IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 3, 2021, accepted February 8, 2021, date of publication February 16, 2021, date of current version February 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3059620

Audio-Visual Self-Supervised Terrain Type Recognition for Ground Mobile Platforms

AKIYOSHI KUROBE[1], (Member, IEEE), YOSHIKATSU NAKAJIMA[1], (Member, IEEE), KRIS KITANI[2], (Member, IEEE), AND HIDEO SAITO[1], (Senior Member, IEEE)

[1]Department of Science and Technology, Keio University, Yokohama 223-8522, Japan
[2]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Corresponding authors: Akiyoshi Kurobe (kurobe@hvrl.ics.keio.ac.jp) and Yoshikatsu Nakajima (nakajima@hvrl.ics.keio.ac.jp)

This work was supported in part by the JST CREST under Grant JPMJCR19F3, and in part by the JST-Mirai Program, Japan, under Grant JPMJMI19B2.

ABSTRACT The ability to recognize and identify terrain characteristics is an essential function required for many autonomous ground robots such as social robots, assistive robots, autonomous vehicles, and ground exploration robots. Recognizing and identifying terrain characteristics is challenging because similar terrains may have very different appearances (*e.g.*, carpet comes in many colors), while terrains with very similar appearance may have very different physical properties (*e.g.*, mulch versus dirt). In order to address the inherent ambiguity in vision-based terrain recognition and identification, we propose a multi-modal self-supervised learning technique that switches between audio features extracted from a microphone attached to the underside of a mobile platform and image features extracted by a camera on the platform to cluster terrain types. The terrain cluster labels are then used to train an image-based real-time CNN (Convolutional Neural Network) to predict terrain types changes. Through experiments, we demonstrate that the proposed self-supervised terrain type recognition method achieves over 80% accuracy, which greatly outperforms several baselines and suggests strong potential for assistive applications.

INDEX TERMS Ground robots, assistive application, self-supervised learning, CNN.

I. INTRODUCTION

Ground robots such as assistive robots (*e.g.*, navigation systems for the visually impaired) and ground exploration robots are often used in open-world environments and must be able to deal with many terrain types. Therefore, the ability to automatically recognize and identify new terrain characteristics is an important function for many applications. However, it is a highly challenging task to recognize terrain types robustly because similar terrains may have very different appearances (*e.g.*, carpet comes in many colors), while terrains with very similar appearance may have very different physical properties (*e.g.*, mulch versus dirt).

Due to the importance of terrain recognition, many vision-based terrain classification approaches have been proposed [14], [18], [25], [32]. Further, audio-based classification has been explored [8], [12], [23], [28], [35]. Besides audio and visual, some researchers have made efforts to recognize terrain types using vibration [1], [5], [9], [37] and tactile sensing [2], [33]. While these existing studies have proved that each modal is effective for recognizing terrain types, ambiguity remains in these methods using only a single sensing modality which may be noisy and may not be able to represent all changes in the terrain across different scenes. Therefore, we focus on an approach based on both audio and visual data, sensing modalities that are inexpensive, practical and easy to use.

We propose a multi-modal self-supervised learning technique that switches between audio features extracted from a microphone attached to a mobile platform's underside and image features extracted by a camera on the platform to cluster terrain types. In our method, we first recognize the characteristics of terrain types by audio-based clustering, which results in a discrete sequence of temporal segments. In order to reduce the noise of the features extracted over each temporal segment, *e.g.*, occlusions in the image or undesired environmental sounds in audio, we then compute average features for each modality within one temporal segment. Since the temporal segments generated by the audio-based clustering tend to over segment the temporal stream of information,

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

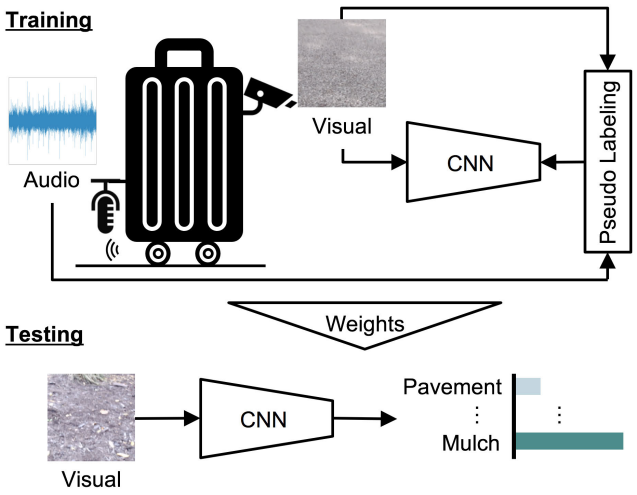


FIGURE 1. Overview of the proposed framework. The proposed method first generates pseudo-labels from audio recorded from a microphone attached to a mobile platform's underside and images captured by an RGB camera. These labels are utilized for training CNNs for terrain type classification in a self-supervised fashion. A more detailed sensor setup is shown in Figure 4. (Note: CNNs = Convolutional Neural Networks).

we implement the second phase of clustering with the averaged features to obtain temporal segments of a larger size. Since our eventual goal is to learn a vision-based terrain classifier, we use the second stage of clustering to assign pseudo labels to each image in each temporal segment. These labels enable us to train an image-based CNN to predict terrain types in a self-supervised fashion. With this scheme, the system can predict terrain types only with visual cues without building a preliminary dataset (see "Testing" in Figure 1), and thus having strong potential for assistive applications (*e.g.*, handheld terrain type notifier for the visually impaired)

We verify the proposed method on our dataset, where each terrain image and audio data is associated with terrain types. In this dataset, the friction sound's audio data is recorded with the super directional microphone heading toward the terrain and wheels. The RGB camera is mounted facing the front terrain. This dataset is available online and would be suitable for research of terrain type classification.

The contributions of this paper are as follow: (i) We present a self-supervised multi-modal clustering method that effectively uses the characteristics of both audio and visual cues to recognize novel terrain types. (ii) We prepare a free-to-use dataset containing labeled terrain images and labeled friction sounds between the terrain and the wheel. (iii) We demonstrate the effectiveness of the proposed clustering method and framework by training and testing a real-time CNN with several comparisons approaches.

II. RELATED WORK

Research for terrain type classification has grown with the development of autonomous driving and navigation systems, where some sensing modalities are utilized. This section describes related works in terms of terrain type recognition method, clustering method, and indoor navigation system.

A. TERRAIN TYPE RECOGNITION

1) VISION BASED

Howard and Seraji [14] presented a vision-based terrain classification method, where they mainly detect an edge of input images, extract a signature, and identify obstacles. Sung *et al.* [32] showed that features with spatial coordinates extracted using Daub2 wavelet in the HSI color space perform well on terrain type recognition. Other methods focus on analyzing terrain textures [25] in visual-spectrum images using Haar wavelet transforms to identify color and texture [18]. The classification accuracy of vision-based terrain recognition is directly affected by appearances, although similar appearances may have very different physical properties (*e.g.*, carpet versus rough concrete in Figure 5). Considering that the field of terrain recognition is vital to navigation solutions for the visually impaired, a more robust approach is desirable.

2) AUDIO BASED

Christie and Kottege [8] presented an audio-based terrain recognizing approach for legged robots using support vector machines (SVM) on audio features extracted during locomotion. Inspired by recent developments in deep neural networks (DNNs), some methods introduce DNNs into the framework of terrain type classifications, achieving high accuracy results [12], [23], [28], [35], [29], [34]. However, these methods generally need to collect a large-scale fully-labeled dataset for training in advance. Also, using a super directional microphone in testing is not practical because it is challenging for the users to set a microphone to obtain ground friction sounds without ambient noise.

3) VIBRATION BASED

Vibration is often a critical information source for recognizing terrain type. Brooks and Iagnemma [5] proposed a vibration-based classification approach, which deals with vibration data by using principal component analysis and linear discriminant analysis. Collins *et al.* [9]'s method classifies terrain types using input frequency responses, which assists autonomous ground vehicle navigation. The approach of Ward and Iagnemma [37] integrates vehicle speed and vibration data for training terrain type SVMs. Recently, Bai *et al.* [1] introduced an approach based on 3D vibrations induced in the rover structure by the wheel-terrain interaction.

4) LiDAR BASED

Due to the significant role of LiDAR sensors in autonomous driving, several methods perform terrain classification with LiDAR sensors for outdoor scenes. Vandapel *et al.* [36] and Lalond *et al.* [22] proposed a terrain classification method focusing on LiDAR point cloud segmentation. Some studies perform terrain classification by combining LiDAR point clouds and camera images [20], [21]. Differently from these approaches, our framework works with an off-the-shelf setup (*i.e.*, RGB camera and microphone) and performs terrain type recognition in both indoor and outdoor scenes.

5) TACTILE BASED

Tactile properties such as roughness and slipperiness also represent terrain characteristics and are used in terrain classification and material estimation tasks. Baishya and Bäuml [2] proposed a deep network-based material estimation method that focuses on a robot finger's tactile sense. The work of Takahashi and Tan [33] addresses the task of recognizing terrain types from visual and tactile sensors, where variational auto-encoders and recurrent neural networks are employed for feature extraction and estimation. As with the LiDAR-based methods, these methods are expensive in terms of introducing cost for tactile sensors.

6) SENSOR FUSION BASED

Sensor fusion techniques for terrain identification and classification in ground mobile robots have also been developed for many ground applications. Laible *et al*. [21] introduced a terrain classification approach by sensor fusion with LiDAR and camera for outdoor robots. Zürn *et al*. [41] proposed a self-supervised terrain segmentation technique with audio-visual information by an acoustic feature learning. However, these sensor fusion-based approaches require a very precise environment for data collection compared to the proposed method's simple setup (see "Testing" in Figure 1).

B. CLUSTERING

For analyzing features representing the target scene and captured data, clustering is a key component and is often applied in computer vision and robotics research. In addition to several traditional approaches, including K-means [24], EM (Expectation–Maximization) clustering [7], and spectral clustering [26], deep variational auto-encoder based clustering approach (VaDE) was proposed in recent years [15]. Further, their extensions for multi-source and cross-modal tasks have been proposed [3], [4], [6], [11], [15], [27], [30], [39], [40]. Contrary to these approaches, our method switches visual- and audio-features by taking noises in terrain features into account, *e.g.*, human legs in images and chatting in audio.

C. INDOOR/OUTDOOR ASSISTIVE SYSTEMS

In recent years, indoor/outdoor assistive systems have been actively developed with the improvement in depth sensors (*e.g.*,, Kinect and LiDAR) and global positioning systems (GPS). Kayukawa *et al*. [16] proposed a collision avoidance system for visually impaired people using both an RGB camera and a depth sensor. Wellhausen *et al*. suggested a self-supervised terrain predicting technique for autonomous navigation of the quadruped robots [38]. Terrain classification is also applied to agricultural fields for assisting agricultural tractors with LiDAR and GPS [19]. Our framework's applications would cover such indoor/outdoor assistive systems, including the extension for slipping and falling avoidance.

III. APPROACH

To realize self-supervised terrain type recognition, we need to perform clustering for labeling each frame (*i.e.*, frames within a same cluster will be assigned the same pseudo label). A central component of our proposed approach is multi-modal clustering, where we use audio-visual cues. Figure 2 shows an overview of the proposed framework. Given input audio and visual data, we extract features from each using a Variational Auto Encoder (VAE) (Section III-A). We then perform EM (Expectation–Maximization) clustering for proposing temporal segments which have the same terrain types, *i.e. sequence* proposal (Section III-B). Next, we average out noises of each feature within each *sequence* (Section III-C1) and compute affinities between each *sequence* (Section III-C2). Finally, we perform an agglomerative clustering based on the calculated affinities to obtain pseudo-labels for each image (Section III-C3).

A. FEATURE EXTRACTION

In this section, we describe the details of feature extraction for both audio and visual data. In this paper, audio and visual data represent the friction sound between the wheel and the terrain (recorded with super-directional microphone) and floor image (recorded with RGB camera), respectively. Figure 4 shows our setup of these sensors.

1) AUDIO

We empirically set the window size for audio features long enough to being robust to undesirable noises (2.8s in experiments). Raw audio data windows are thus too large to treat with neural networks directly, so first, we compress the data. Here, we use a simple audio feature descriptor: Mel-Frequency Cepstrum Coefficients (MFCCs) [10]. We first compute 26 MFCCs, where the step between successive windows is 30 fps (frame rate of RGB camera), the length of the analysis window is 2.8 seconds, and the fast fourier transform (FFT) size is 2^{16}. Then, we apply variational auto-encoder (VAE) feature extraction to 26 MFCCs to compute audio features according to a Gaussian distribution. Figure 3 (upper) shows the VAE network architecture, which mainly consists of fully connected layers. We follow the method of Kingma and Welling [17] for training the VAE. Through this processing, we obtain the latent vector $\{z_t^{\text{audio}} \mid t \in \mathbb{Z}_{\geq 1}\}$.

2) VISUAL

In order to obtain features from terrain appearances, we also extract visual latent vectors from a VAE, as shown in Figure 3 (lower). We resize the input image to 128×128 around the center of the image. By applying these resized images to VAE, we obtain the latent vector $\{z_t^{\text{visual}} \mid t \in \mathbb{Z}_{\geq 1}\}$. We train the VAE with the method of Kingma and Welling [17], as with audio features.

B. SEQUENCE DETECTION

Since clustering for all frames is noise sensitive, we perform clustering on a unit of multiple frames. To propose continuous frames that have the same terrain types, we perform clustering on audio features z_t^{audio}. Here, we employ EM clustering [7]

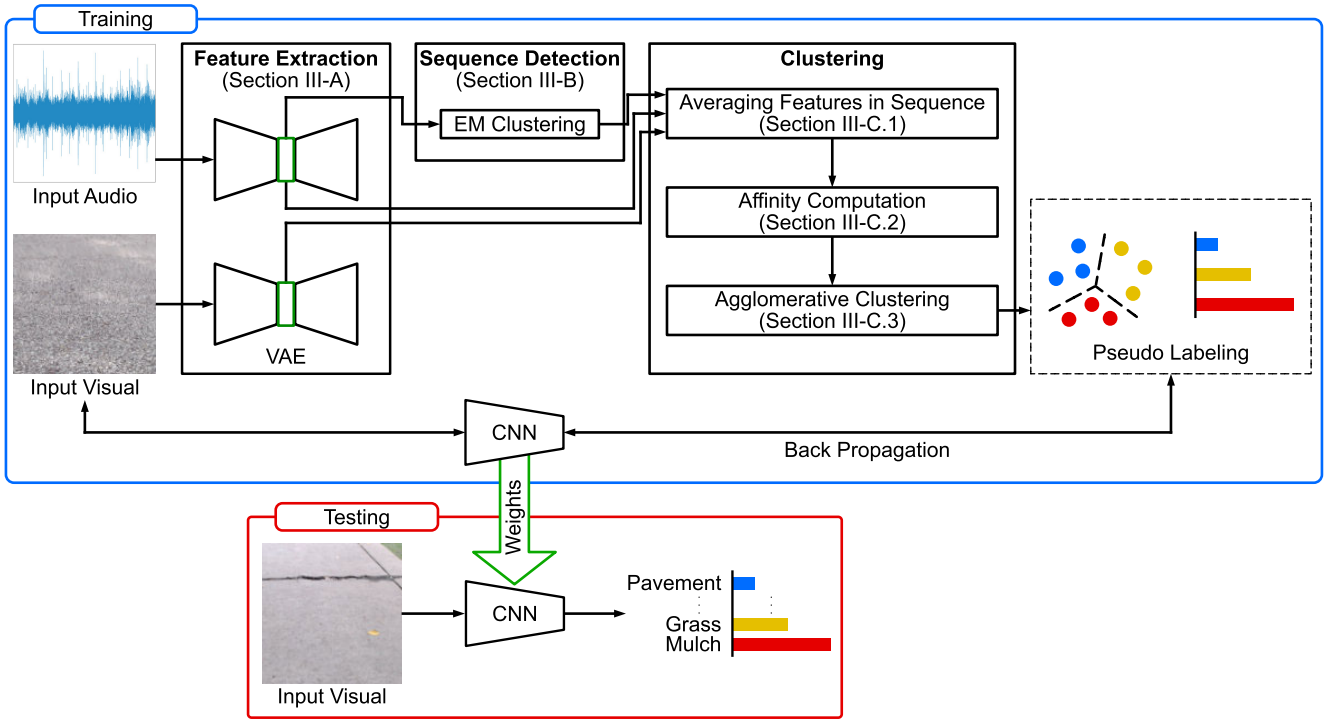


FIGURE 2. **The proposed framework. Our adaptive multi-modal clustering approach, including sequence detection and agglomerative clustering, utilizes audio and visual cues.**

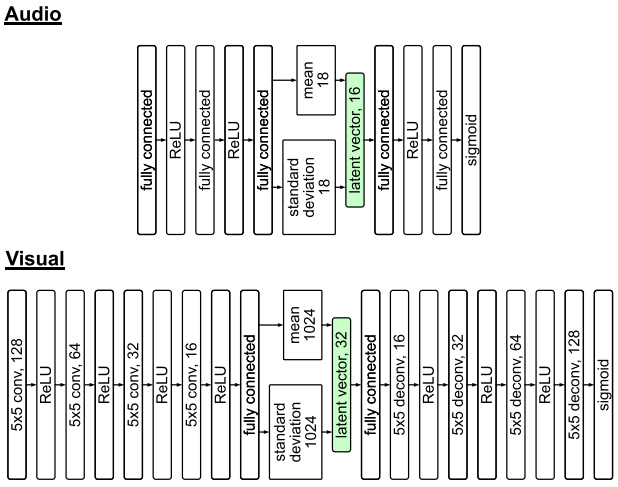


FIGURE 3. **Audio-Visual Feature Extraction. Audio and visual feature are extracted from latent spaces of VAEs.**

since audio features follow a Gaussian distribution after VAE-based feature extraction. We call a set of frames that continuously have the same clustering label *sequence* : S_i. Given the clustering label $\{C_t \mid t \in \mathbb{Z}_{\geq 1}\}$ on each frame t, the i-th *sequence* is defined as follows:

$$S_i = \{t_i \leq t < t_{i+1} \mid t_i, t_{i+1} \in B\},$$
$$B = \left\{0,\ t_i \mid C_{t_i - 1} \neq C_{t_i},\ t_i > 0,\ i \in \mathbb{Z}_{\geq 0}\right\}. \quad (1)$$

Here, B is a set of frames whose cluster changes after the frame.

FIGURE 4. **Sensor Setup. This figure illustrates the mounting positions of a super-directional microphone and RGB camera, surrounded by a green square. The microphone is mounted facing the terrain and wheels in order to record the friction sound clearly. The RGB camera is mounted facing the front terrain.**

C. CLUSTERING

Although audio-based clustering has the advantage of being sensitive to the terrain changes, it tends to over-segment frames by being affected by the change of grain and tile arrangement. The clustering method introduced in this

section merges the over segmented sequences by taking advantage of visual features.

The proposed multi-modal clustering consists of the following three processes: (i) Averaging audio-visual feature in a sequence; (ii) Affinity computation between audio features and visual features; and (iii) Agglomerative clustering. We describe the details of each processing step below.

1) AVERAGING FEATURES IN SEQUENCE

We first reduce external noises by averaging both audio- and visual-features within each sequence S_i. This averaging further enables us to extract audio- and visual-features for each sequence S_i and perform clustering in a unit of sequences, rather than frames. We define representative features of audio $\tilde{z}_i^{\text{audio}}$ and visual $\tilde{z}_i^{\text{visual}}$ of the sequence S_i as follows:

$$\begin{aligned}
\tilde{z}_i^{\text{audio}} &= \frac{1}{|Z_i^{\text{audio}}|} \sum_{z^{\text{audio}} \in Z_i^{\text{audio}}} z^{\text{audio}}, \\
Z_i^{\text{audio}} &= \left\{ z_t^{\text{audio}} \mid t \in S_i \right\}, \\
\tilde{z}_i^{\text{visual}} &= \frac{1}{|Z_i^{\text{visual}}|} \sum_{z^{\text{visual}} \in Z_i^{\text{visual}}} z^{\text{visual}}, \\
Z_i^{\text{visual}} &= \left\{ z_t^{\text{visual}} \mid t \in S_i \right\}, \qquad (2)
\end{aligned}$$

where Z_i^{audio} and Z_i^{visual} denote a set of audio and visual features in S_i.

2) AFFINITY COMPUTATION

In contrast to audio features, visual features do not tend to be affected by tile arrangement changes concerning wheel direction since visual features depend only on their appearances. Our method merges these over-segmented sequences by adaptively switching clustering cues from audio to visual by taking this advantage into account.

Since the noises on visual features are averaged out through the processing described in section III-C1, we switch these feature spaces by merely taking the minimum value of Euclidean distance between audio- and visual-features. The affinity between sequence S_i and S_j is defined as follows:

$$\begin{aligned}
&d(S_i, S_j) \\
&\quad = \min \left\{ \|\tilde{z}_i^{\text{audio}} - \tilde{z}_j^{\text{audio}}\|_2, \ \|\tilde{z}_i^{\text{visual}} - \tilde{z}_j^{\text{visual}}\|_2 \right\}. \qquad (3)
\end{aligned}$$

With this scheme, we are able to merge the sequences where their appearances are close enough. Further, by considering the distance of audio features, this simple strategy is able to handle the difficulty of terrain type recognition: similar terrains may have very different appearances (*e.g.*, carpet comes in many colors) but similar audio profiles.

3) AGGLOMERATIVE CLUSTERING

Finally, in order to obtain labels for each image, we perform agglomerative clustering on the affinity matrix whose element consists of $d(S_i, S_j)$. The clusters are directly utilized to generate pseudo labels for each sequence. Since the frames included in each sequence are known, we obtain labels for all frames by feeding back sequence labels to each frame.

TABLE 1. Dataset Detail. This table shows a number of frames and terrain classes of each scene in our dataset.

	No.	Scene	# frames (train/test)	Classes
Indoor	1	SH	10694 / 7206	Carpet Concrete flooring
	2	NSH	7041 / 7698	Tile Carpet Linoleum
	3	WH	9046 / 8208	Tile Carpet Linoleum
	4	GHC	7736 / 8397	Tile Carpet Concrete flooring Rough concrete
Outdoor	5	Garden	8113 / 6543	Asphalt Pavement Grass
	6	Playground	3822 / 10311	Pavement Grass
	7	Parking	8664 / 7093	Pavement Wood deck Mulch

IV. DATASET

In order to verify our audio-visual self-supervised terrain recognition method, we prepare a diverse terrain classification dataset for indoor/outdoor mobile platforms. This dataset is available online and would be suitable for research of terrain type classification. We record both audio and visual data simultaneously, where each frame is assigned to a terrain type label. Audio data of the friction sound is recorded with the super directional microphone facing the terrain and wheels. Visual data is captured by the RGB camera mounted facing the front terrain. In this section, we describe our sensor setup and the dataset structure in detail.

A. SENSOR SETUP

Figure 4 shows our sensor setup. We put a personal computer on the dolly and connected the RGB camera and super directional microphone to it. The sensors used are: a super directional microphone (Audio-Technica AT897 Line/Gradient Shotgun Condenser Microphone) and an RGB camera (Logitech C922x Pro Stream Webcam – Full 1080p HD Camera). Synchronized audio-visual data is collected by scanning the scene with this dolly. In addition, our outdoor dataset was taken during the daytime with good weather.

B. DATASET DETAIL

Table 1 shows the detail of our dataset. As shown in Figure 5, there are a whole ten classes of terrain types included in our dataset. Each scene comprises about 8000 frames for training and testing CNNs for terrain classification. We prepare test sequences for each scene. These test sequences cover all classes that training sequences include.

V. EXPERIMENT

To demonstrate the proposed method's ability to both recognize and identify terrain types, we experiment on our dataset.

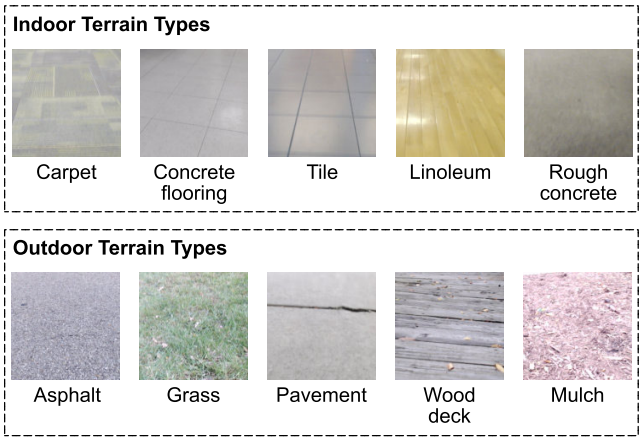


FIGURE 5. Terrain Types. This figure shows each terrain image example included in our dataset.

We first perform the proposed clustering method on each indoor/outdoor training scene and calculate the Normalized Mutual Information (NMI) to verify the effectiveness of the proposed method compared to other clustering approaches. After that, we train ResNet [13] using a set of input visuals linked with pseudo labels of terrain types, where we used SGD optimizer [31] to optimize the network parameters, with an initial learning rate 0.01 and 150 epochs. We then validate the trained CNN with test scenes in terms of prediction accuracy, precision, and recall values. ResNet [13], which achieves real-time processing, is employed for terrain type inference since the response time is one of the essential elements for mobile robots' navigation systems.

A. COMPARISON APPROACH

In order to verify the effectiveness of the proposed method, we experiment with comparison approaches. In this section, we verify the effectiveness (i) using multi-source (audio-visual) data; (ii) two step clustering (agglomerative clustering after sequence detection (EM clustering)); and (iii) with and without our feature switching scheme.

1) SINGLE SOURCE CLUSTERING

For verifying the effectiveness of multi-source (audio-visual) data, we first test single source approaches, which directly performs EM clustering on z_t^{audio} and z_t^{visual}. These comparisons reveal that a single source tends to be affected by the input noise (visual-only) and over-segmentation (audio-only), compared with multi-source clustering approaches.

2) MULTI SOURCE CLUSTERING

In addition to multi-source, the proposed method employs sequence-based clustering, not frame-based. Hence, we reveal the effectiveness of this processing by comparing with simple multi-source clustering, which performs EM clustering on features concatenating z^{audio} and z^{visual}, which we call Audio-Visual clustering. Additionally, in order to verify the effectiveness of our feature switching scheme (mentioned in Section III-C2), we compare our method with the method of clustering on features concatenating $\tilde{z}_i^{\text{audio}}$ and $\tilde{z}_i^{\text{visual}}$, which does not switch feature space but uses both audio and visual.

3) DEEP NETWORK BASED CLUSTERING

As mentioned in Section II, deep network-based clustering methods have been developed. In our experiment, we employ a state-of-the-art deep network-based clustering approach: VaDE [15] as a representative method. We perform VaDE [15] on z_t^{audio}, z_t^{visual}, and features concatenating z_t^{audio} and z_t^{visual}.

B. CNN TRAINING

To evaluate the proposed framework's practicality, we train ResNet50 [13] using our dataset with pseudo labeling based on the output of each scene's proposed clustering method. Through our experiments, the resolution of input images is 128×128.

C. RESULTS

In this section, we experimentally demonstrate the performance of the proposed self-supervised multi-modal terrain type recognition method on test scenes of our dataset. In order to generate pseudo labels for training a CNN, we perform the proposed clustering method on all training scenes. After that, we train the CNN, ResNet50 [13], with the pair of pseudo labels and images, and then test on all test scenes. Through this experiment, we demonstrate the performance of (i) the proposed clustering method by comparing our approach with several baselines in terms of NMI; and (ii) terrain type prediction trained with the proposed framework by measuring accuracy, precision, and recall values of the trained CNN.

1) CLUSTERING

We first demonstrate and analyze the performance of the proposed clustering method, quantitatively and qualitatively. For quantitative comparison, we measure NMI using the proposed training dataset. Table 2 and Table 3 show the results. In Table 2, we compare the proposed method with two single source clustering approaches, where Audio-only and Visual-only features are used for EM clustering, and two multi-source clustering approaches, where Audio-Visual features are used for EM clustering and a state-of-the-art deep clustering method (VaDE). The proposed method outperforms all comparison approaches, with an average accuracy of over 80%. Compared to the Visual-only approach, Audio-only can cluster terrain more accurately, which shows that audio features are more robust to noise than visual features by setting window size long to reduce undesirable noises. We next compare single source clustering (Visual-only and Audio-only) with multi-source clustering (Ours, Audio-Visual, and Audio-Visual VaDE). Regarding Visual-only as a criterion, the accuracy of Audio-Visual is improved, while Audio-Visual does not outperform Audio-only. This shows the

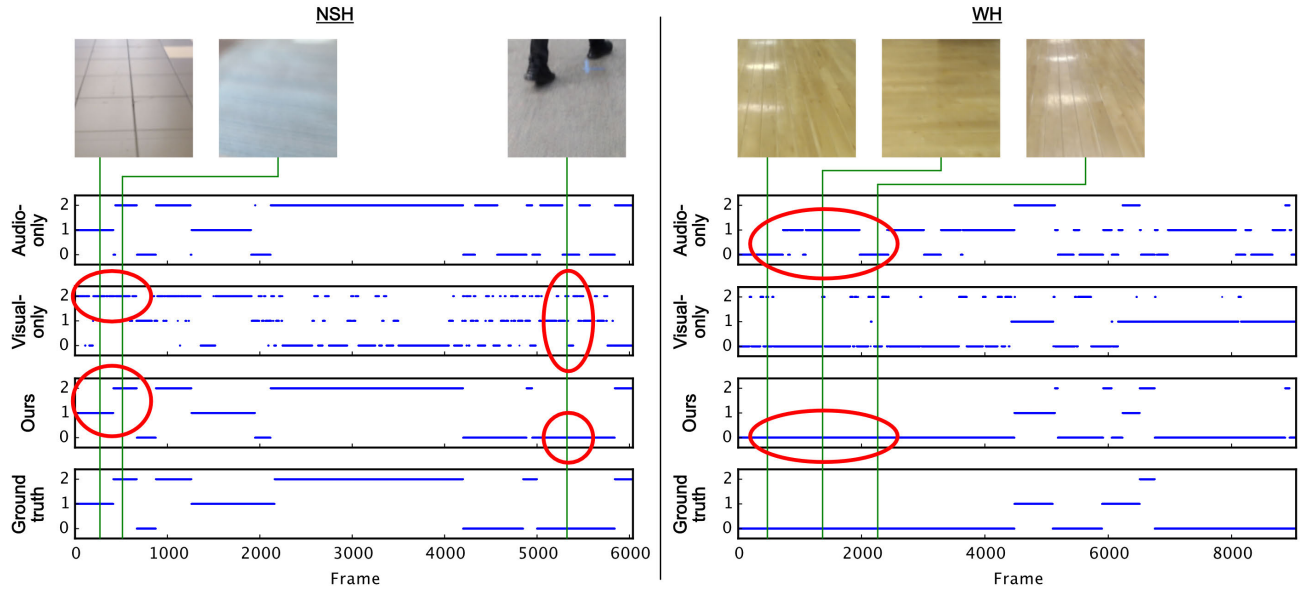


FIGURE 6. **Qualitative clustering comparison of clustering. In this comparison, we demonstrate the effectiveness of switching audio-visual features. In the NSH scene (left), we focus on the comparison with Visual-only clustering. It tends to be affected by terrain appearance (color and texture) and noise such as human feet or wall, which is circled with a red circle in the figure. In the WH scene (right), we focus on comparing with Audio-only clustering, where it tends to be over-segmented when the grain and tile arrangement changes with respect to the wheel.**

TABLE 2. **Quantitative comparison. Single source (Audio-only EM and Visual-only EM), multi-source (Audio-Visual EM), and deep clustering (Audio-Visual VaDE) versus ours.**

No.	Ours	Visual-only EM [7]	Audio-only EM [7]	Audio-Visual EM [7]	Audio-Visual VaDE [15]
1	**88.9**	3.1	82.4	1.8	5.7
2	**81.9**	14.2	56.6	14.3	54.2
3	**64.9**	12.3	31.7	10.0	19.3
4	**94.3**	36.2	90.1	48.9	69.1
5	**90.7**	36.3	90.7	63.3	76.8
6	**92.2**	48.6	88.6	83.9	77.2
7	**54.1**	21.3	39.7	30.3	30.4
Total	**81.0**	24.6	68.5	36.1	50.6

TABLE 3. **Ablation study on effects of clustering approaches and feature switching.**

Feature	Clustering	Accuracy
Audio	K-means [24]	63.7
	EM [7]	68.5
	VaDE [15]	56.9
Visual	K-means [24]	22.1
	EM [7]	24.6
	VaDE [15]	21.7
Audio-Visual	K-means [24]	33.3
	EM [7]	36.1
	VaDE [15]	45.2
Ours	w/o feature switching (eq. 3)	50.6
	w/ feature switching (eq. 3)	**81.0**

importance of using multi-source data for clustering and the effectiveness of the proposed method's switching technique.

Table 3 shows a comparison between applied clustering algorithms, including K-means [24], EM [7], and VaDE [15]. The results suggest that EM clustering is superior to K-means clustering. This is because extracted features follow a Gaussian distribution in the latent space. In our method, we measure NMI in both our proposal (w/ feature switching) and a different approach, which concatenates z^{audio} and z^{visual} instead of switching features (w/o feature switching). The results show that our proposed switching system greatly contributes to highly accurate clustering.

Figure 6 qualitatively shows two clustering results on two scenes, where Audio-only, Visual-only, and Ground truth are presented. Focusing on the red circles in the NSH scene (left), we observe that visual features are sensitive to noise (human feet) and highly dependent on terrain appearance. In the WH scene (right), Audio-only tends to be over-segmented because the floor grain changes with respect to the wheel (*i.e.*, from vertical to parallel), while the proposed method is much accurate by switching the clustering cue to visuals. These qualitative results verify that the proposed switching scheme is able to utilize multi-source and solve the problem of Audio-only and Visual-only approaches.

2) PREDICTION

In Table 4, we present the quantitative evaluation of the terrain type prediction in terms of precision, recall, f1-score, and accuracy on the proposed test scenes. Our method's average accuracy is over 85%, demonstrating the practicality of the proposed framework through all scenes. As we experiment on both indoor/outdoor scenes, our analysis suggests that the proposed framework can be used in applications in diverse scenes. Further, as we achieved much high accuracy (over 85% in total), it could be argued that our framework is able even to handle delicate tasks such as assistive systems.

Figure 7 presents the qualitative results of CNN predictions on terrain images. Since the pseudo-labels used for CNN training are based on multi-source clustering, it is verified

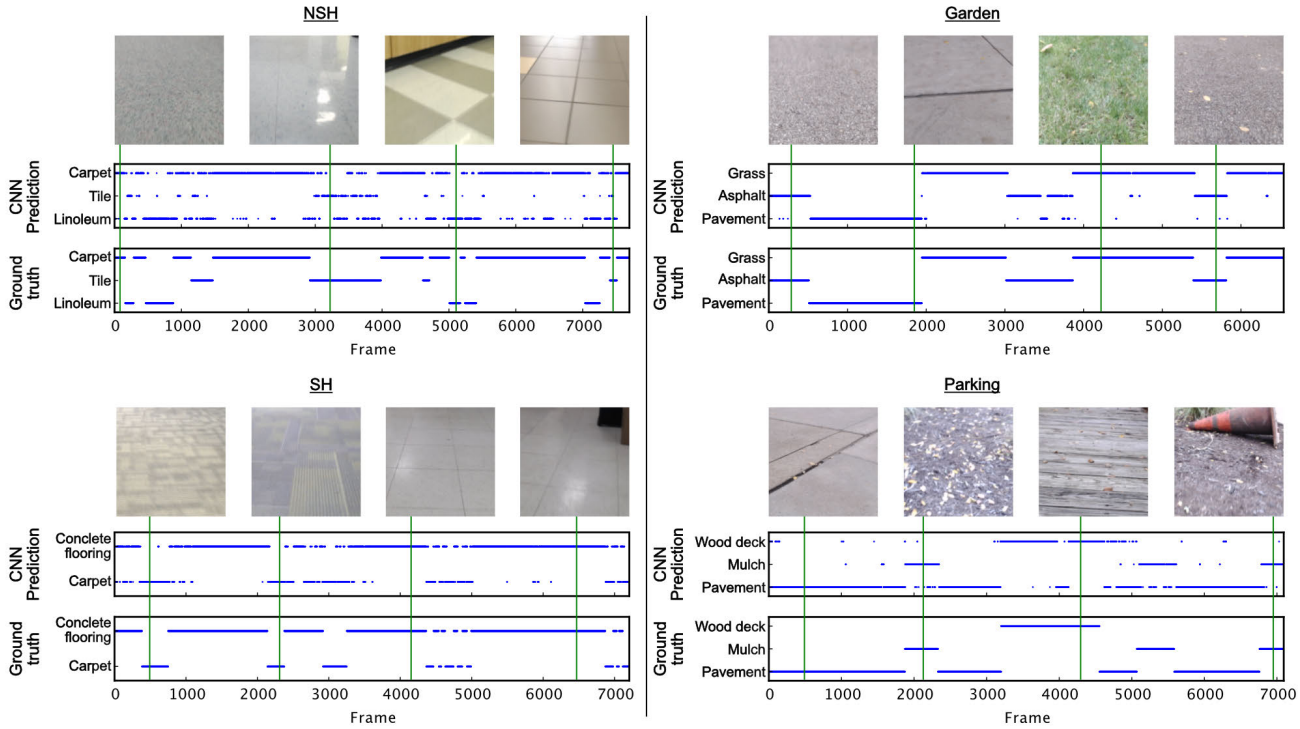


FIGURE 7. **Qualitative comparison of terrain type predictions. The results of CNN prediction and ground truth label are visualized with blue lines. We demonstrate that CNN correctly predicts each terrain type, although the input images have a similar color or texture. This is because pseudo labels used for training the CNN are based on adaptive switching multi-source clustering.**

TABLE 4. **Quantitative evaluation of terrain type predictions, in terms of precision, recall, F1-score, and accuracy.**

No.	Classes	Precision	Recall	F1-score	Accuracy
1	Carpet	65.4	87.0	74.6	87.3
	Concrete flooring	96.1	87.4	91.5	
2	Tile	80.1	37.7	51.3	74.2
	Carpet	88.5	84.3	86.3	
	Linoleum	40.8	80.8	54.2	
3	Tile	63.9	37.8	47.5	88.3
	Carpet	46.5	68.7	55.4	
	Linoleum	92.1	95.7	93.9	
4	Tile	17.0	27.7	21.1	73.6
	Carpet	99.6	71.5	83.2	
	Concrete flooring	56.6	89.3	69.3	
	Rough concrete	92.8	68.4	78.7	
5	Asphalt	95.5	89.7	92.5	95.7
	Pavement	89.8	98.7	94.1	
	Grass	98.7	97.7	98.2	
6	Pavement	92.5	98.4	95.6	95.5
	Grass	98.5	92.9	95.6	
7	Pavement	91.7	91.0	91.4	89.2
	Wood deck	92.7	84.3	88.3	
	Mulch	78.9	87.9	83.2	

that terrain type can be recognized correctly even if terrain appearances are similar.

VI. LIMITATION AND FUTURE WORK

While our outdoor dataset was taken during the daytime with good weather, the system needs to work correctly even in rainy and dark conditions, considering a real scenario. These adverse environments are expected to cause slipping, skidding, and posture instability and significantly affect training and testing accuracy since our method generates pseudo-labels based on audio and visual features. In the future, it is necessary to verify the proposed method's limitations and develop an algorithm that can even work robustly under such situations.

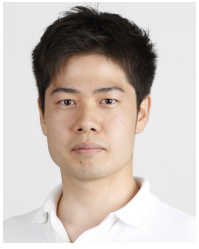
In addition, as the proposed method's extension for more practical applications, we would like to consider a new terrain type discovery approach that can recognize unknown terrain types not included in the training dataset. The incremental classification may also be a useful extension of the proposed method to reduce the prepossessing cost. Besides, since recognizing multiple terrain types at the same frame may be essential for detecting terrain types' changes to a real scenario, the proposed method has a sufficient potential to be extended to be such an application by utilizing the terrain class probabilities of CNN's final layers.

VII. CONCLUSION

Towards the development of ground assistive robots, we present a novel self-supervised multi-modal terrain classification method, CNN based framework, and terrain diverse dataset. We demonstrate that the proposed clustering method is able to cluster terrain by switching between audio and visual features adaptively. Further, the proposed framework's practicality is verified by reporting the accuracy of terrain type classification with a CNN, ResNet50, which is trained through pseudo labels generated by the proposed clustering method.

REFERENCES

[1] C. Bai, J. Guo, and H. Zheng, "Three-dimensional vibration-based terrain classification for mobile robots," *IEEE Access*, vol. 7, pp. 63485–63492, 2019.
[2] S. S. Baishya and B. Bauml, "Robust material classification with a tactile skin using deep learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 8–15.
[3] A. Miguel Bautista, A. Sanakoyeu, E. Tikhoncheva, and B. Ommer, "CliqueCNN: Deep unsupervised exemplar learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3846–3854.
[4] P. Bojanowski and A. Joulin, "Unsupervised learning by predicting noise," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 517–526.
[5] C. A. Brooks and K. Iagnemma, "Vibration-based terrain classification for planetary exploration rovers," *IEEE Trans. Robot.*, vol. 21, no. 6, pp. 1185–1191, Dec. 2005.
[6] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 132–149.
[7] G. Celeux and G. Govaert, "A classification EM algorithm for clustering and two stochastic versions," *Comput. Statist. Data Anal.*, vol. 14, no. 3, pp. 315–332, Oct. 1992.
[8] J. Christie and N. Kottege, "Acoustics based terrain classification for legged robots," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 3596–3603.
[9] E. G. Collins and E. J. Coyle, "Vibration-based terrain classification using surface profile input frequency responses," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2008, pp. 3276–3283.
[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.
[11] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2051–2060.
[12] R. Hadsell, S. Samarasekera, and A. Divakaran, "Audio based robot control and navigation," U. S. Patent 8 532 863, Sep. 10, 2013.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[14] A. Howard and H. Seraji, "Vision-based terrain characterization and traversability assessment," *J. Robotic Syst.*, vol. 18, no. 10, pp. 577–587, 2001.
[15] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1965–1972.
[16] S. Kayukawa, K. Higuchi, J. Guerreiro, S. Morishima, Y. Sato, K. Kitani, and C. Asakawa, "Bbeep: A sonic collision avoidance system for blind travellers and nearby pedestrians," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, p. 52.
[17] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114
[18] N. Kingry, M. Jung, E. Derse, and R. Dai, "Vision-based terrain classification and solar irradiance mapping for solar-powered robotics," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 5834–5840.
[19] M. Kragh, N. R. Jørgensen, and H. Pedersen, "Object detection and terrain classification in agricultural fields using 3D lidar data," in *Proc. Int. Conf. Comput. Vis. Syst.*, 2015, pp. 188–197.
[20] S. Laible, Y. N. Khan, K. Bohlmann, and A. Zell, "3D LIDAR-and camera-based terrain classification under different lighting conditions," in *Autonomous Mobile Systems*. Springer, 2012, pp. 21–29.
[21] S. Laible, Y. N. Khan, and A. Zell, "Terrain classification with conditional random fields on fused 3D LIDAR and camera data," in *Proc. Eur. Conf. Mobile Robots*, Sep. 2013, pp. 172–177.
[22] J.-F. Lalonde, N. Vandapel, D. F. Huber, and M. Hebert, "Natural terrain classification using three-dimensional ladar data for ground robot mobility," *J. Field Robot.*, vol. 23, no. 10, pp. 839–861, 2006.
[23] J. Libby and A. J. Stentz, "Using sound to classify vehicle-terrain interactions in outdoor environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 3559–3566.
[24] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.
[25] P. Mathur and K. S. Pandian, "Terrain classification for traversability analysis for autonomous robot navigation in unknown natural terrain," *Int. J. Eng. Sci. Technol. (IJEST)*, vol. 4, no. 1, pp. 38–49, 2012.
[26] Y. Andrew Ng, I. Michael Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.
[27] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 69–84.
[28] L. Ojeda, J. Borenstein, G. Witus, and R. Karlsen, "Terrain characterization and classification with a mobile robot," *J. Field Robot.*, vol. 23, no. 2, pp. 103–122, 2006.
[29] M. Prágr, P. Cizek, J. Bayer, and J. Faigl, "Online incremental learning of the terrain traversal cost in autonomous exploration," in *Robotics: Science and Systems*. 2019.
[30] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434
[31] H. Robbins and S. Monro, "A stochastic approximation method," in *The Annals of Mathematical Statistics*. JSTOR, 1951, pp. 400–407.
[32] G.-Y. Sung, D.-M. Kwak, and J. Lyou, "Neural network based terrain classification using wavelet features," *J. Intell. Robotic Syst.*, vol. 59, nos. 3–4, pp. 269–281, Sep. 2010.
[33] K. Takahashi and J. Tan, "Deep visuo-tactile learning: Estimation of tactile properties from images," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8951–8957.
[34] A. Valada and W. Burgard, "Deep spatiotemporal models for robust proprioceptive terrain classification," *Int. J. Robot. Res.*, vol. 36, nos. 13–14, pp. 1521–1539, Dec. 2017.
[35] A. Valada, L. Spinello, and W. Burgard, "Deep feature learning for acoustics-based terrain classification," in *Robotics Research*. Springer, 2018, pp. 21–37.
[36] N. Vandapel, D. F. Huber, A. Kapuria, and M. Hebert, "Natural terrain classification using 3-d ladar data," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Apr./May 2004, pp. 5117–5122.
[37] C. C. Ward and K. Iagnemma, "Speed-independent vibration-based terrain classification for passenger vehicles," *Vehicle Syst. Dyn.*, vol. 47, no. 9, pp. 1095–1113, Sep. 2009.
[38] L. Wellhausen, A. Dosovitskiy, R. Ranftl, K. Walas, C. Cadena, and M. Hutter, "Where should i walk? Predicting terrain properties from images via self-supervised learning," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1509–1516, Apr. 2019.
[39] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 478–487.
[40] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5147–5156.
[41] J. Zürn, W. Burgard, and A. Valada, "Self-supervised visual terrain classification from unsupervised acoustic feature learning," *IEEE Trans. Robot.*, early access, Nov. 3, 2020, doi: 10.1109/TRO.2020.3031214.



AKIYOSHI KUROBE (Member, IEEE) received the B.E. and M.Sc.Eng. degrees in information and computer science from Keio University, Japan, in 2017 and 2018, respectively, where he is currently pursuing the Ph.D. degree in science and technology. His research interests include computer vision, robotics, and multimodal processing.

YOSHIKATSU NAKAJIMA (Member, IEEE) received the Ph.D. degree from Keio University, in 2020. He was with the Research and Development Center, Sony Corporation, in 2020. His research interests include computer vision, 3D/RGB-D perception, and mixed reality.

KRIS KITANI (Member, IEEE) received the B.S. degree from the University of Southern California and the M.S. and Ph.D. degrees from The University of Tokyo. He is currently an Associate Research Professor and the Director of the M.S. in Computer Vision program with the Robotics Institute, Carnegie Mellon University. His research projects span the areas of computer vision, machine learning, and human–computer interaction. His research interests include intersection of first-person vision, human activity modeling, and inverse reinforcement learning. His work has been awarded the Marr Prize honorable mention at ICCV 2017, the Best Paper Honorable Mention at CHI 2017 and CHI 2020, the Best Paper at W4A 2017 and 2019, the Best Application Paper ACCV 2014, and the Best Paper Honorable Mention ECCV 2012.

HIDEO SAITO (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Keio University, Japan, in 1992. Since 1992, he has been with the Faculty of Science and Technology, Keio University. From 1997 to 1999, he was as a Visiting Researcher with the Virtualized Reality Project, Robotics Institute, Carnegie Mellon University. Since 2006, he has been a Full Professor with the Department of Information and Computer Science, Keio University. His research interests include computer vision and pattern recognition, and their applications to augmented reality, virtual reality, and human–robotic interaction. His recent activities in academic conferences include being the Program Chair of ACCV 2014, the General Chair of ISMAR 2015, the Program Chair of ISMAR 2016, and the Program Chair of EuroVR2020. He is a Fellow of the Institute of Electronics, Information and Communication Engineers and the Virtual Reality Society of Japan.