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

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

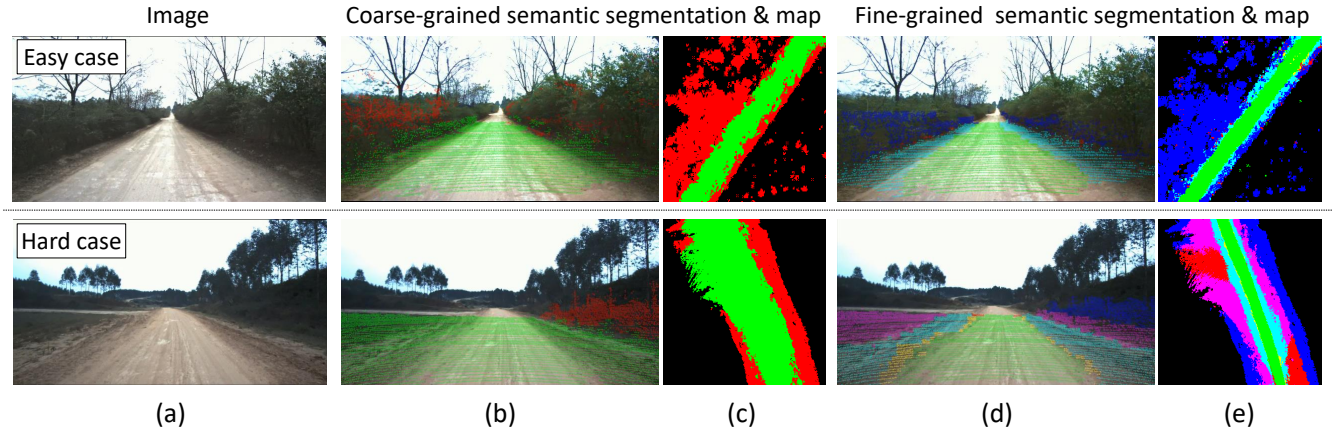

Fig. 1: The significance of fine-grained semantic segmentation and mapping in off-road environment, where coarse-grained results can hardly adapt diverse scenes with unified threshold. (a) scene image. (b) coarse-grained semantic segmentation (binary classification). (c) coarse-grained semantic map (bird's-eye-view).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

***Abstract*— Fine-grained off-road scene understanding for traversability analysis is of great importance for self-driving platforms with different trafficability. However, off-road scenes are unstructured and full of area with semantic ambiguity. It causes difficulties for problem definition and fine-grained semantic labeling, which further leads to the absence of large-scale pixel-level datasets for off-road semantic segmentation and mapping. To address these challenges, this work proposes a fine-grained off-road semantic segmentation and mapping method. Depending on a small number of low-cost anchor annotations without concrete semantic labels, the proposed contrastive learning pipeline can learn discriminative feature representations for semantic segmentation, with no need for pixel-level human annotations. The experiments on diverse off-road datasets illustrate the reasonability of our framework and the validity of our fine-grained semantic segmentation and mapping results.**

I. Introduction

Recent years, a considerable development has grown up around the theme of intelligent vehicles [1]. Driving scene understanding plays a critical role in the preconditions of decision planning and self-driving [2]. For the moment, a considerable amount of studies for urban scenes has been published [3]. 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 [4]. 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 [5][6] 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), with uniform parameters, the drivable area is acceptable for the easy case, but too wide to provide human preferred area of the hard case, which is powerless for further meticulous motion planning. Expected fine-grained semantic segmentation and mapping as shown in Fig. 1(d-e) could extract different semantic regions in both scenes, which improves the traversability analysis ability for off-road driving. Deep learning methods [3] 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 [7][8][9] have been made in computer

*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.

vision researches, which can learn effective feature representations through unsupervised pipeline. It has been proved to support fine-grained image classification tasks in large-scale datasets like ImageNet [10]. The contrastive learning pipeline can learn discriminative representations without relying on fine semantic annotations, which is suitable for ambiguous definition and annotating challenges in off-road scenes.

To solve the challenges in fine-grained off-road semantic segmentation and mapping, i.e. ambiguous problem definition and the scarcity of elaborate annotated datasets, this work proposed a fine-grained off-road semantic segmentation and mapping 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, 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 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. Traditional Methods

Traditional methods mainly make use of geometrical and visual features to extract semantic meanings. Meanwhile, previous research for off-road environments are mostly coarse-grained understanding that focuses on road detection, which can be broadly divided into rule-base and segmentation-based methods.

Some rule-based methods utilize global priors like vanishing point [11][12], which primarily depend on edge cues. Another part rule-based methods assume the road region as geometric triangular [13] or trapezoidal [14] shapes.

Segmentation-based methods formulate the problem as pixel level segmentation tasks. Some studies [15] assume the region at bottom of images as road data or collect vehicle trajectories as drivable area [16]. Other methods [17][18] depend on fixed models like AdaBoost, and make use of various texture [17] or hybrid features [18] for segmentation.

B. Deep Learning Methods

Benefit by developments of deep networks [19] and large-scale annotated datasets like Cityscapes [20], deep learning methods are able to get fine-grained semantic segmentation or maps for autonomous driving. However, most studies face to urban scenes. Due to the lack of large-scale datasets and ambiguous problem definition, research in off-road environment [21] is still limited.

Existing studies for off-road scenes usually attempt to reduce the demand of fine-annotated data, such as weakly and semi-supervised learning [22][23], and transfer learning [24][25]. One mainstream idea is automatically generating training data from other sensor modalities. For example, label traversable area from human-driven trajectories and obstacles from 3D LiDAR data[26][23], classify different terrains by audio features [27] or force–torque signals [4]. Another idea is transfer knowledge of deep networks from existing urban datasets [24] or synthetic data [25] to off-road environment. Nevertheless, transferred models still need some fine-annotated data for finetuning, and the performance is limited by domain gaps. Meanwhile, auto-generated labels and synthetic data could be more accessible, but their label granularity can hardly support fine-grained semantic segmentation and mapping.

C. Contrastive Learning

Recent progress in contrastive learning [7][8][9] demonstrates that discriminative representations could be learnt through unsupervised pipeline, by contrasting positive and negative samples. Various sample definitions make contrastive learning suitable for diverse domains like natural language [7] and images [28]. Zhao et al. [29] introduce contrastive learning to semantic segmentation task, but rely on pixel-level labeled data for initial contrastive learning, then generate pseudo labels for unlabeled images.

Different from settings in [29], this work only rely on a small number of low-cost anchor annotations without concrete semantic labels for contrastive learning. It can alleviate labeling difficulties in ambiguous off-road scenes, while learning discriminative feature representations without large quantities of pixel-level human annotations.

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 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

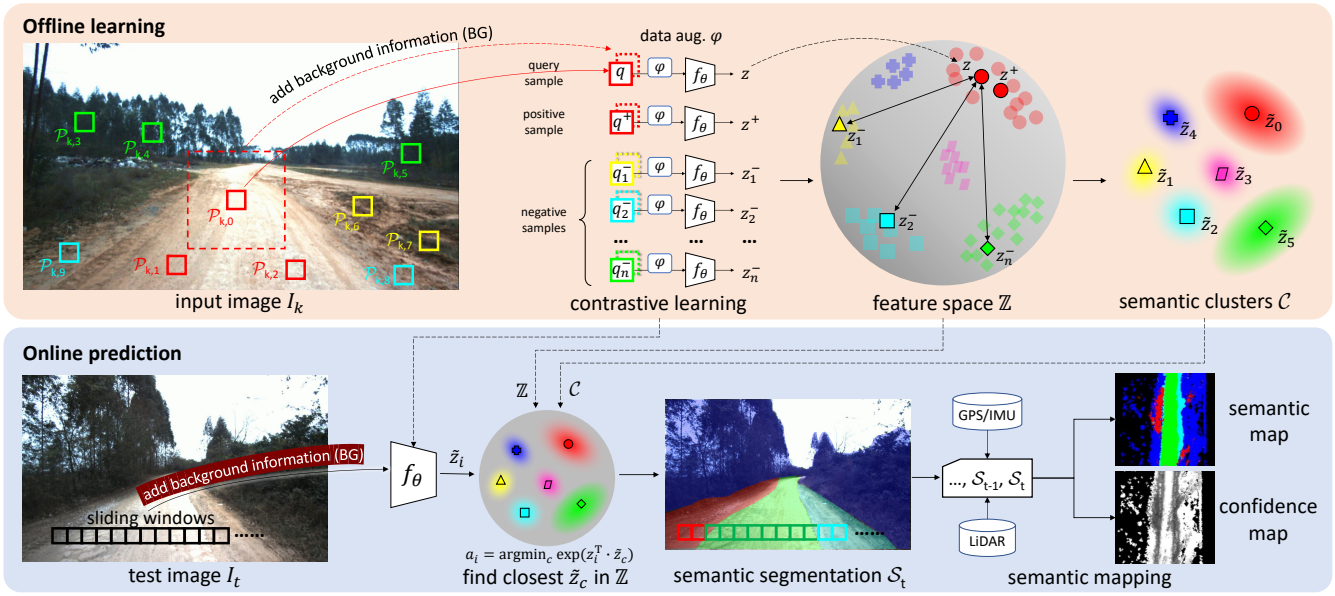

Fig. 2: The proposed pipeline for fine-grained off-road semantic segmentation and mapping via contrastive learning.

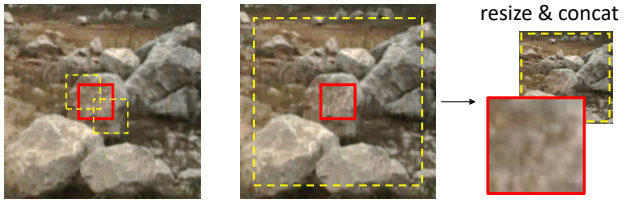

Fig. 3: Illustration of (a) neighborhood sampling strategy, and (b) how to add background information with the foreground image patch.

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 th 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 [30] 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 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 [31] 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 the feature representations with reasonable memory consumption. In practice, unlike the typical contrastive learning studies [32] 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 offline 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 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 $\mathcal{I}$, semantic segmentation $\mathcal{S}$ 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 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. All frames are evaluated when testing, while frames with human annotations are taken into account in quantitative evaluation.

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 datasets, which represent different off-road environments.

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 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:

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

Essentially, it can be seen as Rand Index [33], 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

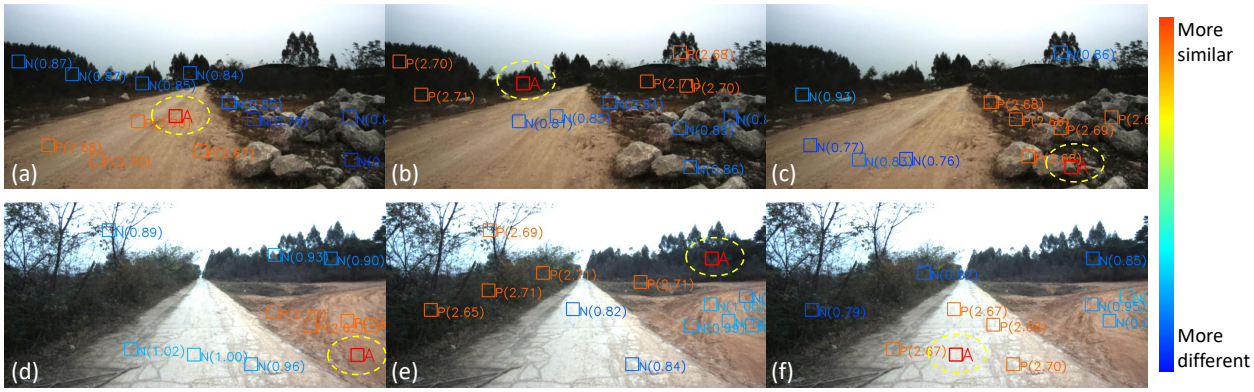

Fig. 5: Visualization of feature distances of query anchor (A) to its positive (P) and negative (N) samples. Query anchors are circled by yellow rings. P/N is according to human annotated anchor labels, and numbers in parentheses measure samples' similarity to the query anchor.

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.

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, make concrete case study and statistical results to show our 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

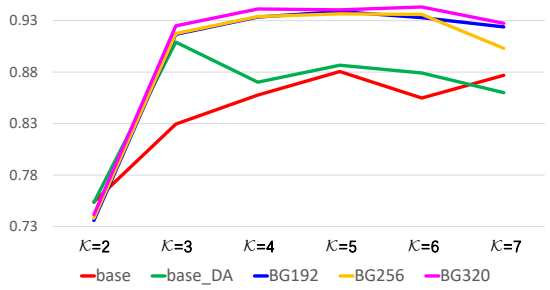

Fig. 6: Average $\mathcal{R}$ of models under different clustering number $\mathcal{K}$.

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

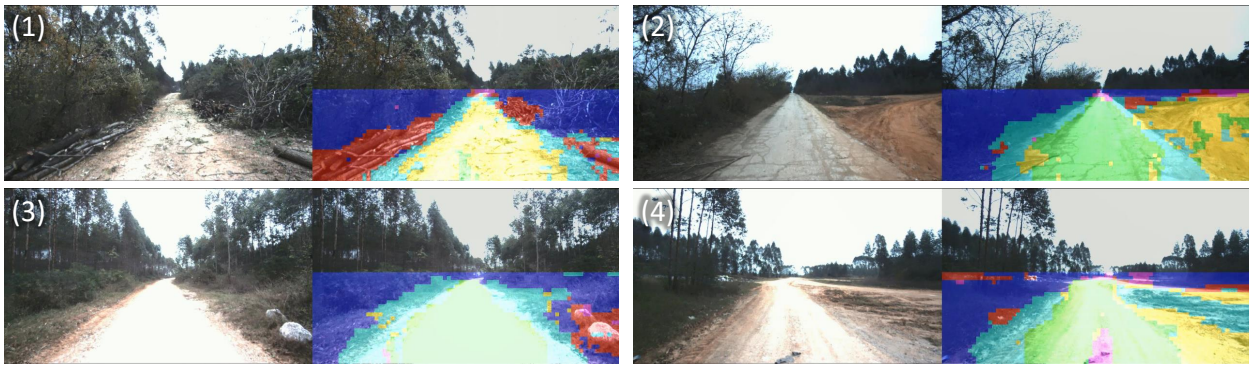

Fig. 7: Some cases of fine-grained semantic segmentation.

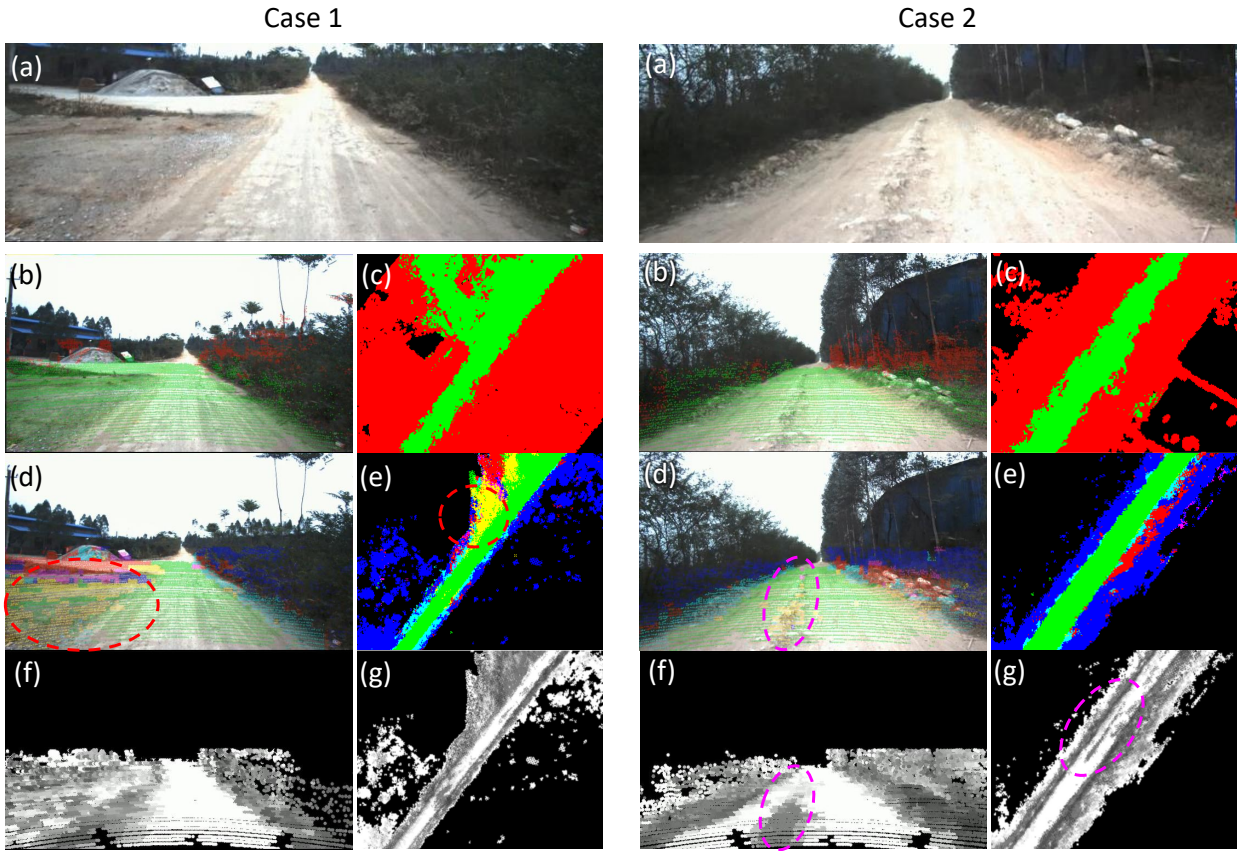

Fig. 8: Case study of fine-grained semantic map and confidence map, compared with coarse-grained road extraction results. (a) video image. (b) coarse-grained segmentation. (c) coarse-grained bird's-eye-view semantic map. (d) fine-grained semantic segmentation projected by point clouds. (e) fine-grained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

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 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

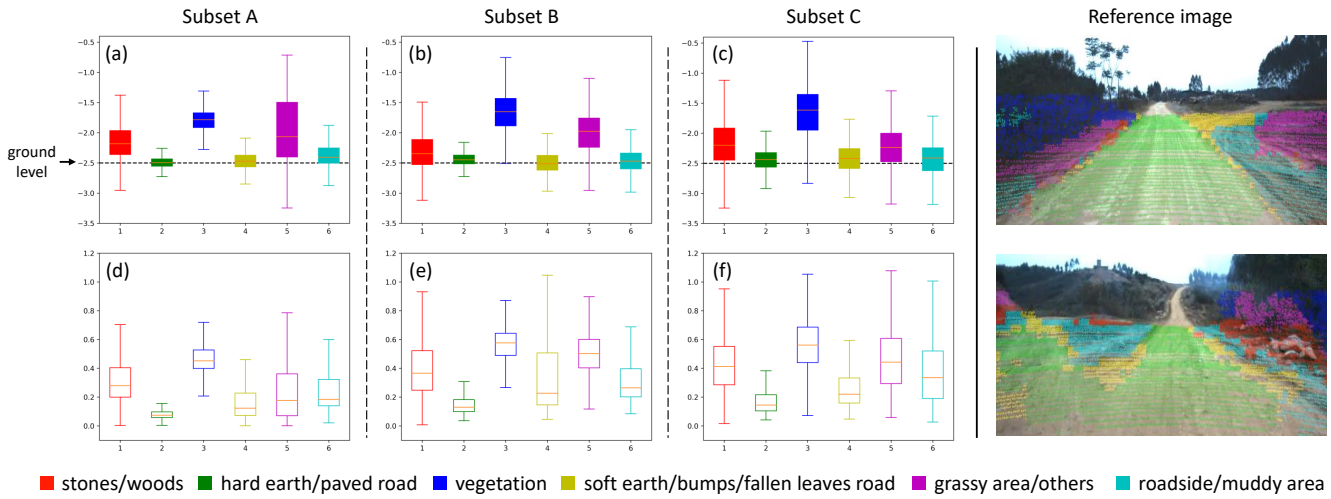

Fig. 9: Traversability analysis of semantic clusters based on point clouds. (a-c) boxplots of points average height, indicate height distribution of different categories. (d-f) boxplots of points height variance, indicate surface flatness and traversability cost.

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 show 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 Fig. 8(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 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, 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 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 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 significantly solve the challenges in off-road semantic segmentation and mapping tasks, i.e. ambiguous problem definition, difficulties in labeling data, and the scarcity of large-scale human annotations. We design a contrastive learning pipeline, which can automatically learn feature representations with only a small number of low-cost anchor annotations, then predict fine-grained semantic segmentation results with no demand of laborious pixel-level annotations. The case study and cross validation on diverse off-road subsets prove the validity of our fine-grained semantic segmentation and mapping results. Future work will be addressed on improving computational efficiency by temporal and spatial consistency, while exploring OOD sample detection and incremental learning mechanism for long-term deployment on off-road self-driving platforms.

Fig. 10: 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] 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.

[5] Sebastian Thrun, Mike Montemerlo, Hendrik Dahlkamp, David Stavens, Andrei Aron, James Diebel, Philip Fong, John Gale, Morgan Halpenny, Gabriel Hoffmann, et al. Stanley: The robot that won the DARPA Grand Challenge. *Journal of Field Robotics*, 23(9):661–692, 2006.

[6] Wende Zhang. LiDAR-based road and road-edge detection. In *Intelligent Vehicles Symposium*, pages 845–848. IEEE, 2010.

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

[8] 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.

[9] 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.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Daniel Maturana, Po-Wei Chou, Masashi Uenoyama, and Sebastian Scherer. Real-time semantic mapping for autonomous off-road navigation. In *Field and Service Robotics*, pages 335–350. Springer, 2018.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

[29] 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.

[30] 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.

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

[32] 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.

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