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

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

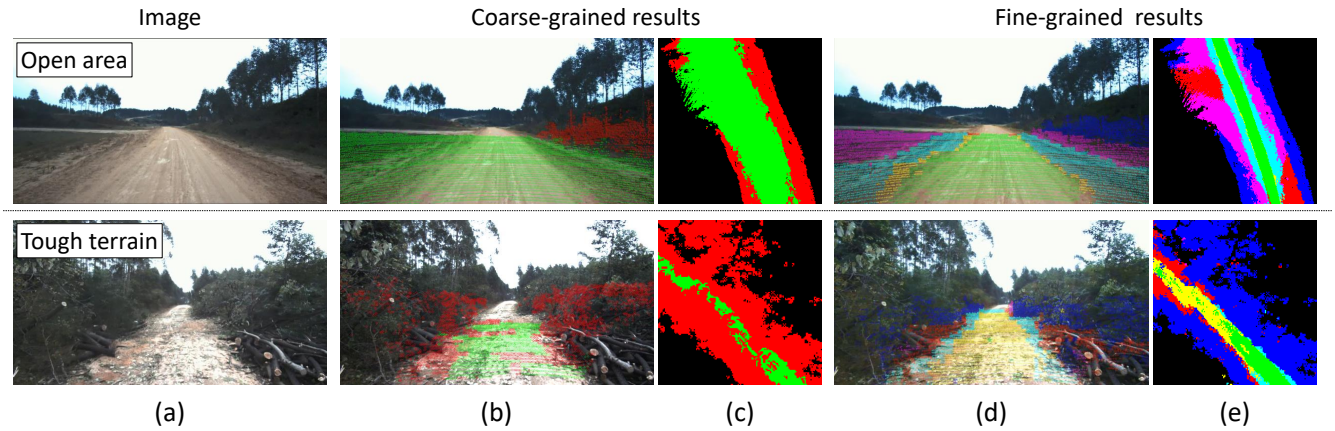

Fig. 1: The significance of fine-grained semantic segmentation and mapping in off-road environment, where coarse-grained results can hardly adapt diverse scenes. (a) scene image. (b) coarse-grained semantic segmentation (binary classification). (c) coarse-grained semantic map (bird's-eye-view). (d) fine-grained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

***Abstract*— Driving scene understanding plays a vital role for mobile robots, especially when traversing in complex off-road scenes. Early researches usually formulated the problem as a coarse-grained segmentation, i.e., road and non-road area. Nevertheless, the mechanical performance of off-road robots can be very diverse, leading to different definition of safe regions to traverse, which requests a fine-grained understanding ability. For this purpose, we propose a contrastive learning based method to achieve fine-grained semantic segmentation and mapping of off-road scenes. It does not need pixel-level labels, which is time-consuming and difficult to annotate at natural scenes, but only relies on a few sparse anchor patches indicating regions with different semantic attributes. The experiments across diverse off-road scenes illustrate the validity of our method and show its potential in applications for off-road environments.**

I. Introduction

Mobile robotic and autonomous driving techniques have been witnessed of tremendous progress in recent years [1]. Driving scene understanding plays a vital role as a prerequisite for the decision making and planning of a robot to traverse in complex environment [2]. Nowadays researches are mainly oriented to the applications at structural scenes such as indoor, parking lots, urban streets, highways, etc. [3], whereas research on understanding off-road environments is rare. Off-road environment is unstructured, dominated by natural objects, lacking artificial features, and its terrain conditions are various and complex. One of the fundamental techniques of an off-road robot is to detect safe region (hereinafter called *off-road*) to traverse, which has also been termed as traversable surface [4], drivable corridor [5], etc., in literature. Comparing with the roads in structured environments, where functional attributes are clearly defined by artificial features such as pavement, barrier and markings, off-roads are ill-defined [6].

Early methods of off-road detection are usually developed by assuming color, texture, boundary of the target, where rule-based methods of extracting vanishing point and subsequently road boundaries [7] [8], and segmentation-based methods of extracting continuous regions based on certain road models are developed [9] [10]. These methods are called *coarse-grained* ones as the problem is formulated as a binary classification, e.g. labelling each image pixel to *road* or *non-road*. As illustrated in Fig. 1(b-c), such methods may fail to detect any region to traverse at tough terrains or extract too wide region that lacks efficiency in prompting the best choice at open area. Moreover, the mechanical performance of off-road robots can be very different, leading to different definitions and selections of safe regions to traverse. Understanding scenes with find-grained labels is needed for off-road robots [11]. On the other hand, deep learning methods have been studied in recent years [12]. Semantic segmentation using deep learning techniques infers scenes at

*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.

pixel- or point-levels [13], where large-scale datasets such as Cityscapes [14], SemanticKITTI [15] with fine-grained labels and massive annotations are needed. There is no such dataset at off-road scenes. How to define fine-grained labels to achieve meaningful scene understanding for a robot to traverse off-road is still an open question.

This research proposes a contrastive learning method to achieve fine-grained semantic segmentation and mapping of off-road scenes as shown in Fig. 1(d-e). It is difficult to define fine-grained categories that are generalized at diverse off-road scenes and it is further hard for a human operator to assign fine-grained labels to each image pixel, where the definitions could be very ambiguous at natural scenes. However, it is not difficult for a human operator to annotate images by sparse anchor patches as illustrated in Fig. 2 to indicate the regions with different semantic attributes on their traversability. Inspired by impressive progress and promising results of contrastive learning [16] [17] [18], this research learns a feature representation to discriminate regions with different semantic attributes using contrastive learning, which is used to develop a method of find-grained semantic segmentation and mapping for off-road applications. Experiments are conducted using our off-road dataset including over $12,000$ frames and 3 subsets with diverse off-road scenes. With less than 100 training frames, our method can achieve nearly 90% anchor accuracy on test sets in cross-scene validation. The fine-grained results are further demonstrated by additional LiDAR data analysis. The experimental results prove the validity of our proposed method, and show its potential in applications for off-road environments.

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. Rule/Segmentation-based Methods

Rule/segmentation-based methods are mainly developed by assuming color, texture, boundary of the target region, and these researches are mostly coarse-grained understanding that formulate the problem as a binary classification. They can be broadly divided into rule-based and segmentation-based methods.

Some rule-based methods utilize global priors like vanishing point [7] [8], which primarily depend on edge cues to obtain road area. The others assume the road region as geometric triangular [19] or trapezoidal [20] shape.

Segmentation-based methods formulate the problem as pixel-level segmentation tasks. Some studies [21] assume the region at bottom of images as road data or collect vehicle trajectories as drivable area [22], then label similar regions as road. Other methods [9] [10] depend on fixed road models and make use of hybrid features to extract continuous region.

B. Deep Learning Methods

Benefit by developments of deep networks [13] and large-scale datasets with fine-grained labels like Cityscapes [14] and SemanticKITTI [15], deep learning methods are able to get fine-grained semantic segmentation or maps. However, most existing datasets and studies are designed for urban scenes, and research in off-road environments is still limited.

Due to the lack of datasets, studies for off-road scenes attempt several ways to reduce the demand of fine-annotated data, such as weakly and semi-supervised learning [23] [24], and transfer learning [25] [26]. One mainstream idea is automatically generating training data from other sensor modalities, such as 3D LiDAR data [27] [24], audio features [28] and force–torque signals [11]. Another idea is transfer knowledge of deep networks from existing urban datasets [25] or synthetic data [26] to off-road environment. Nevertheless, transferred models still need some fine-annotated data for finetuning, and the performance is limited by domain gaps. Meanwhile, labels from other modalities or synthetic data are too limited to support fine-grained semantic segmentation and mapping.

C. Contrastive Learning

Recent progress in contrastive learning [16] [17] [18] demonstrates that discriminative representations could be learnt through a self-supervised pipeline, by contrasting positive and negative samples. Various sample definitions make contrastive learning suitable for diverse domains like natural language [16] and images [29]. Zhao et al. [30] introduce contrastive learning to semantic segmentation task, but rely on pixel-level labeled data for initial contrastive learning and generating pseudo labels for unlabeled images.

Inspired by the promising results of contrastive learning, but different from settings in [30], this work only relies on a small number of sparse anchor annotations without pixel-level labels to learn feature representations to discriminate regions with different semantic attributes, which is further used to develop a method of fine-grained semantic segmentation and mapping.

III. Methodology

A. Problem Formulation

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 normalized 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

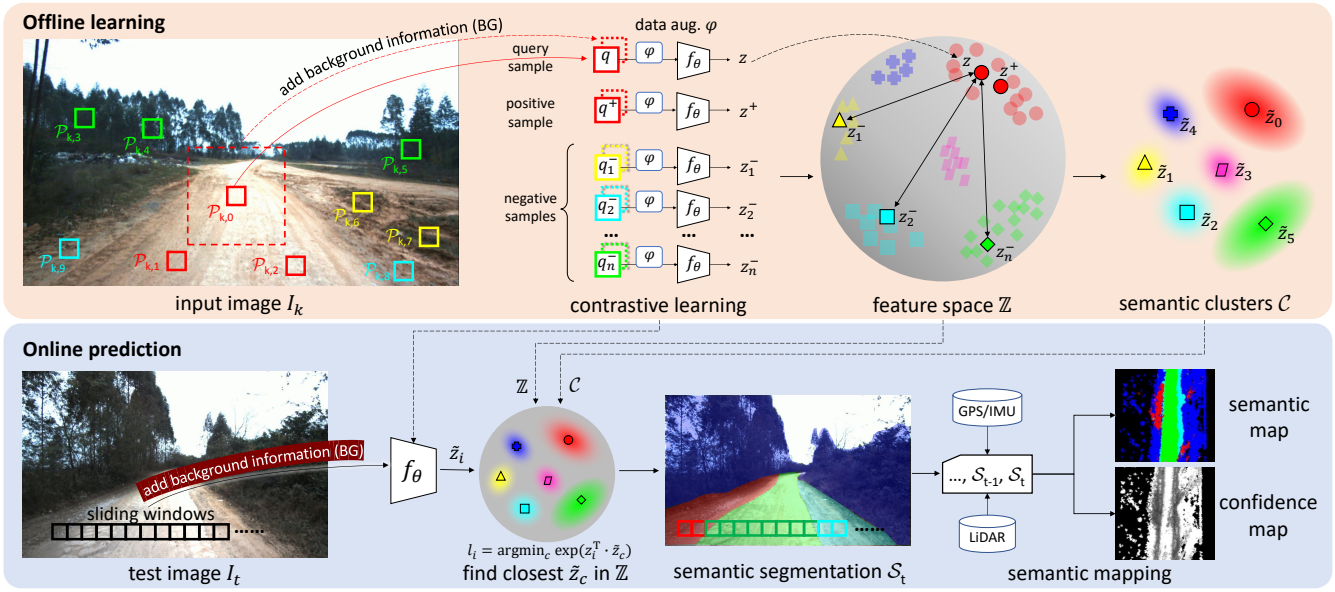

Fig. 2: The proposed pipeline for fine-grained off-road semantic segmentation and mapping via contrastive learning.

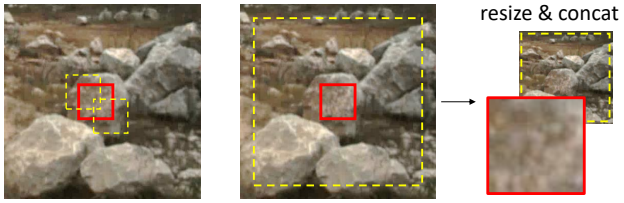

Fig. 3: Illustration of (a) neighborhood sampling strategy, and (b) how to add background information with the foreground image patch.

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 and mapping 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, then 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 in 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 randomly selected 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 randomly clipped neighborhood patches should have the center points 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, sample data is composed in the same way for query, positive and 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 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 [31] is used to model f_θ, which converts the 6-channel tensor of a query, positive or negative sample to a normalized low-dimensional feature vector $z \in \mathbb{Z}^D$. Contrastive learning is used to find θ in 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 [32] is implemented:

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

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 feature representations with reasonable memory consumption. In practice, unlike the typical

contrastive learning studies [33] using a 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 offline learning and online prediction, while the latter is composed of further two steps: semantic segmentation of single images and semantic mapping using multiple images.

1) Offline 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 normalized 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 $\mathcal{I}_t$, semantic segmentation $\mathcal{S}_t$ 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 normalized 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 as the semantic label that has the best match on its feature vector, i.e. $l_i = \arg\min_c exp(z_i^T \cdot \tilde{z}_c)$.

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 to indicate the confidence of 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.

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

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

Fig. 4: Typical scenes in three sub-datasets, which include diverse off-road scenarios.

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 and annotated 973 anchors for training, which only account for about 10% of total frames. All frames are evaluated in testing, while frames with human annotations are taken into account in quantitative evaluations.

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.

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 patch $\mathcal{P}_{k,i}$ and $\mathcal{P}_{k,j}$ are positive samples of each other and classified into the same semantic cluster, then $Pos(i, j) = 1$. Otherwise, if they are not classified to the same semantic 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 called *anchor accuracy* to evaluate how well the clustering results fit human annotations:

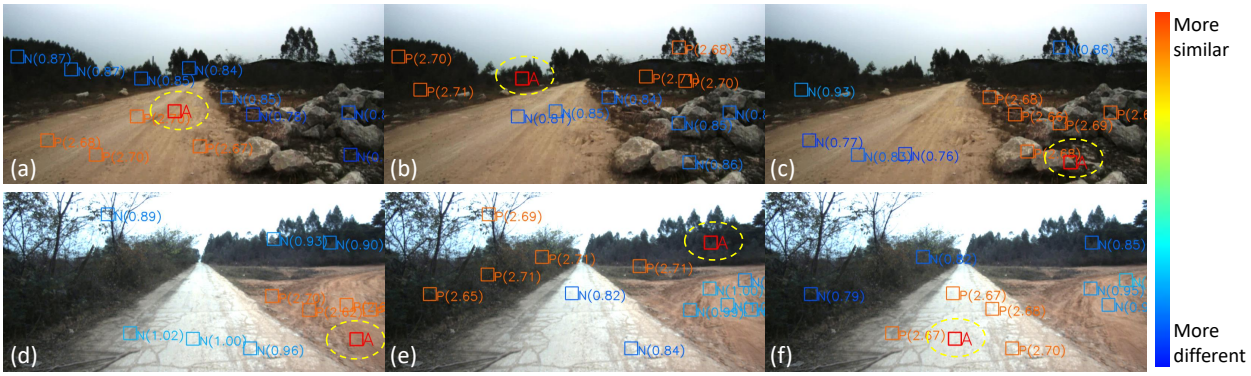

Fig. 5: Visualization of feature distances of query anchor (A) to its positive (P) and negative (N) samples. Query anchors are circled by yellow rings. Numbers in parentheses measure samples' similarity to the query anchor.

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**: pipeline without data augmentation or background information; **base_DA**: use data augmentation, without background information; **BG192/256/320**: complete pipeline with different background size; $\bar{\mathcal{R}}$: average anchor accuracy $\mathcal{R}$.

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

Essentially, it can be seen as Rand Index [34], 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 exploring the effects of different module settings. (3) fine-grained semantic segmentation and mapping, make concrete case study and statistical analysis from additional LiDAR data to show the validity of our fine-grained results.

1) Feature Distance Measurement: The core module in our proposed pipeline is the feature encoder f_θ, which projects a high-dimensional image patch to a low-dimensional feature vector in space $\mathbb{Z}$. Its purpose is to make feature distance closer between similar image patches, while farther between different image patches. In Fig. 5, we visualize some concrete cases. In all images, the query anchors are circled by yellow rings, while the other anchor patches are randomly sampled and colorized by their feature distances to the query anchor. For example, in Fig. 3(a), the query anchor is located on the earth road. We can find that patches on the earth road are closer to red, and other

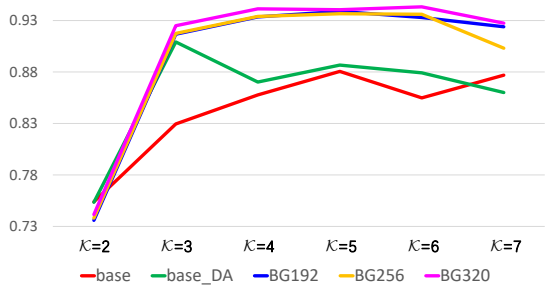

Fig. 6: Average $\mathcal{R}$ of models under different clustering number $\mathcal{K}$.

patches located on different semantic areas are generally blue, which indicates the farther distance to the query anchor. The feature distance distribution is in accord with human-annotated anchor labels. Similar situations are general in other images. 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 evaluations of the proposed method, we make cross-validation on models with different settings, and the statistics are shown in Table II. The table cells are colorized along column data when training and testing on different subsets. The last column lists the average anchor accuracy $\bar{\mathcal{R}}$ on test sets (subsets different with the training one). It is obvious that *BG320* has the best performance on test sets, and all three models with background information have $\mathcal{R}$ over 0.85

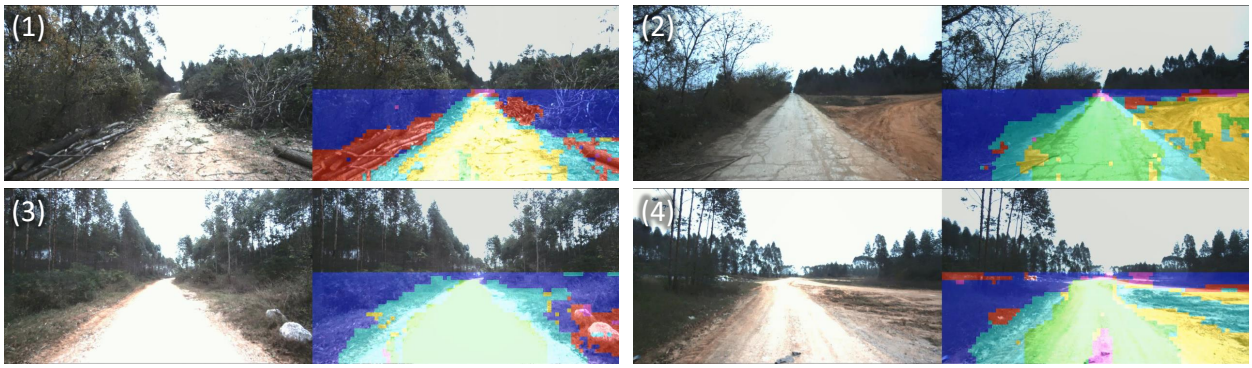

Fig. 7: Some results of fine-grained semantic segmentation.

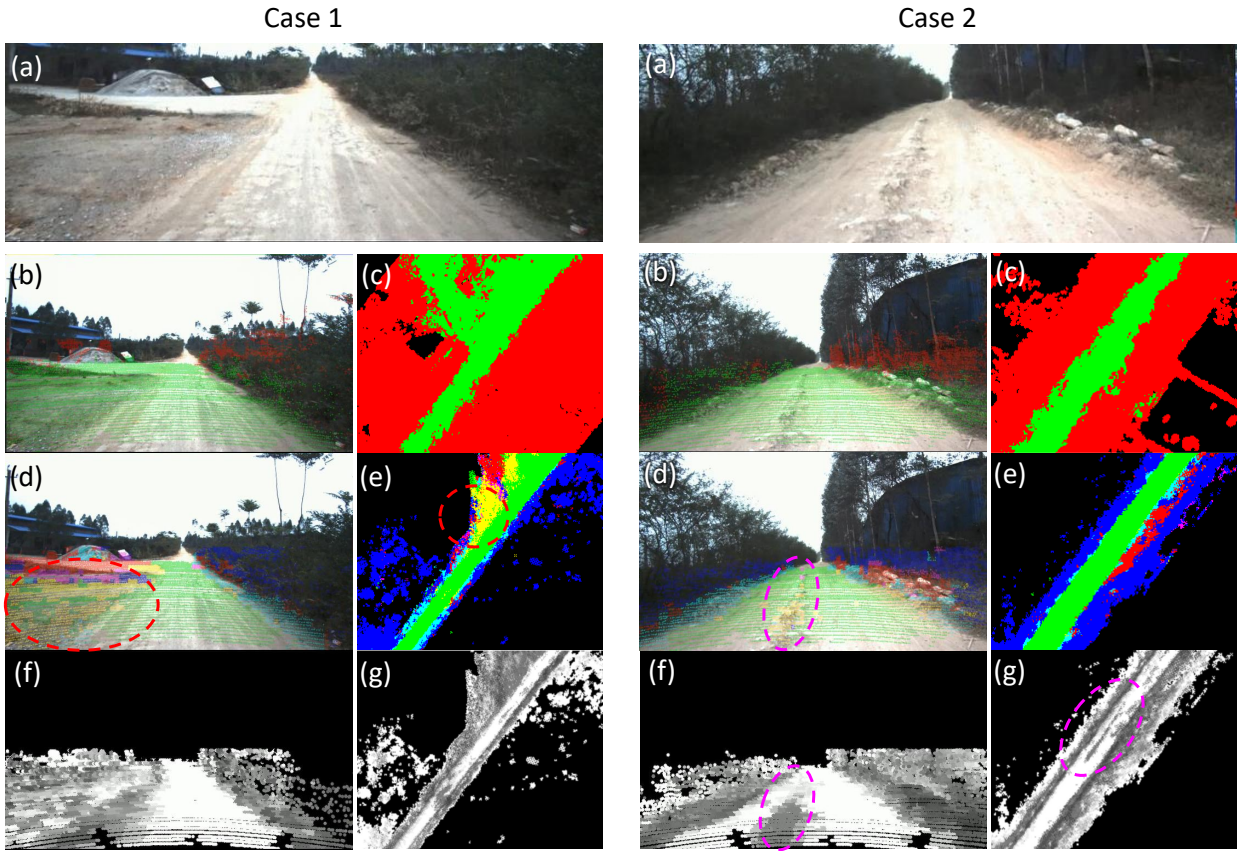

Fig. 8: Case study of fine-grained semantic map and confidence map, compared with coarse-grained results. (a) scene image. (b) coarse-grained segmentation (binary classification). (c) coarse-grained semantic map (bird's-eye-view). (d) fine-grained semantic segmentation (projected on point clouds). (e) fine-grained semantic map (bird's-eye-view). (f) confidence map projected to camera-view, whiter pixels indicate higher confidence. (g) bird's-eye-view confidence map.

among all conditions, which demonstrates the robustness of our proposed method.

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

By the way, above experiments uniformly used clustering number $\mathcal{K} = 6$. How does $\mathcal{K}$ 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 approximately order the same as Table II. Therefore, we choose $\mathcal{K} = 6$ as other experiments' setting to balance the fine-grained demand and model performance.

3) Fine-Grained Semantic Segmentation and Mapping: Due to the absence of ground truth for fine-grained off-road semantic segmentation task, we next analyze the validity of our fine-grained results through case studies of semantic segmentation and mapping. Besides, we compare traversability cost of different categories by additional LiDAR data

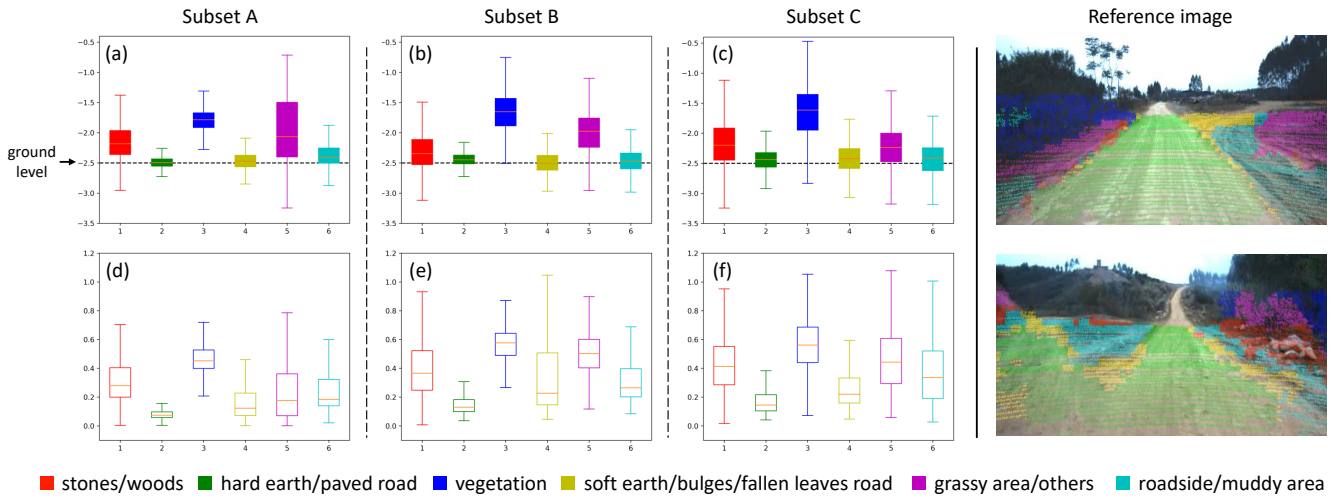

Fig. 9: Traversability analysis of semantic clusters based on 3D LiDAR data. (a-c) boxplots of points average height, indicate height distribution of different categories. (d-f) boxplots of points height variance, indicate surface flatness and traversability cost.

analysis. 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 the bottom half of the image is predicted for simplicity. The semantic labels are not pre-designed, but we can find their intrinsic 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 soft earth, red pixels are stones or woods, etc. Different colors represent different clusters, and they can generally distinguish diverse semantic meanings.

To show the consistency of fine-grained predictions on continuous video frames and get overall scene understanding, we make semantic maps and confidence maps as described in Sec. III-C.3. As shown in Fig. 8(case 1), our fine-grained predictions can label the roadside area (yellow) with higher traversability cost than middle road (green), while the traditional coarse-grained result is unable to distinguish them. In Fig. 8(case 2), let us pay attention to bulges 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 Fig. 8(f)(g) can be helpful to distinguish this subtle traversability difference, where the bumpy area are darker than other flat roads.

More than case studies, a statistical traversability analysis is provided in Fig. 9, which is based on 3D point clouds with labels projected from image semantic segmentation. By the way, semantic meanings of the color table are not pre-defined but concluded from our model's predictions. In Fig. 9(a-c), it is obvious that green, yellow, and cyan mainly distribute around the ground level, which are three primary road types. Furthermore, in Fig. 9(d-f), we can find their different traversability cost, where the green box have the narrowest variance distribution, corresponding to the most easily passable paved road and hard earth. Yellow and cyan boxes are longer, indicating bumpier road surface. They are mainly soft earth, bulges 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 and mapping.

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

In this paper, we propose a fine-grained off-road semantic segmentation and mapping method based on contrastive learning techniques. The proposed method can learn discriminative feature representations with only a small number of sparse anchor annotations, with no demand of laborious pixel-level labels, which significantly reduces labeling difficulties and cost in off-road environment. We examine our fine-grained results across diverse off-road scenes and prove its validity through additional LiDAR data analysis. Future work will be addressed on improving its computational efficiency by temporal and spatial consistency, while exploring OOD sample detection mechanism and incremental learning ability for long-term deployment on off-road robots.

Fig. 10: A challenging case: when meeting unseen semantic categories.

REFERENCES

[1] Di Feng, Christian Haase-Schuetz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *Transactions on Intelligent Transportation Systems*, 2020.

[2] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius Brito Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago Meireles Paixao, Filipe Mutz, et al. Self-driving cars: A survey. *Expert Systems with Applications*, page 113816, 2020.

[3] Mennatullah Siam, Sara Elkerdawy, Martin Jagersand, and Senthil Yogamani. Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges. In *International Conference on Intelligent Transportation Systems*, pages 1–8. IEEE, 2017.

[4] Shengyan Zhou, Junqiang Xi, Matthew W McDaniel, Takayuki Nishihata, Phil Salesses, and Karl Iagnemma. Self-supervised learning to visually detect terrain surfaces for autonomous robots operating in forested terrain. *Journal of Field Robotics*, 29(2):277–297, 2012.

[5] Ara V Nefian and Gary R Bradski. Detection of drivable corridors for off-road autonomous navigation. In *International Conference on Image Processing*, pages 3025–3028. IEEE, 2006.

[6] Marek Ososinski and Frédéric Labrosse. Automatic driving on ill-defined roads: An adaptive, shape-constrained, color-based method. *Journal of Field Robotics*, 32(4):504–533, 2015.

[7] Hui Kong, Jean-Yves Audibert, and Jean Ponce. Vanishing point detection for road detection. In *Conference on Computer Vision and Pattern Recognition*, pages 96–103. IEEE, 2009.

[8] Jinjin Shi, Jinxiang Wang, and Fangfa Fu. Fast and robust vanishing point detection for unstructured road following. *Transactions on Intelligent Transportation Systems*, 17(4):970–979, 2015.

[9] Yaniv Alon, Andras Ferencz, and Amnon Shashua. Off-road path following using region classification and geometric projection constraints. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 689–696. IEEE, 2006.

[10] Jian Wang, Zhong Ji, and Yu-Ting Su. Unstructured road detection using hybrid features. In *International Conference on Machine Learning and Cybernetics*, volume 1, pages 482–486. IEEE, 2009.

[11] Lorenz Wellhausen, Alexey Dosovitskiy, René Ranftl, Krzysztof Walas, Cesar Cadena, and Marco Hutter. Where should I walk? predicting terrain properties from images via self-supervised learning. *Robotics and Automation Letters*, 4(2):1509–1516, 2019.

[12] Thiago Rateke, Karla A Justen, Vito F Chiarella, Antonio C Sobieranski, Eros Comunello, and Aldo Von Wangenheim. Passive vision region-based road detection: A literature review. *ACM Computing Surveys*, 52(2):1–34, 2019.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The CityScapes dataset for semantic urban scene understanding. In *Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[15] 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 *International Conference on Computer Vision*, pages 9297–9307. IEEE, 2019.

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[19] Shengyan Zhou and Karl Iagnemma. Self-supervised learning method for unstructured road detection using fuzzy support vector machines. In *International Conference on Intelligent Robots and Systems*, pages 1183–1189. IEEE, 2010.

[20] Hong Jeong, Yuns Oh, Jeong-Ho Park, BS Koo, and Sang Wook Lee. Vision-based adaptive and recursive tracking of unpaved roads. *Pattern Recognition Letters*, 23(1-3):73–82, 2002.

[21] Keyu Lu, Jian Li, Xiangjing An, and Hangen He. A hierarchical approach for road detection. In *International Conference on Robotics and Automation*, pages 517–522. IEEE, 2014.

[22] Jilin Mei, Yufeng Yu, Huijing Zhao, and Hongbin Zha. Scene-adaptive off-road detection using a monocular camera. *Transactions on Intelligent Transportation Systems*, 19(1):242–253, 2017.

[23] Benjamin Suger, Bastian Steder, and Wolfram Burgard. Traversability analysis for mobile robots in outdoor environments: A semi-supervised learning approach based on 3D-LiDAR data. In *International Conference on Robotics and Automation*, pages 3941–3946. IEEE, 2015.

[24] Biao Gao, Anran Xu, Yancheng Pan, Xijun Zhao, Wen Yao, and Huijing Zhao. Off-road drivable area extraction using 3D LiDAR data. In *Intelligent Vehicles Symposium*, pages 1505–1511. IEEE, 2019.

[25] Christopher J Holder, Toby P Breckon, and Xiong Wei. From on-road to off: transfer learning within a deep convolutional neural network for segmentation and classification of off-road scenes. In *European Conference on Computer Vision*, pages 149–162. Springer, 2016.

[26] Suvash Sharma, John E Ball, Bo Tang, Daniel W Carruth, Matthew Doude, and Muhammad Aminul Islam. Semantic segmentation with transfer learning for off-road autonomous driving. *Sensors*, 19(11):2577, 2019.

[27] Li Tang, Xiaqing Ding, Huan Yin, Yue Wang, and Rong Xiong. From one to many: Unsupervised traversable area segmentation in off-road environment. In *International Conference on Robotics and Biomimetics*, pages 787–792. IEEE, 2017.

[28] Jannik Zürn, Wolfram Burgard, and Abhinav Valada. Self-supervised visual terrain classification from unsupervised acoustic feature learning. *Transactions on Robotics*, 2020.

[29] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[30] Xiangyun Zhao, Raviteja Vemulapalli, Philip Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label-efficient semantic segmentation. *arXiv preprint arXiv:2012.06985*, 2020.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[33] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Conference on Computer Vision and Pattern Recognition*, June 2018.

[34] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.