

An Active and Contrastive Learning Framework for Fine-Grained Off-Road Semantic Segmentation

Biao Gao¹, *Member, IEEE*, Xijun Zhao², *Member, IEEE*, Huijing Zhao¹, *Member, IEEE*,

Abstract—Off-road semantic segmentation with fine-grained labels is necessary for autonomous vehicles to understand driving scenes, as the coarse-grained road detection can not satisfy off-road vehicles with various mechanical properties. Fine-grained semantic segmentation in off-road scenes usually has no unified category definition due to ambiguous nature environments, and the cost of pixel-wise labeling is extremely high. Furthermore, semantic properties of off-road scenes can be very changeable due to various precipitations, temperature, defoliation, etc. To address these challenges, this research proposes an active and contrastive learning-based method that does not rely on pixel-wise labels, but only on patch-based weak annotations for model learning. There is no need for predefined semantic categories, the contrastive learning-based feature representation and adaptive clustering will discover the category model from scene data. In order to actively adapt to new scenes, a risk evaluation method is proposed to discover and select hard frames with high-risk predictions for supplemental labeling, so as to update the model efficiently. Experiments conducted on our self-developed off-road dataset and DeepScene dataset demonstrate that fine-grained semantic segmentation can be learned with only dozens of weakly labeled frames, and the model can efficiently adapt across scenes by weak supervision, while achieving almost the same level of performance as typical fully supervised baselines.

Index Terms—off-road, semantic segmentation, active learning, contrastive learning

I. INTRODUCTION

SEMANTIC segmentation is one of the key perception techniques for an autonomous driving agent to navigate safely and smoothly in complex environments [1]. There has been a large body of studies on semantic segmentation, while most of them are addressed in structured urban scenes [2]. Such scenes are composed of many man-made objects such as paved roads, lane markings, traffic signals, buildings, etc. These objects belong to semantically interpretable categories and their data have fairly clear boundaries. Despite the large needs of fine-grained perception for autonomous driving at off-road scenes [3]–[5], semantic segmentation in such scenes has far less been studied. Off-road scenes are composed of natural objects in various shapes and of indistinct semantic category, diverse terrain surfaces, and changed topographical conditions [6]. Semantic segmentation in such scenes remains an open challenge.

According to the granularity of scene understanding, the methods of off-road semantic segmentation can be divided into

*This work was supported in part by the National Natural Science Foundation of China under Grant 61973004 and High-performance Computing Platform of Peking University.

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

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

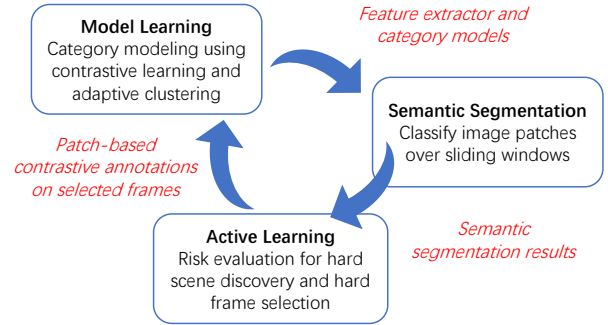


Fig. 1. The proposed active and contrastive learning framework for fine-grained off-road semantic segmentation.

two groups: *coarse-grained* and *fine-grained* ones. Coarse-grained methods formulate the problem as a binary [7][8] or triple classification [9], or road detection by labeling each pixel as road or non-road. They usually rely on prior rules, such as vanishing points [10] or certain road models [11]. However, the mechanical properties of autonomous vehicles are various, requiring a fine-grained understanding of terrain properties that can lead to a measure for the difficulty of terrain negotiation [12]. With the development of deep learning techniques in recent years, many deep semantic segmentation models are developed [13]. These models can be learned end-to-end on large-scale datasets with pixel-wise annotation, while both the size and diversity of the datasets are crucial to the model’s performance [14]. Most of the open datasets in this scope describe urban scenes, such as Cityscapes [15] and SemanticKITTI [16]. The few off-road ones [17][18] are of limited size and different definitions of semantic categories. Fine-grained semantic segmentation in off-road scenes faces the following challenges: 1) There has been no unified category definition in nature scenes due to the diverse objects and ambiguous semantic interpretability; 2) Pixel-wise annotation of fine-grained labels is very hard because a large part of the pixels could suffer from severe semantic ambiguity, which makes manual annotation almost impractical; 3) Off-road scenes can be very different, and even in the same location, semantic properties can be changed greatly due to precipitations, temperature, defoliation, etc.

Facing the challenges, this research proposes a framework of fine-grained off-road semantic segmentation based on active and contrastive learning as illustrated in Fig. 1 and detailed in Fig. 2. It has the following features:

- No pixel-wise annotated datasets: a patch-based annotation is devised to generate contrastive pairs of image patches that have different semantic attributes for weak

supervision, and subsequently a sliding-window-based semantic segmentation is exploited;

- No predefined semantic categories: semantic categories are discovered and modeled on scene data by using contrastive learning for feature representation and adaptive clustering for category modeling;
- Adaptation to new scenes actively: a risk evaluation method is developed to discover scenes where the model results suffer from high-risk and the hard frames where the model is the most uncertain, so as to update the model actively and efficiently.

An off-road dataset is developed in this research containing three subsets of different scenes with a total of 8000 image frames. Extensive experiments are conducted to examine the performance of both key modules and the system flow of passive-active learning on both the self-developed and DeepScene [17] datasets. Experimental results show that a model of fine-grained off-road semantic segmentation can be learned through weak supervision on dozens of annotated image frames, when performance degradation is detected, active learning can be automatically triggered to update the model with additional annotations on no more than 40 selected hard frames. DeepScene experiments show that the proposed weakly supervised method achieves almost the same level of performance as the typical fully-supervised ones.

This paper is organized as follows. Related works are introduced in Section. II. Section. III describes the proposed contrastive and active learning method. In Section. IV, the experimental design are illustrated. Section. V shows experimental results. Finally, we provide the conclusion in Section. VI.

II. RELATED WORKS

A. Off-Road Semantic Segmentation

Early researches were mainly coarse-grained, which are usually formulated as a binary classification problem. These methods depend on priors like vanishing point [10], vehicle trajectories [6] or assume the road area as geometric shapes [19][20], or utilize fixed road models [11][8].

Benefiting from advances in deep learning, stronger feature representation results has led to fine-grained semantic segmentation capabilities. Rothrock et al. [21] firstly implemented FCN [22] for terrain classification of the Martian surface. After that, more studies [23]–[26] focused on off-road scenes have been developed. Some of them deal with the challenges from various illumination and visual features in off-road scenes by combining multi-modal information with RGB images, such as stereo camera [27], NIR [17] and LiDAR [28][29]. Due to the limitation of public datasets and the difficulty of off-road labeling, several studies tried to reduce the demand for fine-annotated data by transfer learning [30][31] from urban or synthetic data. For autonomous platforms with multiple sensors, weak supervision can be obtained from other modalities, such as LiDAR [32][9], audio features [33] and force-torque signals [12]. However, these automatically-generated labels are usually limited to certain specific categories and cannot meet fine-grained requirements. In addition, few studies

consider how to effectively adapt the model to the new off-road environments while avoiding pixel-wise annotations and network architecture changes.

B. Contrastive Learning

Contrastive learning has proven its promising ability to learn discriminative feature representations through a self-supervised pipeline by comparing positive and negative samples. This idea has been widely used in many fields such as natural language processing [34] and typical visual tasks [35]. These methods usually treat each instance and its augmented version as a positive pair, while other randomly selected instances are regarded as negative samples. In this setting, a large number of negative samples are required to ensure the effectiveness of the learned feature representation. The memory bank is usually used to store the features of the training data [35]–[37].

Recent studies [38] proposed a supervised contrastive learning framework for the image classification task, which uses class labels to generate positive and negative samples. This idea has been extended to pixel-level semantic segmentation tasks by [39][40]. However, pixel-wise annotations are extremely rare and expensive in off-road scenes. Different from the setting of [39][40], this paper does not require pixel-level labels but only use a few sparse image patch-based annotations to distinguish similar or different regions of an image, and the features learned by contrastive learning are further used to generate fine-grained semantic segmentation results.

C. Active Learning

The core idea of active learning [41] is to let the trained model actively select the hardest or most informative samples to query manual annotations. According to [42], researches can be categorized by different query strategies: uncertainty-based approach [43]–[45], diversity-based approach [46]–[48], and expected model change [49]–[51]. The uncertainty-based methods select samples with the highest uncertainty, which can be estimated by entropy [43][44] or softmax probability from deep neural networks [45]. MC Dropout [46] and ensemble methods [52] can be used to improve uncertainty estimation. Diversity-based methods [46]–[48] tend to select samples in accord with input distributions, but it may lead to increased labeling costs. The methods of expected model change [49]–[51] predict the influence of an unlabeled example on future model decisions, and choose the examples leading to more expected model change as an informative sample.

The active learning approaches for semantic segmentation usually use regions or entire images as the sampling unit. Region-level methods [53][54] rely on a pre-segmentation to retrieve super-pixels, but due to insufficient or over-segmentation, the segmentation algorithm may not be able to separate appropriate semantic regions for labeling. Image-level methods [55][56] use the entire image as the sampling unit. [57] incorporates the semantic difficulty to measure the informativeness and select samples at the image level. In this work, we also select samples at the image level. However, we do not require high-cost pixel-wise labels, but only query patch-wise weak annotations.

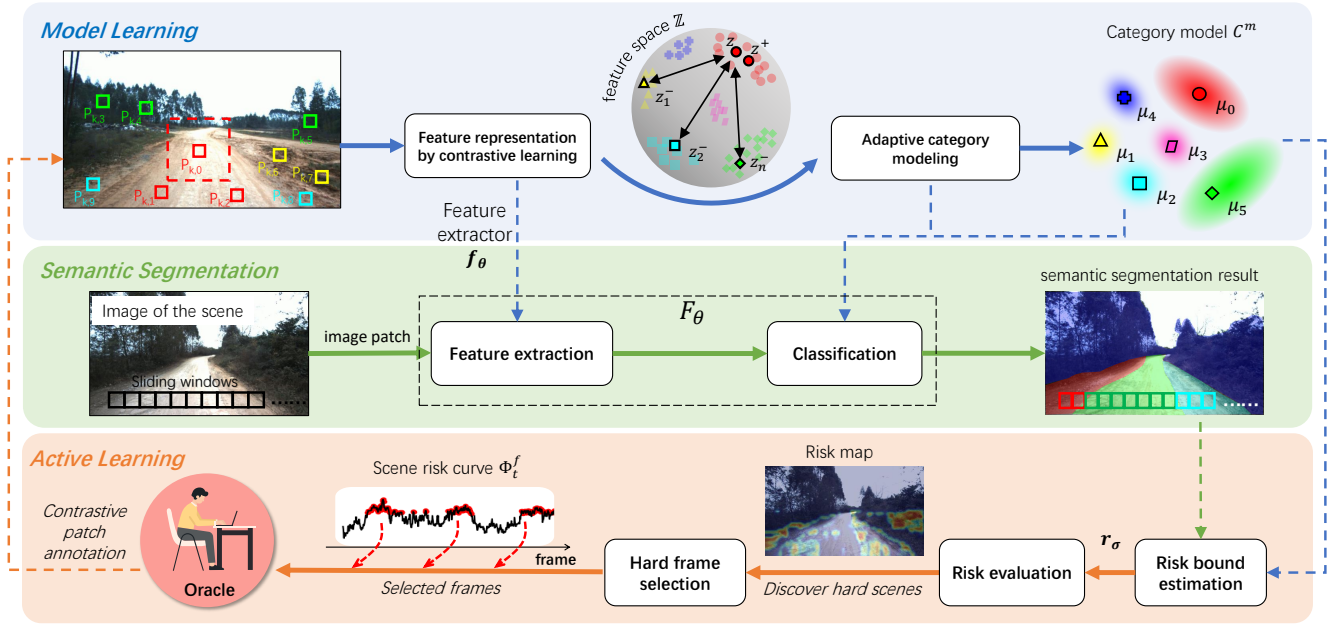


Fig. 2. The proposed active and contrastive learning framework for fine-grained semantic segmentation of off-road scenes.

III. METHODOLOGY

A. Outline

As illustrated in Fig. 3(a), many deep semantic segmentation models F_θ take the entire image I as input and map to semantic masks corresponding to each image pixels, which usually requires pixel-wise supervision. This research exploits a different flow as Fig. 3(b) to take patch-based annotations as weak supervision. Given an image I , generate a sliding window and find the semantic label y for each image patch x through the classifier F_θ . In this research, F_θ consists of a feature extractor f_θ that discriminates a given contrastive image patches in the feature space, and a maximum likelihood classifier based on the category model C^m that is learned by adaptive clustering of the training features.

The model F_θ is trained at scene D^{train} by a set of patch-based annotations $\{A_k\}$. At a new scene D^{new} , F_θ could be exposed to data that is substantially different from those in training, resulting in performance degradation. Such situation is very dangerous for safety-critical applications like autonomous driving. The agent needs to be aware of this performance degradation and require the model to be updated to accommodate the new scenes. To this end, a risk evaluation method is developed to discover when the model is no longer valid and the results are high-risk, and triggers the process of active learning. In the active learning process, the hard frames which are the most uncertain for the model are tend to be selected for human annotation, and update the model F_θ subsequently.

B. Model Learning

1) *Problem Formulation*: One training image frame I_k includes several anchor patch annotations $A_k = \{A_{k,i} = \langle p_{k,i}, a_{k,i} \rangle\}$. An anchor patch $A_{k,i}$ consists of an image patch

$p_{k,i}$ and a label $a_{k,i}$. Different from common defined semantic labels that map a label ID to a specific category among the whole dataset, in this research, the labels of anchor patches are comparable only if they belong to the same image. In other words, this $a_{k,i}$ only identifies image patches with similar or different semantic attributes in the current image. It provides great convenience for off-road data labeling, because it is difficult to determine a unified category list in advance for diverse off-road scenes.

Denoting $z = f_\theta(p)$ as an encoder that converts a high-dimensional image patch p to a normalized D -dimensional feature vector $z \in \mathbb{Z}^D$, then the exponential cosine similarity $sim(p_i, p_j)$ is used to evaluate the similarity of two image patches via their features z_i and z_j :

$$sim(p_i, p_j) = exp(z_i^T \cdot z_j) \quad (1)$$

The contrastive learning method is used to optimize f_θ , which making the similarity between anchor patches $sim(p_{k,i}, p_{k,j})$ be higher for $A_{k,i}$ and $A_{k,j}$ sharing the same label, i.e. $a_{k,i} = a_{k,j}$, and vice versa.

Through the optimized f_θ and extracted feature vectors $Z = \{z_1, z_2, \dots, z_N\}$, the category modeling aims to find the most applicable class number m and corresponding model parameters $C^m = \{c_1, c_2, \dots, c_m\}$ by adaptive clustering.

2) Feature Representation by Contrastive Learning:

a) *Contrastive Samples*: The core idea behind contrastive learning is to learn a f_θ that separates samples with different semantic meanings. At each step in the training process, contrastive learning requests a query sample q following one corresponding positive sample q^+ and n negative samples $\{q_i^- | i = 1, \dots, n\}$. Here, one query sample is an anchor patch $A_{k,i}$. Its corresponding positive and negative samples are all from the same image.

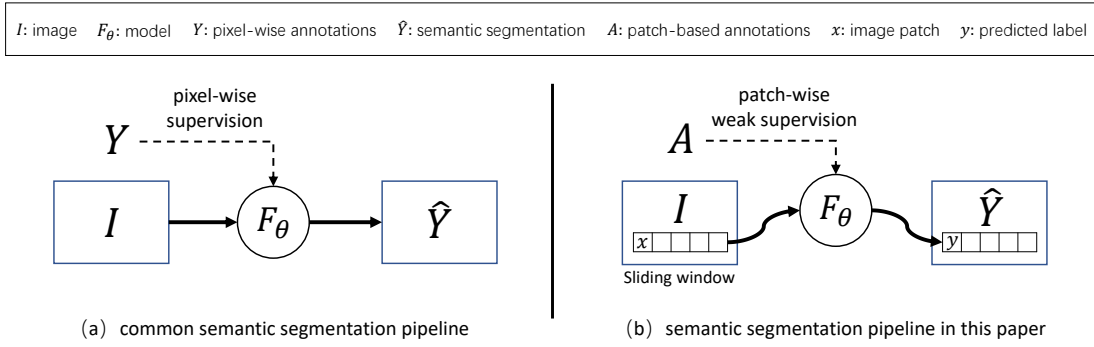


Fig. 3. Semantic segmentation pipeline in this paper.

Given a query sample with label $a_{k,i}$ in image frame I_k , other anchor patches in I_k can be divided to two sets according to the patch label $a_{k,i}$: one is positive anchor set $\{A_{k,i}^+\}$ with patches sharing the same label $a_{k,i}$; the other is negative anchor set $\{A_{k,i}^-\}$ including the rest patches. Positive and negative samples are selected from the aforesaid two sets respectively. The detailed sampling strategy is described in Section III-B2c.

b) Network Design and Loss Function: Use a convolutional neural network backbone to model f_θ , i.e. AlexNet [58], convert the tensor of query, positive or negative samples into a normalized feature vector z in low-dimensional embedding space \mathbb{Z}^D . The parameter θ is optimized by contrastive learning, aiming to increase the exponential cosine similarity of z s that share the same label, while decreasing those with different labels.

A contrastive loss function InfoNCE [59] is implemented:

$$\mathcal{L} = -\log \frac{\exp(z^T \cdot z^+ / \tau)}{\exp(z^T \cdot z^+ / \tau) + \sum_{i=1}^n \exp(z^T \cdot z_i^- / \tau)} \quad (2)$$

where τ denotes a temperature hyper-parameter, z^+ and z_i^- are feature vectors of positive and negative samples.

Unlike the typical contrastive learning setting [60], which uses a memory bank to save features of training samples, in this research, updated positive and negative sample features are calculated at each training step. This is because positive and negative samples are only comparable within the same image frame, which makes it possible to compute features with reasonable memory cost.

c) Sampling Strategy: To enrich data variety of a limited number of anchor patches, assuming that neighbor regions in the off-road environment are semantically similar, positive and negative samples of a query sample q are randomly drawn from the neighbor regions of q 's positive and negative anchors at the frame. As illustrated in Fig. 4(a), the neighbor sample are drawn with its center point locating inside the region of the original sample q .

When composing a sample data, contextual information is also included, which is crucial for classifying objects that lack texture. As shown in Fig. 4(b), given an RGB image patch q , treating it as the foreground q_f containing 3 channels, a background q_b with a larger cropped area is centered at q_f .

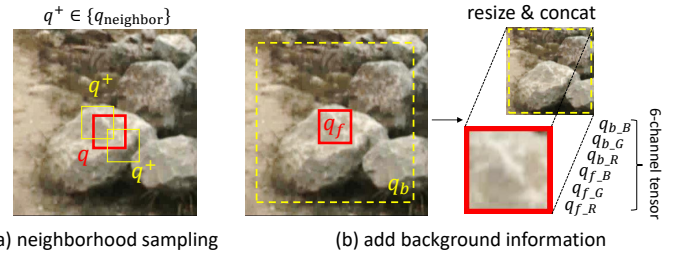


Fig. 4. Illustration of (a) neighborhood sampling strategy, and (b) how to add background information with the foreground image patch.

They are both reshaped to the same size and compose a 6-channel tensor as the input to the feature extractor f_θ . To enhance the robustness of the model in different environments with different illumination conditions, we implement data augmentation on the 6-channel tensor for each sample before feeding it to f_θ . Concretely, data augmentation contains random greyscale, random flip and color jitter (randomly changing the brightness, contrast, and saturation of an image).

3) Adaptive Category Modeling: Given a set of N D -dimensional data points $Z = \{z_1, z_2, \dots, z_N\}$ that are feature vectors extracted by f_θ on image patches $\mathcal{P} = \{p_1, p_2, \dots, p_N\}$, category modeling is to find the category number m and model parameters of the clusters $C^m = \{c_1, c_2, \dots, c_m\}$ that has the maximum likelihood on Z .

We consider the case where data following the multivariate Gaussian distribution, each cluster c_k is modeled by a mean vector μ_k and a covariance matrix Σ_k . The likelihood of data point z_i under cluster c_k is

$$\gamma(z_i; c_k) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(z_i - \mu_k)^T \Sigma_k^{-1} (z_i - \mu_k)\right\} \quad (3)$$

In the feature space \mathbb{Z}^D , the maximum likelihood function of the mixed category distribution C^m on data points Z can be given as follows [61],

$$L(C^m; Z) = \prod_{i=1}^N \sum_{k=1}^m \gamma(z_i; c_k), \quad (4)$$

For a certain cluster number m , the model parameters C^m can be estimated by maximizing $L(C^m; Z)$, where the EM

algorithm [62] is among the most popular approaches for parameter estimation.

To determine the number of clusters m , the Bayesian information criterion (BIC) [63] is used. Perform EM for each number of clusters $m = 2, \dots, M$, where M is an empirical value representing the maximal number of clusters. The BIC value is estimated as follows

$$BIC(C^m; Z) = -2 \log L(C^m; Z) + u \log(N) \quad (5)$$

where u is the number of model parameters. The m that leads to the decisive first local minimum of BIC value is found as the optimal number of clusters.

C. Semantic Segmentation and Active Learning

1) *Semantic Segmentation*: During semantic segmentation inference, for a given image frame I_t , a sliding window is conducted to generate image patches $\mathcal{P}_t = \{p_1, p_2, \dots, p_{N_t}\}$. For each image patch p_i , a feature vector is first extracted by $z_i = f_\theta(p_i)$, then a category label y_i is assigned by matching z_i with the category model C^m as below.

$$k^* = \arg \max_k (\gamma(z_i; c_k)) \quad (6)$$

$$r_i = 1 - \gamma(z_i; c_{k^*}) \quad (7)$$

$$y_i = \begin{cases} k^* & \text{if } r_i \leq \mathbf{r}_\sigma; \\ \phi & \text{otherwise} \end{cases} \quad (8)$$

where k^* is the cluster ID that has the maximal likelihood with z_i . r_i is the risk of classifying z_i to the most likely cluster k^* . k^* is assigned to category label y_i if and only if the risk r_i is below a certain risk bound \mathbf{r}_σ . Otherwise, y_i will be assigned a special label ϕ indicating the unknown class.

It is generally believed that all pixels in one patch belong to the same category. When acquiring image patches by a $s_p \times s_p$ sliding window with step size ξ , if higher resolution semantic segmentation is required, usually set $\xi < s_p$, so that the patches are partially overlapped. As a result, each pixel may get multiple predictions from different patches. The final label of each pixel is determined by the weighted voting method. When calculating, the closer a pixel is to the center of the patch, the higher the voting weight of the patch label.

After obtaining the semantic segmentation of the entire image, the widely used Dense Conditional Random Field (DenseCRF [64]) is used as an optional post-processing module to refine the predictions. In scenes with clear region boundaries, segmentation can be effectively refined.

2) *Risk Bound Estimation*: After learning the category model C^m on the training set \mathcal{D}^{train} including a total of N^{train} patches set \mathcal{P}^{train} , a confidence level δ of the learned model can be given by the operator's experience. It means that the proportion of risky classification is less than $1 - \delta$. Therefore, the risk bound \mathbf{r}_σ can be estimated by solving the following constrained optimization.

$$\begin{cases} \min & \mathbf{r}_\sigma \\ \text{s.t.} & \frac{|\{r_i > \mathbf{r}_\sigma\}|}{N^{train}} \leq 1 - \delta \end{cases} \quad (9)$$

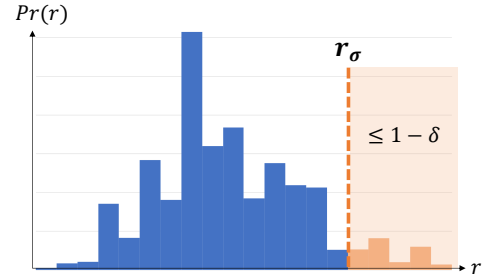


Fig. 5. Illustration of risk bound estimation.

An analytical solution is shown in Fig. 5. Generate a histogram $Pr(r)$ over the set of $\{r_i\}$, where each r_i is computed over $p_i \in \mathcal{P}^{train}$. Minimize \mathbf{r}_σ while satisfying $\frac{|\{r_i > \mathbf{r}_\sigma\}|}{N^{train}} \leq 1 - \delta$, equivalent to $\sum_{r > \mathbf{r}_\sigma}^{r_{\max}} Pr(r) \leq 1 - \delta$. The resulting risk bound \mathbf{r}_σ is used for the following risk evaluation.

3) *Risk Evaluation*: During inference, given a set of image patches $\mathcal{P}_t = \{p_1, p_2, \dots, p_{N_t}\}$ belonging to image frame I_t , along with the category model C^m with m clusters and a risk bound \mathbf{r}_σ , a set of labels $Y_t = \{y_1, y_2, \dots, y_{N_t}\}$ will be estimated, where $y_i \in \{1, \dots, m, \phi\}$. For convenience, we denote the classification of each data point by $y_i = \mathbf{y}(p_i | C^m)$. Let $\mathcal{P}_t^* \subset \mathcal{P}_t$ be the subset containing patches classified as ϕ ,

$$\mathcal{P}_t^* = \mathcal{P}_t^*(\mathcal{P}_t; C^m) = \{p_i \in \mathcal{P}_t \wedge \mathbf{y}(p_i | C^m) = \phi\} \quad (10)$$

The proportion of \mathcal{P}_t^* represents the degree of uncertainty or risk of the model in the scene described by the image frame I_t . An index describing the model uncertainty at the image frame level, i.e. **frame-level risk**, is subsequently defined:

$$\Phi_t^f = \Phi_t^f(\mathcal{P}_t; C^m) = \frac{|\mathcal{P}_t^*(\mathcal{P}_t; C^m)|}{N_t}, \quad (11)$$

where $|\mathcal{P}_t^*(\mathcal{P}_t; C^m)|$ is the size of the set \mathcal{P}_t^* .

When deploying on the autonomous vehicle, the data to be predicted is usually a sequence of T_s image frames $\{I_1, \dots, I_{T_s}\}$ including image patches $S = \{\mathcal{P}_1, \dots, \mathcal{P}_{T_s}\}$. Let $S^* \subset S$ be the subset containing risky frames that the model is uncertain, i.e. exceeding the risk level ϵ .

$$S^* = S^*(S; C^m) = \{P_t \in S \wedge \Phi_t^f(P_t; C^m) > \epsilon\} \quad (12)$$

The proportion of S^* represents the degree of risk, in other words, the uncertainty of the model on the dataset described by the sequence of image frames S . An index describing **sequence-level risk** is then defined.

$$\Phi^s = \Phi^s(S; C^m) = \frac{|S^*(S; C^m)|}{T_s} \quad (13)$$

In this research, sequence-level risk is a measure of discovering when a model is no longer valid and trigger the active learning process, while frame-level risk is to find the hard frames that the model is uncertain, which are requested for human annotation.

4) *Workflow*: The workflow of active learning for semantic segmentation is described below.

§1. Offline learning

§1-1. Initialization (Model learning)

Given a training dataset including anchor patches set A^{train} , learn a model f_θ by contrastive learning (Section III-B2), then find category model C^m by adaptive clustering (Section III-B3).

§1-2. Risk Bound Estimation

Given the learned category model C^m on the training patches \mathcal{P}^{train} from A^{train} , estimate a risk bound r_σ (Eqn. 9).

§2. Online Semantic Segmentation

§2-1. Semantic Segmentation

Given a test image frame I_t , generate a set of image patches \mathcal{P}_t using a sliding window. For each image patch p_i , find the corresponding label y_i (Eqn. 8) and risk r_i (Eqn. 7).

§2-2. Risk Evaluation

For consecutive T_s test image frames $\{I_1, \dots, I_{T_s}\}$ including image patches $S = \{\mathcal{P}_t\}$, estimate $\{\mathcal{P}_t^*\}$ (Eqn. 10), $\{\Phi_t^f\}$ (Eqn. 11), S^* (Eqn. 12) and Φ^s (Eqn. 13). If the sequence-level risk Φ^s exceeds a certain threshold, the active learning module will be triggered to update the current model.

§3. Active Learning

§3-1. Hard Frame Selection

Choose a batch of \mathcal{B} image frames $\{I_{u_1}, \dots, I_{u_{\mathcal{B}}}\}$ with image patches $S_u = \{\mathcal{P}_{u_1}, \dots, \mathcal{P}_{u_{\mathcal{B}}}\} \subset S$ to meet

$$\left\{ \begin{array}{l} \max \sum_{t=u_1}^{u_{\mathcal{B}}} \Phi_t^f \\ \text{s.t.} \quad \forall \mathcal{P}_{u_i}, \mathcal{P}_{u_j} \in S_u, |u_i - u_j| > \Delta \end{array} \right. , \quad (14)$$

where Δ is a threshold to avoid selection on neighbor frames that provide repetitive information.

§3-2. Human Annotation

For selected \mathcal{B} frames, annotate contrastive image patches and corresponding labels to obtain $\mathcal{A} = \{A_{u_1}, \dots, A_{u_{\mathcal{B}}}\}$, where $A_u = \{\langle p_{u,i}, a_{u,i} \rangle\}$, then update the supplemental annotations $A^{AL} \leftarrow A^{AL} + \mathcal{A}$.

§3-3. Model Update

Fine-tune f_θ on the A^{AL} , then estimate category model C^m and risk bound r_σ . Finally, continue semantic segmentation process (go to §2-1).

IV. EXPERIMENTAL DESIGN

A. Notations

1) *Experiment Stage*: For clear interpretation, the following notations are introduced:

- *Learning*: the initial training stage of semantic segmentation models. Given a training set containing anchor annotations, train the feature extractor f_θ and the corresponding category model C^m by contrastive learning.
- *Active Learning*: Given a learned model, implement it on a new dataset and use active learning to obtain supplemental annotations, then update the model.

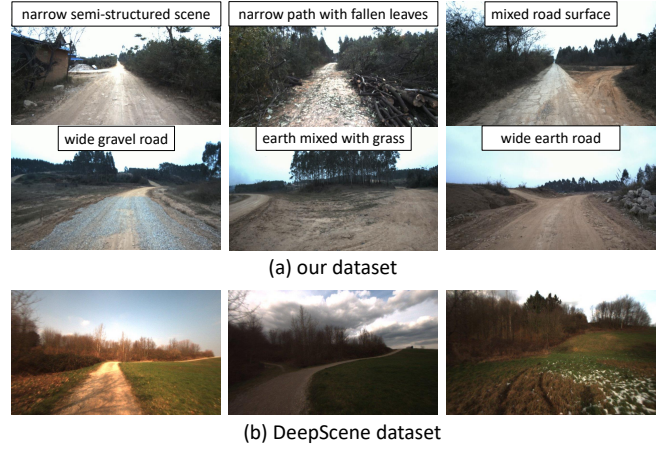


Fig. 6. Typical scenes in (a) our off-road dataset and (b) DeepScene dataset [17].

- *Test*: evaluate the performance of semantic segmentation models.
- 2) *Model*: The notations of semantic segmentation models:
 - M_A : the model trained on dataset A ;
 - M_A^B : based on the initial model M_A , the model updated by active learning on dataset B .

When emphasizing the semantic granularity of annotations, the following notations are used:

- M_{ALv1} : the model trained by annotations with $Lv1$ (Level 1) semantic granularity on dataset A . In experiment, the granularity of anchor annotations include 3 levels, i.e. $Lv1$, $Lv2$ and $Lv3$.

When indicating the frame number for training, the following notations are used:

- M_{A50} : the model trained by 50 image frames of dataset A ;
- M_{A50}^{B20} : based on the initial model M_{A50} , the model updated by active learning on 20 frames of dataset B .

When comparing different methods, M is usually replaced by the method's abbreviation, such as $Base_A$, $Rand_{A50}^{B20}$ and $Ours_{A50}^{B20}$.

B. Dataset

1) *Our Dataset*: We developed an off-road dataset for experimental validation of the proposed method. As shown in Fig. 6(a), the images are collected by a front-view monocular RGB camera mounted on a moving vehicle. As shown in Table I, our dataset includes 3 sub-datasets (noted as A/B/C) for different experimental stages.

Subset A contains 5064 frames for experiments of the *Learning* stage. We randomly sample 50 frames for patch-based annotations and use them to train a contrastive learning-based feature extractor and obtain adaptive category modeling, evaluating its performance by randomly selecting patches on other images.

Subset B includes 1639 frames for evaluating the active learning pipeline. The risk evaluation module will check the results of M_A . When the active learning is activated, X frames

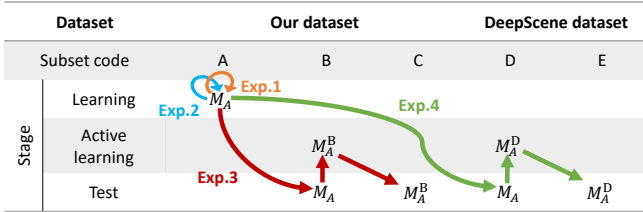
TABLE I
DATASET STATISTICS

Dataset	Ours			DeepScene		
	Subset code	A	B	C	D	E
Exp. stage	Learning	Active learning	Test	Active learning	Test	
Annotated anchor frames	50	10~30	-	20~40	-	
Total frames	5064	1639	1600	230	136	

* The train and test set of DeepScene dataset are noted as D and E respectively.

Exp.	Purpose	Dataset usage		
		Learning	Active learning	Test
Exp. 1	Contrastive learning for feature representation	A	-	-
Exp. 2	Adaptive category modeling	A	-	-
Exp. 3	Active learning	A	B	C
Exp. 4	Active learning across datasets	A	D	E

(a) Experimental design and dataset usage



(b) Experimental pipeline

Fig. 7. Experimental design and pipeline. (a) experimental pipeline. (b) experimental design and dataset usage.

will be selected from hard frames for human annotation and the model is updated to M_A^B .

Subset C includes 1600 frames for evaluating the improvements from active learning. Concretely, the effectiveness of active learning can be examined by comparing the semantic segmentation results and risk-based metrics between the initial model M_A and the updated model M_A^B .

2) *DeepScene Dataset*: Besides our dataset, the proposed method is also evaluated on the DeepScene dataset [17]. It contains several types of camera data, while this paper only uses monocular RGB images. As shown in Fig. 6(b), compared to our dataset, the DeepScene dataset has different environments and illumination to examine the generalization ability of the proposed method in different scenes.

The train and test set of DeepScene are noted as dataset D and E, including 230 and 136 frames respectively. Each image frame has pixel-wise semantic annotations. Similar to dataset B/C, the dataset D/E are respectively used for training and testing of the active learning models.

C. Experimental Design

Four experiments are designed as shown in Fig. 7, which are introduced as follows:

1) *Exp.1 Contrastive learning for Feature Representation*: It aims to evaluate the proposed contrastive learning method for feature extraction. Train the model M_A by annotations on 50 randomly selected frames from dataset A. The results are shown in Section V-A.

2) *Exp.2 Adaptive Category Modeling*: It is designed to examine the results of adaptive category modeling. In the experiments, we use dataset A labeled by 3 granularity levels, noted from coarse-grained to fine-grained as A^{Lv1} , A^{Lv2} and A^{Lv3} . The results are shown in Section V-B.

3) *Exp.3 Active Learning*: It aims to evaluate the proposed active learning pipeline. The initial model M_A trained on dataset A is deployed on dataset B. After activating the active learning module to select a few image frames for human annotation, the model is updated from M_A to M_A^C and tested on dataset C. The results are shown in Section V-C.

4) *Exp.4 Active Learning across Datasets*: It aims to demonstrate the cross-dataset generalization ability of the proposed active learning method. The initial model M_A is deployed on dataset D. After activating the active learning module to select a few image frames for human annotation, the model is updated from M_A to M_A^D and tested on dataset E. The results are shown in Section V-D.

D. Evaluation Metrics

On our dataset, two risk-based metrics are used to evaluate models' performance.

- **FLR** (frame-level risk). Note the frame-level risk Φ_t^f (Equation 11) as **FLR**, then **mFLR** indicates the mean **FLR** of a data sequence, which describes the average proportion of high risk patches per frame. The lower **FLR** value means the better model performance.
- **Sc** (scene coverage), defined as $\mathbf{Sc} = 1 - \Phi_t^s$, where sequence-level risk Φ_t^s is described by Equation 13. It means the certainty of the model over the entire data sequence, i.e. the proportion of non-risk frames. The higher **Sc** value means the better performance.

On the DeepScene dataset, 5 common metrics for semantic segmentation tasks are introduced:

- **mIoU** (mean Intersection over Union), **PA** (Pixel Accuracy), **PRE** (Precision), **REC** (Recall), **FPR** (False Positive Rate).

V. EXPERIMENTAL RESULTS

A. Exp. 1: Contrastive Learning for Feature Representation

In the initial training stage of the semantic segmentation model, the feature extractor f_θ trained by contrastive learning is dedicated to narrowing down the semantically similar image patches in the feature space, while pushing away different image patches. Fig. 8 shows the patch similarity (Equation 1) calculated by the feature extractor of model M_A on non-training image frames. Similarity values are marked above the image patches.

As shown in Fig. 8(a), several image patches are randomly selected in each image, one of which is regarded as an anchor patch (marked as red A). Different colors visualize feature

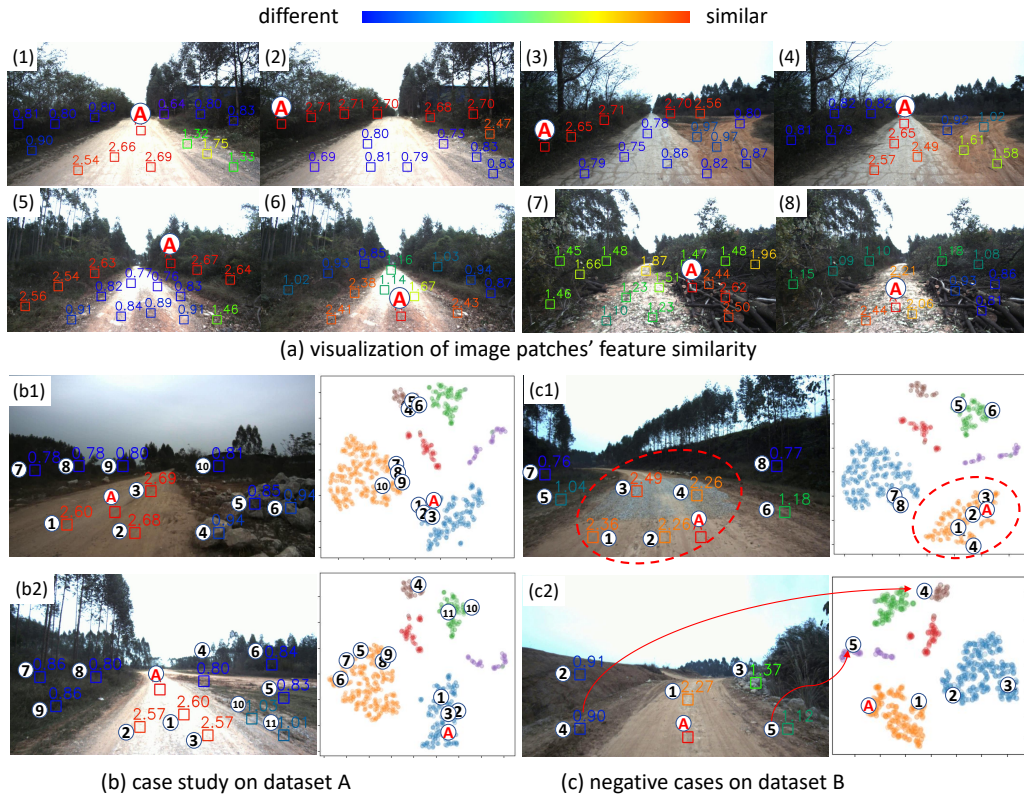


Fig. 8. Visualization of image patches' feature similarity. Randomly choose one patch as an anchor patch (red A), other patches' color indicate their similarity to the anchor patch (Equation 1). (a) image patches' feature similarity at different scenes. (b) case study on dataset A: the corresponding positions of the image patches in the feature space. (c) negative cases on dataset B: feature similarity does not match semantic meanings.

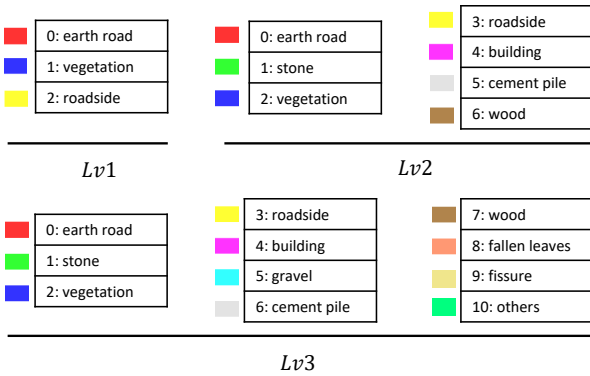


Fig. 9. The reference label definitions of different semantic granularity.

similarities between the anchor patch and other patches. Colors closer to red indicate higher similarity.

Fig. 8(b) selects two cases in dataset A for concrete analysis. The feature vector of each image patch is reduced by t-SNE[65], and then drawn on the right. The patch IDs in the two images correspond to each other. The other points in the feature map are from the training patches. For both cases in Fig. 8(b), the feature similarity between each image patch and anchor is consistent with their semantic relationship, and image patches of the same category are relatively concentrated in the feature map. Such distribution provides a prerequisite for subsequent adaptive category modeling.

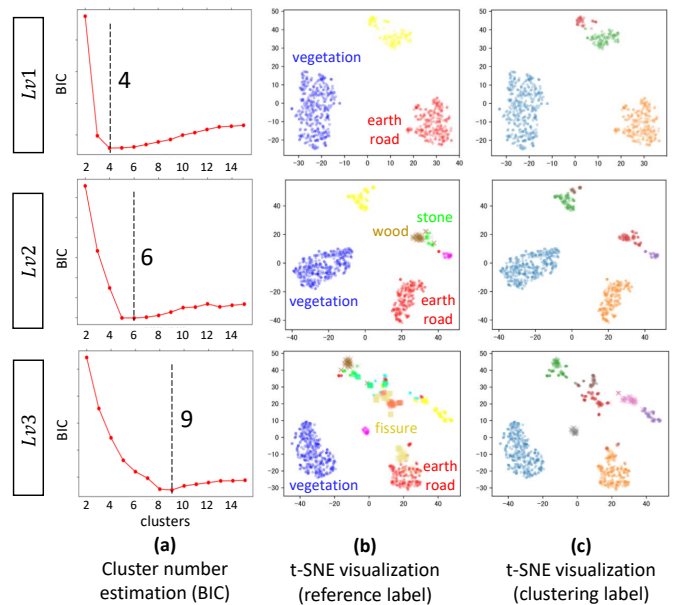


Fig. 10. The cluster number estimated by BIC and adaptive category modeling results under different granularity annotations. (a) cluster number estimation. (b) category modeling results colored by reference labels. (c) category modeling results colored by clustering labels.

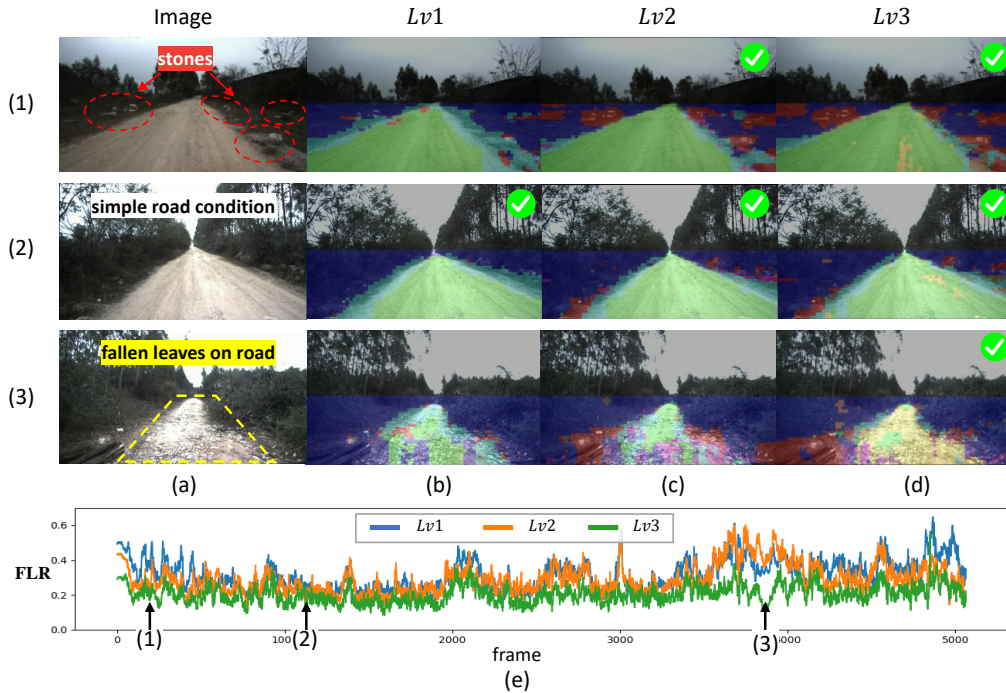


Fig. 11. Analysis of adaptive category modeling results under different granularities. (a) input image; (b-d) predictions from the model under granularity $Lv1$ - $Lv3$; (e) The **FLR** curves of models trained by three granularities on dataset A.

Fig. 8(c) shows some negative cases of model M_A in an unknown scenario (dataset B). In Fig. 8(c1), the gravel road (patch ID 3,4) in the red circle does not appear in the training set, while the earth road samples (patch ID 1,2) have a high-level similarity with them and distribute close in the feature map. It indicates that the current features extracted by f_θ are difficult to distinguish between the two categories. In Fig. 8(c2), the image patch 4 and 5 (muddy areas on the roadside) of the same category are distributed farther in the feature map, which also reflects the weakness of f_θ . In the new scenes, the generalization ability of the features extractor is insufficient, and active learning needs to be introduced to update the model.

B. Exp. 2: Adaptive Category Modeling

The mechanical properties of unmanned platforms are different, which requests different granularities for semantic segmentation, i.e. the types of terrain that need to be distinguished are also different. According to the mechanical properties of the platform, we define three reference label sets with different granularities, as shown in Fig. 9. It needs to be emphasized that it is very difficult to annotate pixel-wise labels based on these label sets due to the ambiguity between classes. However, they can provide guidance for labeling patch-based positive and negative samples.

Fig. 10(a) shows the result of using Bayesian Information Criterion (BIC) to determine the number of clusters in adaptive category modeling. Under the three granularity annotations $Lv1$, $Lv2$ and $Lv3$, the training samples are adaptively clustered into 4, 6, and 9 categories, respectively. Figure 10 (b-c) is the visualization result of the image patch features in the

training data after dimensional reduction by t-SNE[65], which is colored according to the reference label and clustering label.

From the clustering distribution of Fig. 10(b-c), under all three semantic granularities, *earth road*, *vegetation* and other dominated categories adaptive clustering results are basically consistent with the reference label. Small-sized clusters are usually from categories with low frequency, such as *gravel*, *fallen leaves*, etc.

There exist differences between these clustering results and the reference labels, which are mainly divided into two situations: one is the category with complex and diverse appearance. For example, the *fissure* samples in Fig. 10(b)- $Lv3$ has multiple scattered clusters, indicating high variance within the category. The second situation is the category with similar features to other clusters. For example, in Fig. 10(b)- $Lv2$, the distribution of *wood* samples is relatively concentrated, but very close to *stone* samples. In category modeling, they can easily be misclassified into the same category.

Concrete cases are shown in Fig. 11. The *stones* in case (1) have been effectively modeled at granularity $Lv2$ and $Lv3$, but failed to be segmented at coarse-grained $Lv1$. In case (2), such simple road condition only includes basic categories like earth roads and vegetation. The corresponding semantic segmentation results under different granularity are basically the same. From the **FLR** curve in Fig. 11(e), it can also be seen that the risk values of different models in case (2) are close and low, indicating that a basic category modeling is enough to handle such simple scenes. In case (3), the roads covered by *fallen leaves* are marked as yellow in the semantic segmentation results. Only the fine-grained $Lv3$ model can distinguish this category. From the corresponding **FLR** curves in Fig. 11(e), it can be seen that the curve of $Lv3$

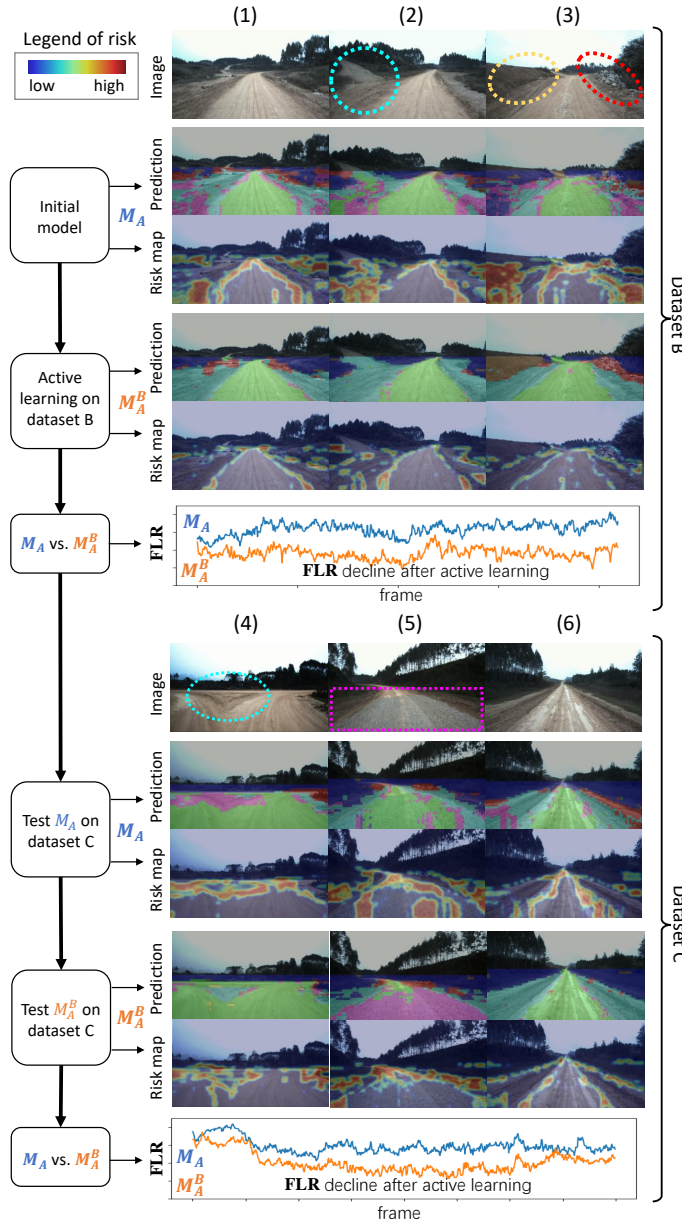


Fig. 12. Exp. 3: Comparison of model's semantic segmentation results before and after active learning.

is significantly reduced. The fine-grained category modeling effectively reduces prediction risk.

C. Exp. 3: Results of Active Learning

Fig. 12 shows the semantic segmentation results at different stages in the active learning process. First, deploy the model M_A originally trained on dataset A on dataset B, trigger the active learning module to select a small number of frames in difficult scenes, and after manual supplementary annotation, the model is updated to M_A^B , then compare its performance to M_A . Finally, the semantic segmentation performance of the model M_A and M_A^B are compared on dataset C. Fig. 12(1-3) and (4-6) show the semantic segmentation results and **FLR** curves of the two models on dataset B and C respectively.

TABLE II
PERFORMANCE OF ACTIVE LEARNING METHODS

	model	mFLR ↓		Sc	
		Dataset B	Dataset C	Dataset B	Dataset C
comparison of frame selection strategy	M_A	54.25%	59.04%	35.50%	8.75%
	$Rand_A^{B20}$	46.52%	49.50%	78.00%	66.25%
	$Unif_A^{B20}$	44.43%	44.68%	83.63%	73.00%
	$Ours_A^{B20}$	44.34%	40.67%	85.00%	83.88%
comparison of labeled frame numbers	$Ours_A^{B10}$	51.57%	50.70%	44.88%	61.50%
	$Ours_A^{B15}$	45.58%	40.58%	77.88%	81.50%
	$Ours_A^{B20}$	44.34%	40.67%	85.00%	83.88%
	$Ours_A^{B25}$	43.03%	38.12%	88.13%	84.75%
	$Ours_A^{B30}$	39.22%	35.36%	91.00%	86.38%

¹ All models' subscript A50 is abbreviated to A;

² $Rand_A^{B20}$: random select frames for active learning; $Unif_A^{B20}$: uniformly select frames for active learning;

↓: lower value means better performance.

TABLE III
PERFORMANCE ON DEEPSCENE DATASET

type	class	model	PA	IoU	PRE	REC	FPR↓
weakly sup.	5	M_A	75.28	16.95	32.65	43.43	15.35
		$Ours_A^{D20}$	91.30	49.60	63.84	70.40	5.92
		$Ours_A^{D40}$	92.66	53.96	66.76	71.56	5.02
		$Ours_A^{D40}+CRF$	95.16	61.26	78.72	76.30	3.67
	4	$Ours_A^{D20}$	89.81	61.51	78.85	78.56	6.76
		$Ours_A^{D40}$	91.89	68.81	82.69	84.42	5.73
fully sup.	5	SegNet [66]	88.47	74.81	84.63	86.39	13.53
		FCN [22]	90.95	77.46	87.38	85.97	10.32
		ParseNet [67]	93.43	83.65	90.07	91.57	8.94

¹ **IoU**: Intersection over Union, **PA**: Pixel Accuracy, **PRE**: Precision, **REC**: Recall, **FPR**: False Positive Rate.

In Fig. 12(1-3), it can be found that the predictions of M_A on dataset B contains many noises. For example, the lime soil inside the cyan circle in Fig. 12(2) and the stones inside the red circle in Fig. 12(3) are predicted to be a mixture of multiple categories, and the corresponding positions in risk maps also present a high prediction risk (red). In Fig. 12(3), the earth embankment inside the yellow circle is a new type for model M_A , and its segmentation results are noisy and the risk value is also high. After active learning, the model M_A^B gives better semantic segmentation results in aforesaid scenes, and the high-risk areas in the risk map are also significantly reduced. Among the dataset, the reduction of frame-level risk is reflected by the **FLR** curve. After active learning, the orange curve of M_A^B is significantly lower than the blue curve of M_A .

On dataset C, we also test and compare the performance of model M_A and M_A^B . The muddy triangle area in Fig. 12(4), the gravel road in Fig. 12(5), etc., all show the better semantic

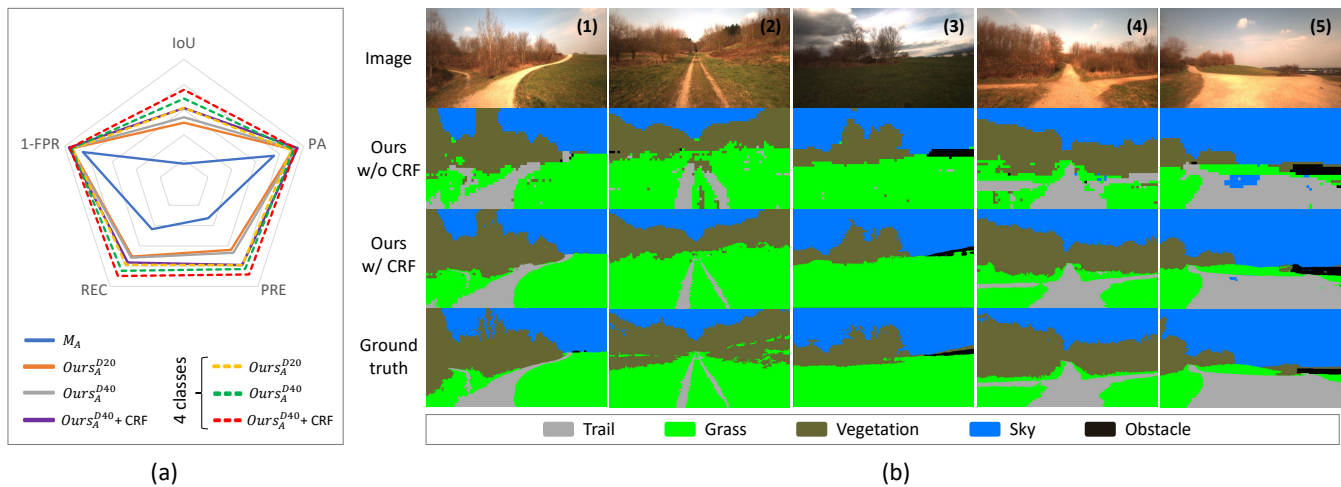


Fig. 13. Semantic segmentation results on the DeepScene dataset. (a) the radar chart visualization of methods' performance shown in Table III. (b) qualitative semantic segmentation results of the proposed method.

segmentation performance of M_A^B , which is also reflected in the lower **FLR** curve on dataset C.

Table II quantitatively compares the performance of different models. The upper part of the table compares different frame selection strategies. Each model selects 20 frames on dataset B for active learning. $Rand_A^{B20}$ randomly selects frames for annotation, $Unif_A^{B20}$ uses uniform sampling, and $Ours_A^{B20}$ uses the hard frame selection strategy proposed in Section III-C4. Under the metrics **mFLR** and **S_c**, the proposed method shows significant advantages. The bottom half of Table II evaluates the effect of the active sampled frame number B . In general, more supplemented annotations lead to better model performance.

D. Exp. 4: Results of Active Learning across Datasets

Table III is the performance comparison of different models on DeepScene dataset. Based on the pixel-wise labels of DeepScene, several classical semantic segmentation metrics are evaluated for comparison. According to the category number, models are divided into two groups: (1) models with 5 categories, consistent with the label definitions of DeepScene dataset; (2) models with 4 categories based on the adaptive category modeling, the ignored label corresponds to *obstacle* in DeepScene. *Obstacle* samples are very rare in the training set of DeepScene, accounting for only 0.33%. The active learning module only selects a few image frames, which makes it difficult for *obstacles* to get annotations. Therefore, it is not considered as a valid independent label during adaptive category modeling.

In the table, the 5-class model $Ours_A^{D20}$ uses only 20 frames of patch-based weak annotations in the new environment, while achieving 16.02% PA and 32.65% IoU improvement over the initial model M_A . The model $Ours_A^{D40}$ further improves the metrics. Since the road boundaries in the DeepScene dataset are relatively clear, the post-processing module DenseCRF[64] can refine the semantic segmentation results. The corresponding model $Ours_A^{D40}+CRF$ achieves the overall

best performance. Under both label definitions, the proposed active learning method brings significant performance gains.

Comparing models with different category numbers: when the numbers of frames for supplementary annotation are the same, the 4-class model obtained by adaptive category modeling outperforms the 5-class model in all indicators. Among them, the optimal models under both category definitions have better FPR and PA than some classic fully supervised methods. The 4-class model $Ours_A^{D40}+CRF$ achieves the same level as the fully supervised methods in all evaluation metrics.

Fig. 13(a) uses a radar chart to visualize model performance in Table III. Some concrete examples are shown in Fig. 13(b). In various scenes under different lighting conditions, the model $Ours_A^{D40}+CRF$ (5 categories) requires only 40 frames of low-cost patch-based annotations, while achieving good semantic segmentation results.

VI. CONCLUSION

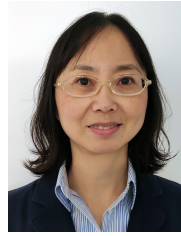
In this paper, we propose a framework of fine-grained off-road semantic segmentation based on active and contrastive learning. Through patch-based weak annotations, a contrastive learning-based feature extractor is learned to discriminate different semantic attributes. After that, an adaptive category modeling method is proposed, then a sliding-window-based semantic segmentation is exploited. To help the model adapt to new scenes efficiently, a risk evaluation method is developed to discover and select hard frames for active learning of new scenes. To evaluate the proposed method, extensive experiments are conducted on the self-developed off-road dataset with a total of 8000 image frames and the public DeepScene dataset. With only dozens of image frames as weak supervision, the fine-grained off-road semantic segmentation model can be learned. When detecting performance degradation in new scenes, the proposed active learning method can effectively select hard frames for the current model by risk evaluation and improve results with no more than 40 frames of patch-based annotations. Experiments on the DeepScene dataset show that the proposed weakly supervised method

can achieve the same level of performance as typical fully supervised ones. Future work will be addressed on preventing models from catastrophic forgetting after adapting to new scenes. Research about continual learning and incremental learning will be explored in the future.

REFERENCES

- [1] C. Badue *et al.*, “Self-driving cars: A survey,” *Expert Systems with Applications*, vol. 165, p. 113 816, 2021.
- [2] 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.
- [3] D. Braid *et al.*, “The terramax autonomous vehicle,” *Journal of Field Robotics*, vol. 23, no. 9, pp. 693–708, 2006.
- [4] H. Shariati *et al.*, “Towards autonomous mining via intelligent excavators,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 26–32.
- [5] H. Fang *et al.*, “Trajectory tracking control of farm vehicles in presence of sliding,” *Robotics and Autonomous Systems*, vol. 54, no. 10, pp. 828–839, 2006.
- [6] 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.
- [7] J. Shi *et al.*, “Fast and robust vanishing point detection for unstructured road following,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 970–979, 2015.
- [8] 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.
- [9] B. Gao *et al.*, “Off-road drivable area extraction using 3d lidar data,” in *IEEE Intelligent Vehicles Symposium*, 2019, pp. 1505–1511.
- [10] H. Kong *et al.*, “Vanishing point detection for road detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 96–103.
- [11] Y. Alon *et al.*, “Off-road path following using region classification and geometric projection constraints,” in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, vol. 1, 2006, pp. 689–696.
- [12] L. Wellhausen *et al.*, “Where should I walk? predicting terrain properties from images via self-supervised learning,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1509–1516, 2019.
- [13] S. Minaee *et al.*, “Image segmentation using deep learning: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [14] B. Gao *et al.*, “Are we hungry for 3d lidar data for semantic segmentation? a survey of datasets and methods,” *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–19, 2021.
- [15] M. Cordts *et al.*, “The cityscapes dataset for semantic urban scene understanding,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.
- [16] J. Behley *et al.*, “SemanticKITTI: A dataset for semantic scene understanding of lidar sequences,” in *IEEE International Conference on Computer Vision*, 2019, pp. 9297–9307.
- [17] A. Valada *et al.*, “Deep multispectral semantic scene understanding of forested environments using multimodal fusion,” in *International Symposium on Experimental Robotics (ISER)*, 2016.
- [18] M. Wigness *et al.*, “A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments,” in *IEEE/RSSJ International Conference on Intelligent Robots and Systems*, IEEE, 2019, pp. 5000–5007.
- [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] B. Rothrock *et al.*, “Spoc: Deep learning-based terrain classification for mars rover missions,” in *AAIA SPACE*, 2016, p. 5539.
- [22] J. Long *et al.*, “Fully convolutional networks for semantic segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [23] I. Sgibnev *et al.*, “Deep semantic segmentation for the off-road autonomous driving,” *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 617–622, 2020.
- [24] Y. Jin *et al.*, “Memory-based semantic segmentation for off-road unstructured natural environments,” *arXiv preprint arXiv:2108.05635*, 2021.
- [25] K. Viswanath *et al.*, “Offseg: A semantic segmentation framework for off-road driving,” in *IEEE International Conference on Automation Science and Engineering*, IEEE, 2021, pp. 354–359.
- [26] T. Guan *et al.*, “Ganav: Group-wise attention network for classifying navigable regions in unstructured outdoor environments,” *arXiv preprint arXiv:2103.04233*, 2021.
- [27] S. Chiodini *et al.*, “Evaluation of 3d cnn semantic mapping for rover navigation,” in *IEEE International Workshop on Metrology for AeroSpace*, IEEE, 2020, pp. 32–36.
- [28] D. Maturana *et al.*, “Real-time semantic mapping for autonomous off-road navigation,” in *International Conference on Field and Service Robotics*, 2018, pp. 335–350.
- [29] D.-K. Kim *et al.*, “Season-invariant semantic segmentation with a deep multimodal network,” in *Field and Service Robotics*, Springer, 2018, pp. 255–270.
- [30] 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.
- [31] S. Sharma *et al.*, “Semantic segmentation with transfer learning for off-road autonomous driving,” *Sensors*, vol. 19, no. 11, p. 2577, 2019.
- [32] L. Tang *et al.*, “From one to many: Unsupervised traversable area segmentation in off-road environment,” in *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, IEEE, 2017, pp. 787–792.
- [33] J. Zürn *et al.*, “Self-supervised visual terrain classification from unsupervised acoustic feature learning,” *Transactions on Robotics*, 2020.
- [34] A. v. d. Oord *et al.*, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [35] Y. Tian *et al.*, “Contrastive multiview coding,” *arXiv preprint arXiv:1906.05849*, 2019.
- [36] K. He *et al.*, “Momentum contrast for unsupervised visual representation learning,” in *Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
- [37] Z. Wu *et al.*, “Unsupervised feature learning via non-parametric instance discrimination,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.
- [38] P. Khosla *et al.*, “Supervised contrastive learning,” in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 18 661–18 673.
- [39] X. Zhao *et al.*, “Contrastive learning for label-efficient semantic segmentation,” *arXiv preprint arXiv:2012.06985*, 2020.
- [40] W. Wang *et al.*, “Exploring cross-image pixel contrast for semantic segmentation,” *arXiv preprint arXiv:2101.11939*, 2021.
- [41] B. Settles, “Active learning literature survey,” 2009.
- [42] P. Ren *et al.*, “A survey of deep active learning,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 9, pp. 1–40, 2021.
- [43] B. Settles *et al.*, “An analysis of active learning strategies for sequence labeling tasks,” in *Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 1070–1079.
- [44] R. Hwa, “Sample selection for statistical parsing,” *Computational linguistics*, vol. 30, no. 3, pp. 253–276, 2004.
- [45] K. Wang *et al.*, “Cost-effective active learning for deep image classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.
- [46] Y. Gal *et al.*, “Deep bayesian active learning with image data,” in *International Conference on Machine Learning*, PMLR, 2017, pp. 1183–1192.
- [47] Y. Guo, “Active instance sampling via matrix partition,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2010, pp. 802–810.
- [48] H. T. Nguyen *et al.*, “Active learning using pre-clustering,” in *International Conference on Machine Learning*, 2004, p. 79.
- [49] A. Freytag *et al.*, “Selecting influential examples: Active learning with expected model output changes,” in *European Conference on Computer Vision*, Springer, 2014, pp. 562–577.
- [50] N. Roy *et al.*, “Toward optimal active learning through monte carlo estimation of error reduction,” *International Conference on Machine Learning*, vol. 2, pp. 441–448, 2001.
- [51] B. Settles *et al.*, “Multiple-instance active learning,” *Advances in Neural Information Processing Systems*, vol. 20, pp. 1289–1296, 2007.
- [52] W. H. Beluch *et al.*, “The power of ensembles for active learning in image classification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9368–9377.

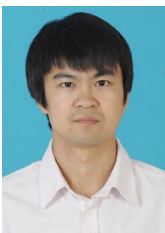
- [53] Y. Siddiqui *et al.*, "Viewal: Active learning with viewpoint entropy for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9433–9443.
- [54] R. Mackowiak *et al.*, "Cereals-cost-effective region-based active learning for semantic segmentation," *arXiv preprint arXiv:1810.09726*, 2018.
- [55] L. Yang *et al.*, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *International conference on medical image computing and computer-assisted intervention*, Springer, 2017, pp. 399–407.
- [56] M. Gorriz *et al.*, "Cost-effective active learning for melanoma segmentation," *arXiv preprint arXiv:1711.09168*, 2017.
- [57] S. Xie *et al.*, "Deal: Difficulty-aware active learning for semantic segmentation," in *Asian Conference on Computer Vision*, 2020.
- [58] A. Krizhevsky *et al.*, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
- [59] A. v. d. Oord *et al.*, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
- [60] Z. Wu *et al.*, "Unsupervised feature learning via non-parametric instance discrimination," in *Conference on Computer Vision and Pattern Recognition*, Jun. 2018.
- [61] C. Fraley, "Algorithms for model-based gaussian hierarchical clustering," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 270–281, 1998.
- [62] G. J. McLachlan *et al.*, "Finite mixture models," *Annual review of statistics and its application*, vol. 6, pp. 355–378, 2019.
- [63] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978.
- [64] P. Krähenbühl *et al.*, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, pp. 109–117, 2011.
- [65] L. Van der Maaten *et al.*, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [66] V. Badrinarayanan *et al.*, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [67] W. Liu *et al.*, "ParseNet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.



Huijing Zhao received B.S. degree in computer science from Peking University in 1991. She obtained M.E. degree in 1996 and Ph.D. degree in 1999 in civil engineering from the University of Tokyo, Japan. From 1999 to 2007, she was a postdoctoral researcher and visiting associate professor at the Center for Space Information Science, University of Tokyo. In 2007, she joined Peking University as a tenure-track professor at the School of Electronics Engineering and Computer Science. She became an associate professor with tenure on 2013 and was promoted to full professor on 2020. She has research interest in several areas in connection with intelligent vehicle and mobile robot, such as machine perception, behavior learning and motion planning, and she has special interests on the studies through real world data collection.



Biao Gao received B.S. degree in computer science (machine intelligence) from Peking University, Beijing, China, in 2017, where he is currently pursuing the Ph.D. degree with the Key Laboratory of Machine Perception (MOE), Peking University. His research interests include intelligent vehicles, 3D LiDAR perception, computer vision, and deep learning.



Xijun Zhao was born in Yanji City, Jilin, China, in 1984. He received the B.S degrees in vehicular engineering from Beijing Institute of Technology, Beijing, China, in 2007 and the Ph.D degree in mechanical engineering from Beijing Institute of Technology, Beijing, China, in 2011. He is now working with China North Vehicle Research Institute. His research interests include perception, localization, motion planning and control of Unmanned Ground Vehicles.