which decides granularity of semantic segmentation.

2) Semantic segmentation: Given the current image $\mathcal{I}$, semantic segmentation is conducted by generating image patches using sliding windows, and predicting a semantic label for each image patch. Given an image patch p_i, a semantic label is predicted as follows. A 6-channel tensor data is first composed, containing both local and global features of the image patch. The data is then projected by f_θ to a lower-dimensional feature vector z_i, which is subsequently compared with the set of feature vectors $\mathcal{C} = \{\tilde{z}_c\}$ representing the $\mathcal{K}$ dominant semantic labels. The image patch is assigned the semantic label that has the best match on its feature vector, i.e. $a_i = \arg\min_c exp(z_i^T \cdot \tilde{z}_c)$.

To make up denser semantic segmentation, we could adjust step size of sliding windows. For example, we can assign the semantic label to $3*3$ pixels centered at each image patch, while setting sliding windows' horizontal/vertical step size to 3 pixels, then get denser semantic segmentation results.

3) Semantic mapping: Centered at the ego vehicle's location in the frame, a horizontal plane is drawn at the ground level and tessellated into regular grids. The pixel labels of the current image are projected to corresponding 3D LiDAR point clouds with the calibration parameters, and then projected onto the grids with additional vehicle localization data at each frame. Since a single grid can have multiple label predictions, let $\sigma_{x,y}^c$ denote the counts of predicting label c of grid (x, y), the semantic label $l_{x,y} = \text{argmax}_c(\sigma_{x,y}^c)$ is assigned to the grid. Meanwhile, a confidence map is estimated too indicating the confidence of the grids' predicted labels. The confidence value of grid (x, y) is assigned as $\max(\sigma_{x,y}^c)/\sum \sigma_{x,y}^c$, which can also serve as a measure to evaluate prediction consistency.

IV. Experimental Results

A. Dataset

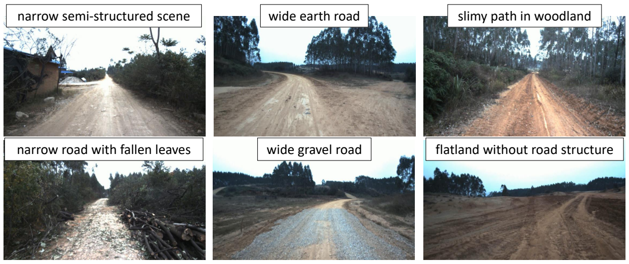
The performance of the proposed method is evaluated on our off-road dataset. The dataset is collected by an instrumented vehicle with a front-view monocular camera, a GPS/IMU suite and a 3D LiDAR. In this work, we mainly use camera images as input data, while the GPS/IMU and

LiDAR data are supplementary for semantic mapping. As shown in Table I, the off-road dataset includes 3 subsets. Take subset A as an example, we randomly selected 50 frames for 973 anchors annotation and training, which only account for about 10% of total 5064 frames. While all frames are evaluated when testing, and frames with human annotations are taken into account in quantitative evaluation.

In addition, as shown in Fig. 4, the 3 subsets represent different typical off-road environments. The scenarios in subset A are mostly narrow roads with bushes aside. Subset B are relatively wide scenes, and subset C includes diverse scenarios like slimy path in woodland and flatland without road structure. In following experiments, we train and test the proposed method on different subsets to evaluate its cross-scene generalization performance.

TABLE I: Statistics of the off-road dataset

	subset A	subset B	subset C
total frames	5064	3239	4098
frames for training	50	100	80
anchors	973	1606	1437


narrow semi-structured scene
wide earth road
slimy path in woodland
narrow road with fallen leaves
wide gravel road
flatland without road structure

(a) subset A (b) subset B (c) subset C

Fig. 4: Typical scenes in three datasets, which represent different off-road environments.

B. Evaluation Metrics

Suppose that there are N anchors in one frame, then any two anchors must be either positive or negative samples of each other. Hence, there exists $N \cdot (N-1)$ pairs anchor constraints. We denote positive samples' constraints as $Pos(i,j)$: if anchor a_i and a_j are positive samples of each other and classified into the same cluster, $Pos(i,j) = 1$. Otherwise, if they are not classified to the same cluster, $Pos(i,j) = 0$. Negative samples' constraints are defined in a similar way, and denoted as $Neg(i,j)$.

We use the following metrics to evaluate how well the clustering results fit human annotated anchors:

$$\mathcal{R} = \frac{\sum_{i,j} Pos(i,j) + \sum_{i,j} Neg(i,j)}{N \cdot (N-1)}, i \neq j \tag{2}$$

Essentially, it can be seen as Rand Index [5], which is a commonly used measurement for clustering.

C. Results on Proposed Method

To evaluate the proposed method, we design the following experiments: (1) feature distance measurement, explore the validity of feature encoder and distance measurement learned by contrastive learning. (2) cross validation and ablation study, verify the performance and robustness of our proposed method in diverse test scenes, while explore the effects of different modules or parameters. (3) fine-grained semantic segmentation and mapping, analyze the fine-grained semantic predictions through case study and semantic mapping, to show the method's validity for fine-grained off-road traversability analysis.

1) Feature Distance Measurement: The core module in our proposed pipeline is the feature encoder f_θ, which projects high-dimensional image patch to low-dimensional feature vector in space $\mathbb{Z}$. Its purpose is making feature distance closer between similar image patches, while farther between different image patches. In Fig. 5, we visualize some case studies. In all images, the query anchors are circled by yellow rings, while the other anchor patches are randomly sampled and colorized by its feature distance to the query anchor. For example, in Fig. 3(a), the query anchor is located on earth road. We can find that patches on earth road are closer to red, and other patches located on different semantic area are generally blue, which indicate farther distance to the query anchor. The feature distance distribution is accord with human annotated anchor labels. Similar situations are general on images (a)-(f). As a result, the learned feature encoder and distance measurement are able to distinguish similar or different image patches.

2) Cross Validation and Ablation Study: For comprehensive evaluation of the proposed method, we make cross validation on models with different settings, and the statistics of $\mathcal{R}$ are shown in Table II. The table cells are colorized by column data, when training and testing on different subsets. The last column lists the average performance $\bar{\mathcal{R}}$ of models on test sets (different subsets with the training one). It is obviously that *BG320* has the best performance on test sets, and all three models with background information have $\mathcal{R}$ over 0.85 among all training/testing combinations, which demonstrates the robustness of our proposed method.

Comparing the basic data augmentation with background information, they can both increase models' performance, while the latter makes more contribution. Besides, increasing background size could sightly improve the overall performance, but not obvious at all situations.

By the way, above experiments uniformly used clustering number $\mathcal{K} = 6$. How does clustering number affects models' performance? An ablation study of $\mathcal{K}$ is made, and the results are shown in Fig. 6. We can find that the models' performance with regard to $\mathcal{K}$ are basically stable when $\mathcal{K} \geq 4$, and slightly decrease when $\mathcal{K} > 6$. In general, models' performance with different $\mathcal{K}$s approximately order the same as Table II. Therefore, we choose $\mathcal{K} = 6$ as other experiments' setting, which balances the fine-grained demand and model performance.

3) Fine-Grained Semantic Segmentation and Mapping: Due to the absent of ground truth for fine-grained off-road semantic segmentation task, we next analyze the validity of our fine-grained results through case study of semantic

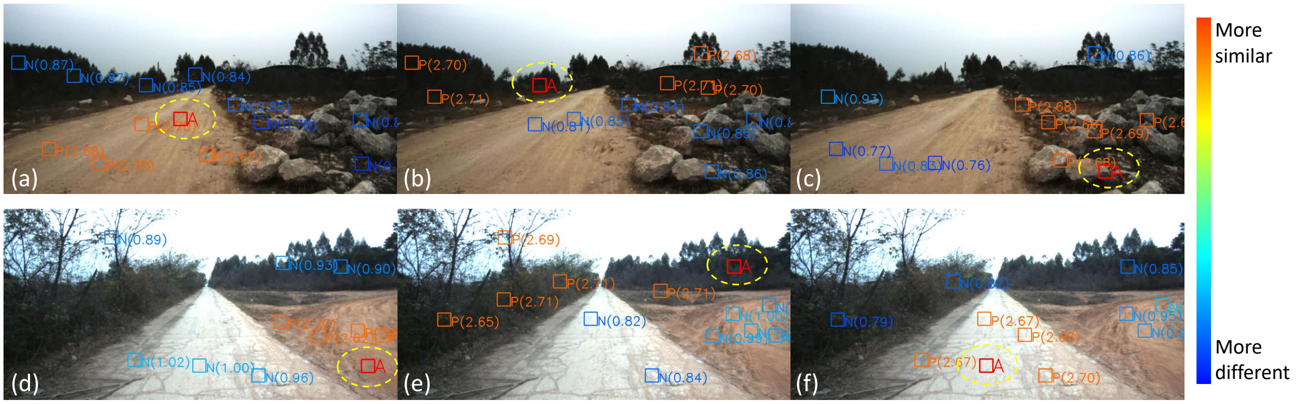


Fig. 5: Visualization of feature distances of query anchor (A) to its positive (P) and negative (N) samples. Query anchors are circled by yellow rings. P/N is according to human annotated anchor labels, and numbers in parentheses measure samples' similarity to the query anchor.

TABLE II: Cross Validation Results ($\mathcal{R}$) on Different Datasets

model	data aug.	BG size	train on subset A			train on subset B			train on subset C			$\bar{\mathcal{R}}$ on test sets
			test on			test on			test on			
			A	B	C	B	A	C	C	A	B	
base	×	×	0.9854	0.8548	0.8509	0.9997	0.7957	0.8492	0.9966	0.8288	0.9258	0.8509
base_DA	✓	×	0.9693	0.8792	0.8422	0.9959	0.8210	0.8625	0.9913	0.8296	0.9119	0.8578
BG192	✓	192	0.9939	0.9330	**0.8650**	0.9994	0.8524	**0.8899**	0.9944	0.8653	0.9468	0.8920
BG256	✓	256	0.9987	0.9360	0.8627	0.9991	0.8577	0.8839	0.9934	0.8665	0.9512	0.8930
BG320	✓	320	0.9986	**0.9433**	0.8559	0.9980	**0.8667**	0.8895	0.9958	**0.8776**	**0.9544**	**0.8979**

* **BG**: background; **base**: basic pipeline without data augmentation or background information; **base_DA**: with basic data augmentation, without background information; **BG192/256/320**: complete pipeline with different background size.

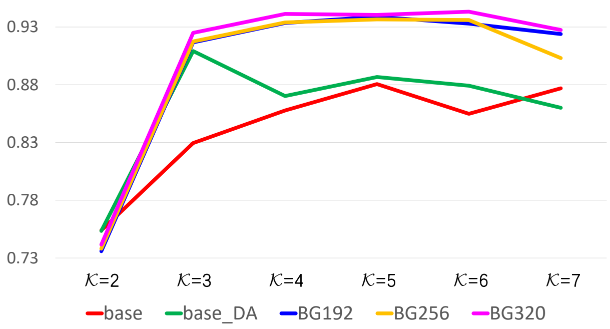


Fig. 6: Average $\mathcal{R}$ of models under different clustering number $\mathcal{K}$.

segmentation and mapping. Besides, we compare different categories traversability cost by additional LiDAR data. The following results are all based on the model trained by 50 frames of subset A.

Fig. 7 shows some cases of fine-grained semantic segmentation. This work focuses on off-road traversability analysis, so we do not pay attention to the sky area, and only bottom half of the image are predicted for simplicity. The semantic labels are not pre-assigned, we can find some uniform semantic meanings through these concrete cases. For example, green indicates hard earth road and paved road, blue pixels are vegetation, yellow pixels are road with fallen leaves or muddy area, red pixels are stones or woods, etc. Different colors represent different clusters, and they can generally distinguish diverse semantic meanings.

To evaluate the overall performance and consistence of the fine-grained prediction on continuous video frames, we make semantic maps and confidence maps as described in Sec. III-C.3. As shown in Fig. 8(d)(e), our fine-grained predictions can label the roadside area (yellow) with higher traversability cost than middle road (green), while the traditional coarse-grained region grow method is unable to distinguish them. In Fig. 8(case 2), let us pay attention to bumps in the middle of the road, which is a hard case. Although it has been separated in the single frame prediction in Fig. 8(d), its segmentation is not stable enough to obtain majority votes in the semantic map. The good news is, confidence map in (f)(g) can be helpful to distinguish this subtle traversability difference, where the bumps area are darker than other well-travelled road.

More than case study of fine-grained semantic, segmentation, a statistical traversability analysis is provided in Fig. 9, which is based on 3D point clouds with labels projected from fine-grained image semantic segmentation. By the way, the semantic meanings of color table is not pre-defined, but concluded from our model's predictions. In Fig. 9(a-c), it is obvious that green, yellow and cyan are mainly distribute around the ground level, which are three primary

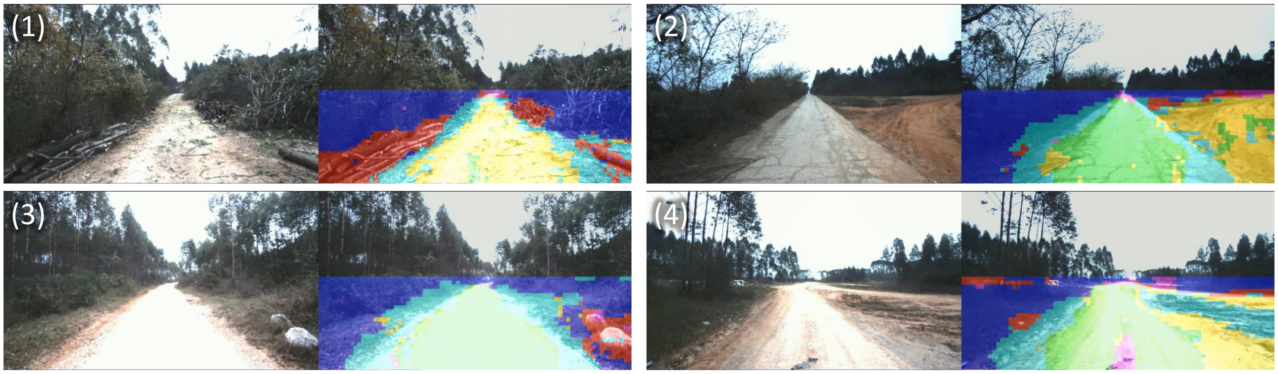


Fig. 7: Some cases of fine-grained semantic segmentation.

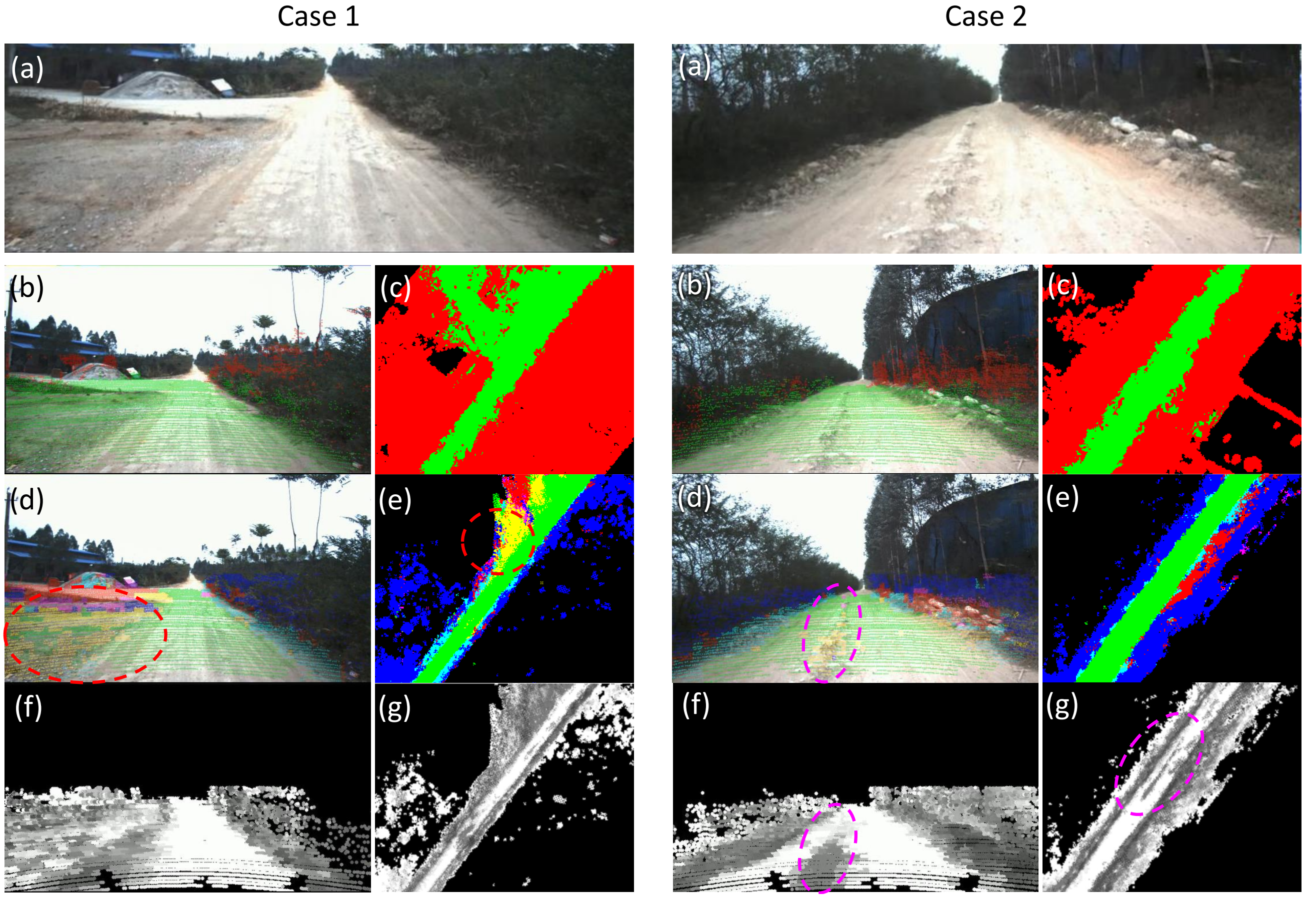


Fig. 8: Case study of fine-grained semantic map and confidence map, compared with coarse-grained road extraction results. (a) video image. (b) coarse-grained segmentation. (c) coarse-grained bird's-eye-view semantic map. (d) fine-grained semantic segmentation projected by point clouds. (e) fine-grained bird's-eye-view semantic map. (f) confidence map projected on camera-view, the whiter the higher confidence. (g) bird's-eye-view confidence map.

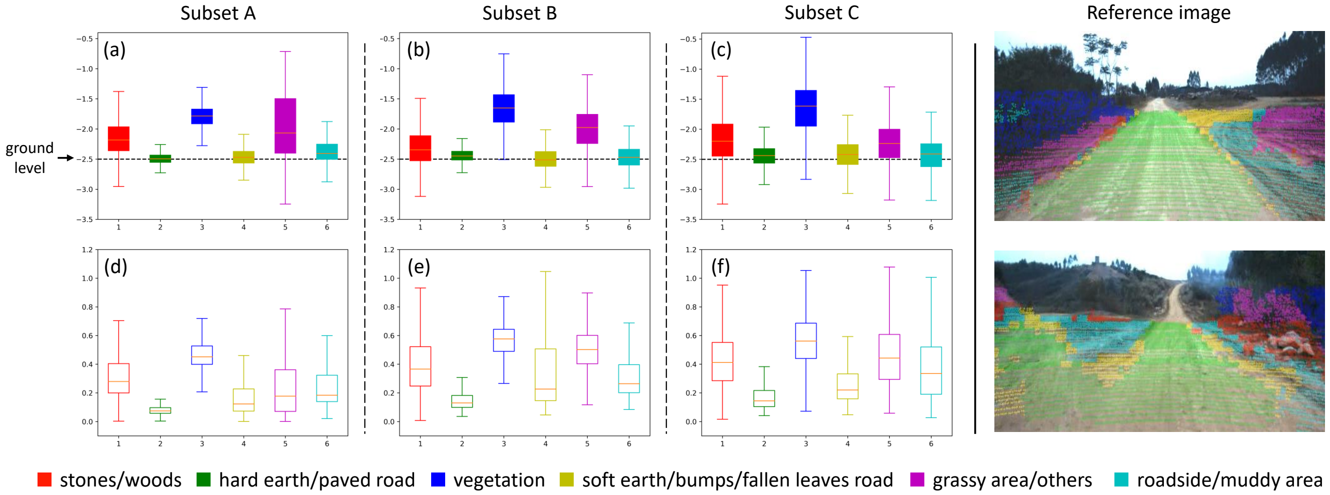


Fig. 9: Traversability analysis of semantic clusters based on point clouds. (a-c) boxplots of points average height, indicate height distribution of different categories. (d-f) boxplots of points height variance, indicate surface flatness and traversability cost.

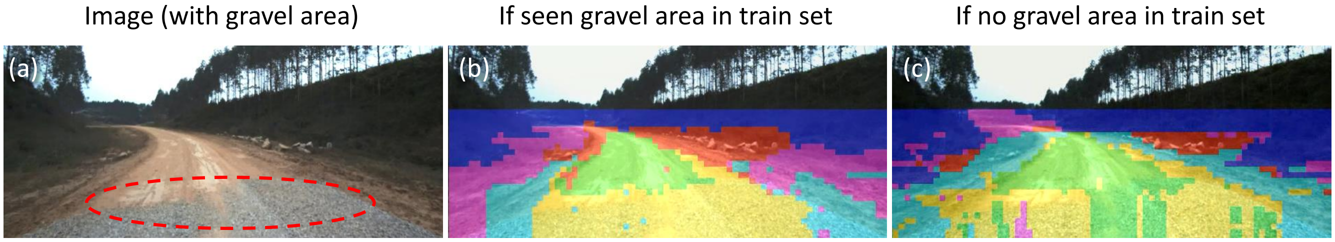


Image (with gravel area)
If seen gravel area in train set
If no gravel area in train set
(a)
(b)
(c)

Fig. 10: Challenging case: when meeting unseen semantic categories.

road types. Furthermore, in Fig. 9(d-f), we can find their different traversability cost, where green points have the narrowest variance distribution, corresponding to the most well-travelled paved road and hard earth road. Yellow and cyan boxes are longer, indicating more bumpy road surface. They are mainly soft earth, bumps or muddy area at the roadside. Blue boxes are mostly bushes and trees, with the highest average height and traversability cost, which is in accord with boxplots distribution. In summary, the statistical analysis of additional 3D LiDAR data can prove the validity of our fine-grained off-road semantic segmentation.

D. Challenges

Currently, there are still some challenges for the proposed method. Firstly, the current pipeline to obtain dense predictions has relatively high computational cost, which can be optimized by temporal and spatial consistency in future works. The second one is unseen semantic categories, or called out of distribution (OOD) samples, as shown in Fig. 10. The current pipeline will not discriminate unseen category samples, but simply classified them into existing clusters, which may lead to confused predictions as Fig. 10(c). To minimize labor cost, the OOD sample detection and incremental training mechanism deserve to be explored in our future works.

V. CONCLUSIONS

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[3] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[5] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.