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

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

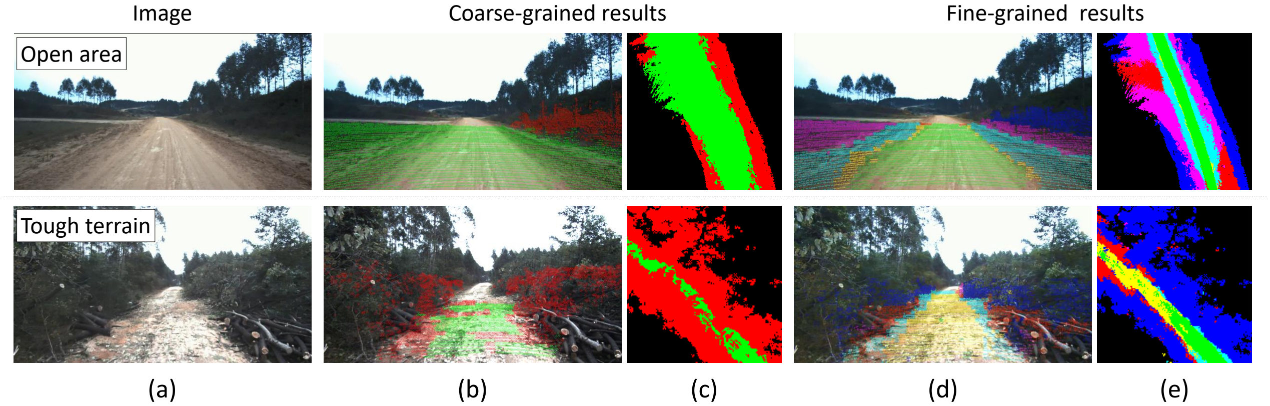


Fig. 1: The significance of fine-grained semantic segmentation and mapping in off-road environment, where coarse-grained results can hardly adapt diverse scenes. (a) scene image. (b) coarse-grained semantic segmentation (binary classification). (c) coarse-grained semantic map. (d) fine-grained semantic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

***Abstract*— Road detection or traversability analysis has been a key technique for a mobile robot to traverse complex off-road scenes. The problem has been mainly formulated in early works as a binary classification one, e.g. associating pixels with road or non-road labels. Whereas understanding scenes with fine-grained labels are needed for off-road robots, as scenes are very diverse, and the various mechanical performance of off-road robots may lead to different definitions of safe regions to traverse. How to define and annotate 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 based method. With a set of human-annotated anchor patches, a feature representation is learned to discriminate regions with different traversability, a method of fine-grained semantic segmentation and mapping is subsequently developed for off-road scene understanding. Experiments are conducted on a dataset of three driving segments that represent very diverse off-road scenes. An *anchor accuracy* of 89.8% is achieved by evaluating the matching with human-annotated image patches in cross-scene validation. Examined by associated 3D LiDAR data, the fine-grained segments of visual images are demonstrated to have different levels of toughness and terrain elevation, which represents their semantical meaningfulness. The resultant maps contain both fine-grained labels and confidence values, providing rich information to support a robot traversing complex off-road scenes.**

I. Introduction

Mobile robotic and autonomous driving techniques have been witnessed with 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 environments [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. The 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 regions (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, boundaries 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. labeling 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 regions that lack efficiency in promoting 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 fine-grained labels is needed

*This work is supported in part by the National Natural Science Foundation of China under Grant 61973004.

[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.

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 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 and annotate 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 fine-grained semantic segmentation and mapping for off-road applications. An off-road dataset is developed containing over 12000 image frames of three driving segments that represent very diverse off-road scenes. With no more than 100 training frames in all experimental settings, the test results in cross-scene validation show an 89.8% *anchor accuracy*, which is a new metric introduced to evaluate the matching with human-annotated image patches. Examined by additionally measured 3D LiDAR data, it is found that the fine-grained segments of visual images are semantically meaningful to represent different levels of toughness and terrain elevation. The resultant maps contain both fine-grained labels and confidence values, providing rich information to support a robot traversing complex 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, boundaries of the target region, and these researches are mostly coarse-grained understanding that formulates 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 roads. Other methods [9][10] depend on fixed road models and make use of hybrid features to extract continuous regions.

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 for 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 [24][27], audio features [28] and force-torque signals [11]. Another idea is to transfer knowledge of deep networks from existing urban datasets [25] or synthetic data [26] to off-road environments. 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 learned 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 similarity $sim(p_i, p_j) = exp(z_i^T \cdot z_j)$ to measure the similarity of two image patches via their

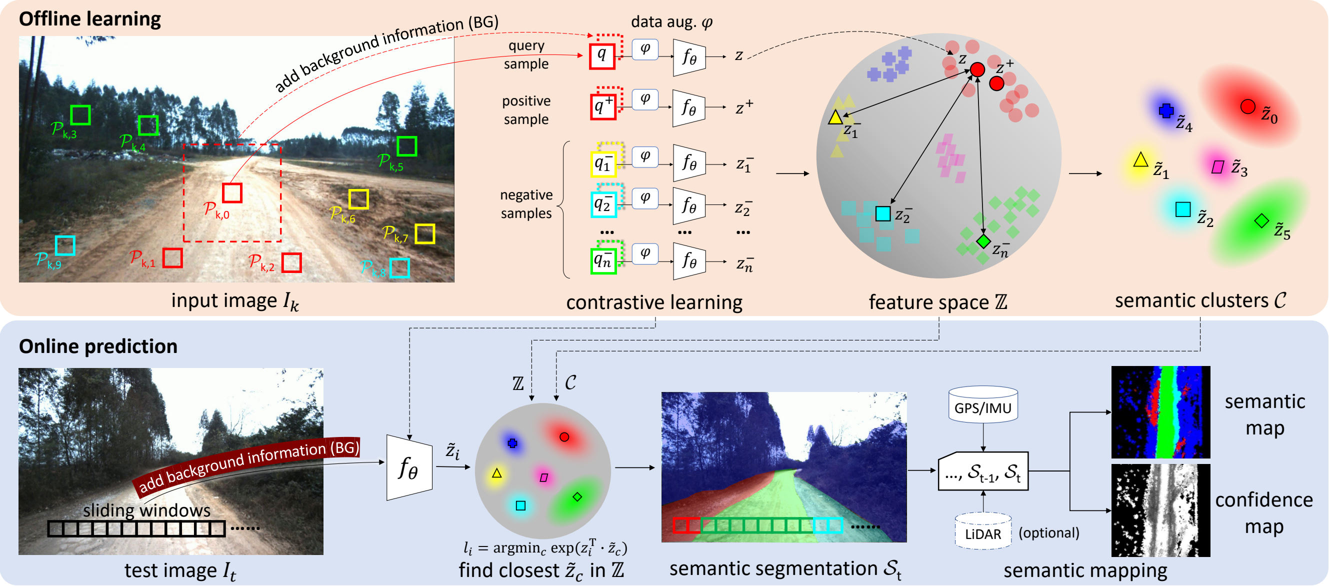
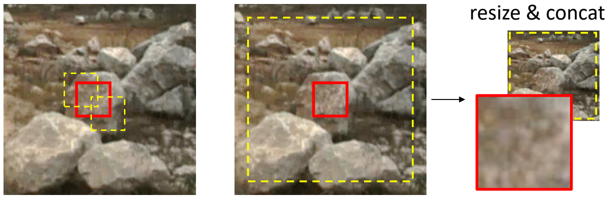


Fig. 2: The proposed pipeline for fine-grained off-road semantic segmentation and mapping via contrastive learning.



(a) neighborhood sampling

(b) add background information

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

low-dimensional feature vectors. Therefore, given an anchor patch $\mathcal{P}_{k,i}$, its similarity to another anchor patch $\mathcal{P}_{k,j}$, i.e. $sim(p_{k,i}, p_{k,j})$, should be higher if they share the same label $a_{k,i} = a_{k,j}$, whereas lower 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 z 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. Assume that an off-road scene is spatially continuous, i.e. nearby regions could be semantically similar. An anchor patch p is first randomly selected from $\{\mathcal{P}_{k,i}^+\}$, where an image patch is randomly clipped from p's 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: As shown in Fig. 3(b), sample data contains foreground and background image patches to describe both local and global features. The foreground is image patch p, while the 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. With an image patch p, sample data is composed in the same way for the query, positive and negative samples.

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 greyscale, 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, i.e. 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 similarity of the z are high if they share the same labels, whereas low for those differences. Following the principle of previous contrastive learning studies [18], 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 workflow 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 learned to convert each image patch to a normalized low-dimensional vector $z \in \mathbb{Z}^D$ in the space of $\mathbb{Z}^D$, the image patches with the same labels are projected close, whereas far for the others.

The z of the anchor patches are then clustered by the 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, clustering number $\mathcal{K}$ is a hyper-parameter, which decides the granularity of semantic segmentation.

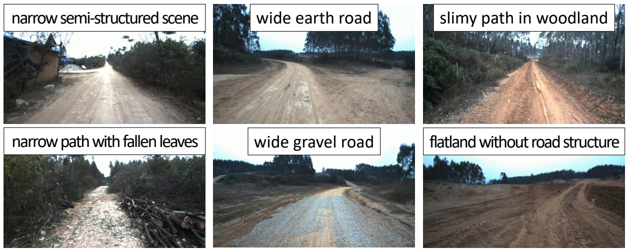
2) Semantic Segmentation: Given the current image $\mathcal{I}_t$, semantic segmentation $\mathcal{S}_t$ is conducted by generating image patches using sliding windows and predicting a semantic label for each image patch. Given an image patch p_i, a semantic label is predicted as follows. A 6-channel 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 l_i 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 the 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 at the frame, a horizontal plane is drawn at the ground level and tessellated into regular grids. The pixel labels of the current image can be projected onto the grids with the camera's calibration parameters. Besides, the pixel labels of early frames can also be projected onto the grids with additionally the vehicle's localization data at each frame. If a 3D LiDAR is associated, the projection can be conducted via LiDAR points, where the up and down of off-road terrain can be taken into calculation. Since a single grid can have multiple label predictions, let $\sigma^c_{x,y}$ denote the counts of predicting label c of grid (x, y), the semantic label $l_{x,y}$ = $\text{argmax}_c(\sigma^c_{x,y})$ 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 $\max(\sigma^c_{x,y})/\sum \sigma^c_{x,y}$, 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


narrow semi-structured scene
wide earth road
slimy path in woodland
narrow path 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 sub-datasets, which include diverse off-road scenarios.

IV. Experimental Results

A. Experimental Data

An off-road dataset is developed to evaluate the proposed method. The dataset is collected by an instrumented vehicle with a front-view monocular RGB camera, a GPS/IMU suite, and a 3D LiDAR. In this work, we use visual images for semantic segmentation, while GPS/IMU provides 6 DoF poses of the ego vehicle, which is used for mapping. 3D LiDAR is mainly used to examine the semantic meaning of the fine-grained segmentation, while it is also used in this research in projecting visual labels to a horizontal plane so as to generate a more accurate map by considering the up and down of off-road terrain.

As shown in Table I, the dataset contains over 12000 image frames of three driving segments that represent very diverse off-road scenes. Take subset A as an example, 50 image frames are randomly selected, which account for 10% of the total 5064 frames of subset A. 973 anchor patches are annotated on the 50 image frames by a human operator, which are used in training. The rest image frames of subset A are used in testing, and the image frames of subset B and C are also used to test the model trained on subset A in the experiment of cross-scene validation. Experiments on subset B and C are conducted in the same way to examine the results of semantic segmentation. To this end, image frames are used for testing and image patches are manually annotated in the same way as the anchors, which are used as *ground truth* to evaluate the accuracy of the results.

The three subsets contain driving data at very different off-road scenes. As illustrated in Fig. 4, the scenarios in

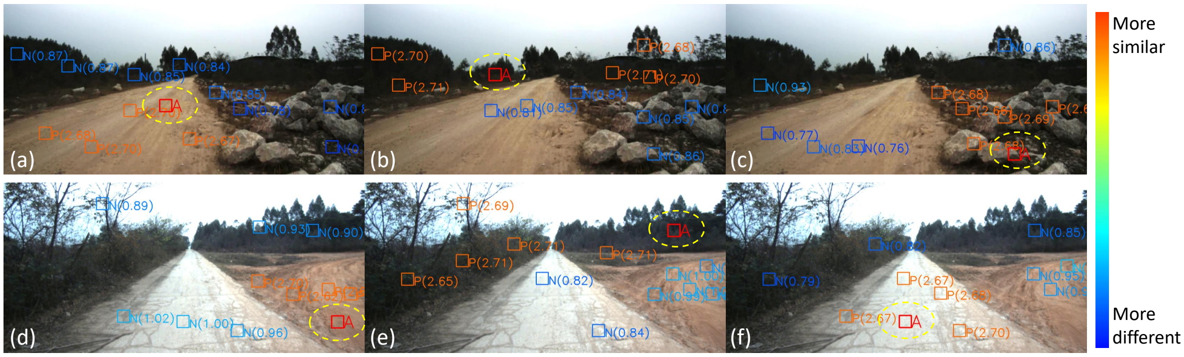


Fig. 5: Visualization of feature similarity of query anchor (A) to its positive (P) and negative (N) samples. Query anchors are circled by yellow rings. Numbers in parentheses measure samples' exponential cosine similarity to the query anchor.

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

model	data aug.	BG size	train on subset A test on A	 B	 C	train on subset B test on B	 A	 C	train on subset C test on C	 A	 B	$\bar{\mathcal{R}}$ on test sets
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 $\mathcal{R}$.

subset A are mostly narrow roads with bushes aside, subset B are relatively wide scenes, and subset C includes diverse scenarios like slimy paths in woodland and flatland without road structure. In the 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:

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

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 similarity measurement, explore the validity of feature encoder and similarity measurement learned by contrastive learning. (2) cross-scene 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.

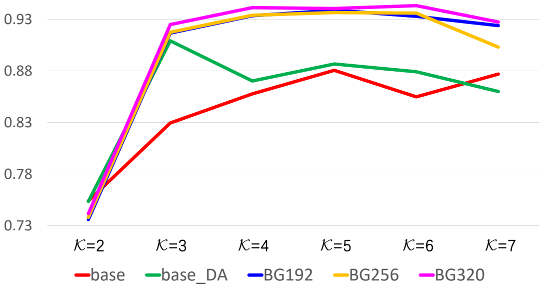


Fig. 6: Average $\mathcal{R}$ of models under different clustering number $\mathcal{K}$.

1) Feature Similarity Measurement: The feature encoder f_θ aims to make similar image patches closer and different image patches farther in feature space. In Fig. 5, we visualize some concrete cases of the learned similarity measurement $sim(p_i, p_j) = exp(z_i^T \cdot z_j)$. In all images, the query anchors are circled by yellow rings, while the other anchor patches are randomly sampled and colorized by their exponential cosine similarity 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 patches located on different semantic areas are

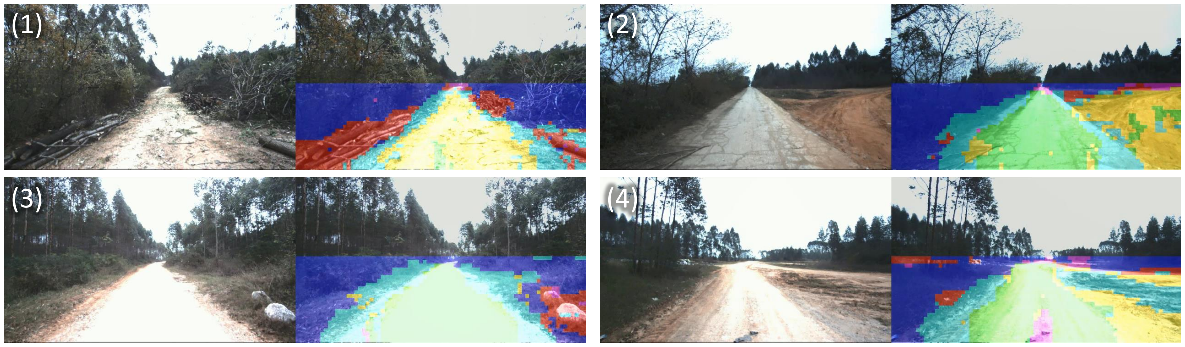


Fig. 7: Some results of fine-grained semantic segmentation.

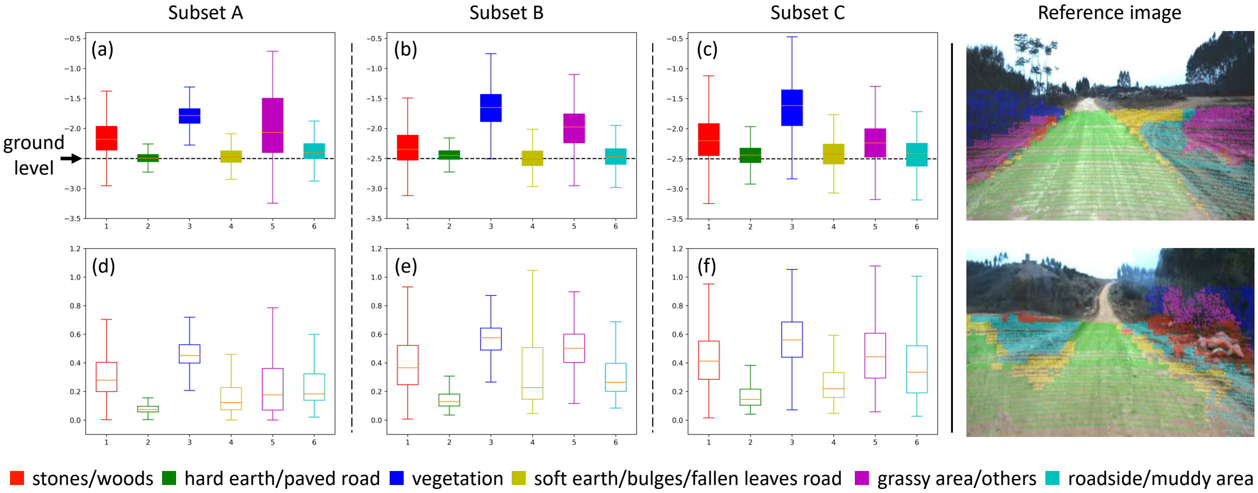


Fig. 8: Traversability analysis of semantic clusters by 3D LiDAR data. (a-c) boxplots of points average height, indicate height distribution of different categories. (d-f) boxplots of points height variance, indicate surface flatness and traversability cost.

generally blue, which indicates the lower similarity to the query anchor. The feature similarity distribution is in accord with the semantic differences. Similar situations are general in other images. **As a result, the feature encoder and similarity measurement learned by contrastive learning are able to distinguish similar or different image patches.**

2) Cross-Scene Validation and Ablation Study: For comprehensive evaluations of the proposed method, we make cross-scene 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 (the 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 85% among all conditions, which demonstrates the robustness of our proposed method. The data augmentation and background information can both increase models' performance, while the latter makes more contribution. Increasing background size could sightly improve overall performance, but is not obvious in all situations.

To explore how clustering number $\mathcal{K}$ affects models' performance, an ablation study is made as 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 orders the same as Table II. Therefore, we choose $\mathcal{K} = 6$ as other experiments' setting to balance the fine-grained demand and model performance.

In summary, the proposed method achieves 89.8% average anchor accuracy in cross-scene validation, and the performance is stable with regard to different clustering numbers, which demonstrates the robustness and generalization of our method.

3) Fine-Grained Semantic Segmentation and Mapping: Due to the absence of pixel-level annotations for the task, we next demonstrate the validity of our fine-grained results through case studies and additional LiDAR data 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. Because this work focuses on off-road traversability analysis, so 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

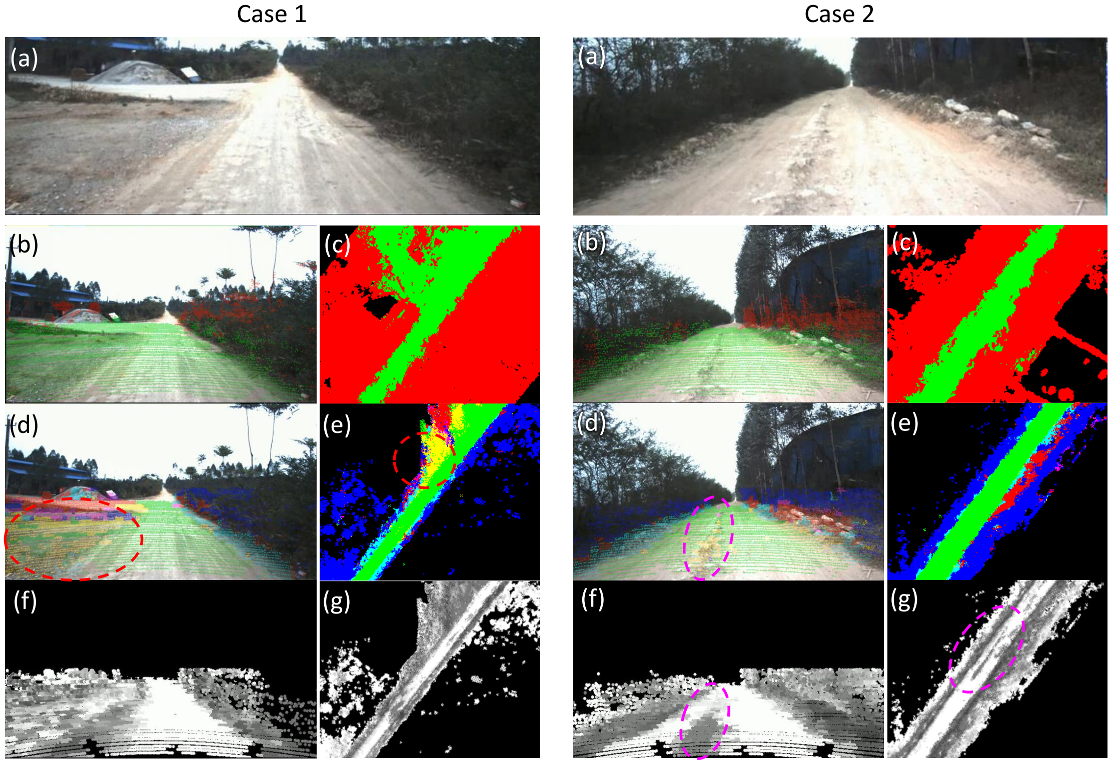


Fig. 9: Case study of fine-grained semantic map and confidence map, compared with coarse-grained results. (a) scene image. (b) coarse-grained segmentation (binary classification). (c) coarse-grained semantic map (bird's-eye-view). (d) fine-grained semantic segmentation (projected on point clouds). (e) 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.

road, *blue* pixels are vegetation, *yellow* pixels are road with fallen leaves or soft earth, *red* pixels are stones or woods, etc. Different clusters can generally distinguish diverse semantic meanings.

Statistical analysis is provided in Fig. 8, which is based on 3D LiDAR data with labels projected from image semantic segmentation. In Fig. 8(a-c), three categories' terrain elevation (*green*, *yellow*, and *cyan*) mainly distribute around the ground level, which are three primary road types. Furthermore, from Fig. 8(d-f), we can find their different traversability cost. The *green* boxes have the narrowest variance distribution, corresponding to the most easily passable paved road and hard earth. The *yellow* and *cyan* boxes are longer, indicating bumpier road surface. The *blue* boxes indicate bushes and trees, with the highest elevation and traversability cost. **In a word, examined by associated 3D LiDAR data, the fine-grained segments of images are proved to have different levels of toughness and terrain elevation, which represents their semantical meaningfulness.**

The semantic maps and confidence maps are shown in Fig. 9. In case 1, our fine-grained predictions label the roadside area (*yellow*) with higher traversability cost than middle road (*green*). In case 2, bulges in the middle of the road are separated in the single frame segmentation, but not stable enough to obtain majority votes in the semantic map. The confidence map can be helpful to distinguish this subtle traversability difference, where the bumpy area is darker than other flat roads. **Therefore, the resultant fine-grained semantic maps and confidence maps can provide rich information for robots to traverse in complex off-road scenes.**

D. Challenges

Currently, there are still some challenges with the proposed method. Firstly, the current pipeline to obtain dense predictions has a high computational cost. Predicting one image patch takes about 30 ms on an NVIDIA TITAN X. Although parallel computing helps to predict more patches at one time, the repetitive computation of overlapped patches 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 semantic clusters, which may lead to confused predictions as Fig. 10(c). To minimize labor costs, the OOD sample detection and incremental training mechanism deserve to be explored in our future works.

V. CONCLUSIONS

In this paper, we propose a contrastive learning based method for off-road fine-grained semantic segmentation and mapping. With a set of human-annotated anchor patches, a feature representation is learned to discriminate regions with different traversability. After that, the fine-grained semantic

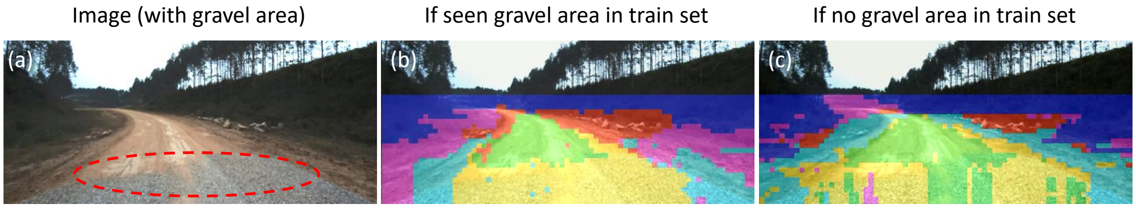


Image (with gravel area)
If seen gravel area in train set
If no gravel area in train set
(a)
(b)
(c)

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

segmentation and mapping pipeline is proposed for off-road scene understanding. For the experimental study of our method, we develop an off-road dataset with three driving segments that represent very diverse off-road scenes. The proposed method achieves 89.8% anchor accuracy in cross-scene validation by evaluating the matching with human-annotated image patches. Examined by associated 3D LiDAR data, the fine-grained segments of visual images are demonstrated to have different levels of toughness and terrain elevation, which represents their semantical meaningfulness. The resultant maps contain both fine-grained labels and confidence values, providing rich information to support a robot traversing complex off-road scenes. Future work will be addressed on improving the computational efficiency by temporal and spatial consistency, and exploring OOD sample detection mechanism and incremental learning ability for long-term deployment on off-road robots.

References

[1] D. Feng *et al.*, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *Transactions on Intelligent Transportation Systems*, 2020.

[2] C. Badue *et al.*, "Self-driving cars: A survey," *Expert Systems with Applications*, p. 113 816, 2020.

[3] M. Siam *et al.*, "Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges," in *International Conference on Intelligent Transportation Systems*, IEEE, 2017, pp. 1–8.

[4] S. Zhou *et al.*, "Self-supervised learning to visually detect terrain surfaces for autonomous robots operating in forested terrain," *Journal of Field Robotics*, vol. 29, no. 2, pp. 277–297, 2012.

[5] A. V. Nefian *et al.*, "Detection of drivable corridors for off-road autonomous navigation," in *International Conference on Image Processing*, IEEE, 2006, pp. 3025–3028.

[6] M. Ososinski *et al.*, "Automatic driving on ill-defined roads: An adaptive, shape-constrained, color-based method," *Journal of Field Robotics*, vol. 32, no. 4, pp. 504–533, 2015.

[7] H. Kong *et al.*, "Vanishing point detection for road detection," in *Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 96–103.

[8] J. Shi *et al.*, "Fast and robust vanishing point detection for unstructured road following," *Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 970–979, 2015.

[9] Y. Alon *et al.*, "Off-road path following using region classification and geometric projection constraints," in *Conference on Computer Vision and Pattern Recognition*, IEEE, vol. 1, 2006, pp. 689–696.

[10] J. Wang *et al.*, "Unstructured road detection using hybrid features," in *International Conference on Machine Learning and Cybernetics*, IEEE, vol. 1, 2009, pp. 482–486.

[11] L. Wellhausen *et al.*, "Where should I walk? predicting terrain properties from images via self-supervised learning," *Robotics and Automation Letters*, vol. 4, no. 2, pp. 1509–1516, 2019.

[12] T. Rateke *et al.*, "Passive vision region-based road detection: A literature review," *ACM Computing Surveys*, vol. 52, no. 2, pp. 1–34, 2019.

[13] J. Long *et al.*, "Fully convolutional networks for semantic segmentation," in *Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[14] M. Cordts *et al.*, "The CityScapes dataset for semantic urban scene understanding," in *Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[15] J. Behley *et al.*, "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *International Conference on Computer Vision*, IEEE, 2019, pp. 9297–9307.

[16] A. v. d. Oord *et al.*, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[17] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, PMLR, 2020, pp. 1597–1607.

[18] K. He *et al.*, "Momentum contrast for unsupervised visual representation learning," in *Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[19] S. Zhou *et al.*, "Self-supervised learning method for unstructured road detection using fuzzy support vector machines," in *International Conference on Intelligent Robots and Systems*, IEEE, 2010, pp. 1183–1189.

[20] H. Jeong *et al.*, "Vision-based adaptive and recursive tracking of unpaved roads," *Pattern Recognition Letters*, vol. 23, no. 1-3, pp. 73–82, 2002.

[21] K. Lu *et al.*, "A hierarchical approach for road detection," in *International Conference on Robotics and Automation*, IEEE, 2014, pp. 517–522.

[22] J. Mei *et al.*, "Scene-adaptive off-road detection using a monocular camera," *Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 242–253, 2017.

[23] B. Suger *et al.*, "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*, IEEE, 2015, pp. 3941–3946.

[24] B. Gao *et al.*, "Off-road drivable area extraction using 3D LiDAR data," in *Intelligent Vehicles Symposium*, IEEE, 2019, pp. 1505–1511.

[25] C. J. Holder *et al.*, "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*, Springer, 2016, pp. 149–162.

[26] S. Sharma *et al.*, "Semantic segmentation with transfer learning for off-road autonomous driving," *Sensors*, vol. 19, no. 11, p. 2577, 2019.

[27] L. Tang *et al.*, "From one to many: Unsupervised traversable area segmentation in off-road environment," in *International Conference on Robotics and Biomimetics*, IEEE, 2017, pp. 787–792.

[28] J. Zürn *et al.*, "Self-supervised visual terrain classification from unsupervised acoustic feature learning," *Transactions on Robotics*, 2020.

[29] Y. Tian *et al.*, "Contrastive multiview coding," *arXiv preprint arXiv:1906.05849*, 2019.

[30] X. Zhao *et al.*, "Contrastive learning for label-efficient semantic segmentation," *arXiv preprint arXiv:2012.06985*, 2020.

[31] A. Krizhevsky *et al.*, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[32] A. v. d. Oord *et al.*, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[33] Z. Wu *et al.*, "Unsupervised feature learning via non-parametric instance discrimination," in *Conference on Computer Vision and Pattern Recognition*, Jun. 2018.

[34] 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.