2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Single-modal Incremental Terrain Clustering from Self-Supervised Audio-Visual Feature Learning

Reina Ishikawa, Ryo Hachiuma, Akiyoshi Kurobe, and Hideo Saito
Department of Information and Computer Science, Keio University, Yokohama, Japan
{reina-ishikawa, ryo-hachiuma, kurobe.akiyoshi, hs}@keio.jp

***Abstract*—The key to an accurate understanding of terrain is to extract the informative features from the multi-modal data obtained from different devices. Sensors, such as RGB cameras, depth sensors, vibration sensors, and microphones, are used as the multi-modal data. Many studies have explored ways to use them, especially in the robotics field. Some papers have successfully introduced single-modal or multi-modal methods. However, in practice, robots can be faced with extreme conditions; microphones do not work well in crowded scenes, and an RGB camera cannot capture terrains well in the dark. In this paper, we present a novel framework using the multi-modal variational autoencoder and the Gaussian mixture model clustering algorithm on image data and audio data for terrain type clustering. Our method enables the terrain type clustering even if one of the modalities (either image or audio) is missing at the test-time. We evaluated the clustering accuracy with a conventional multi-modal terrain type clustering method and we conducted ablation studies to show the effectiveness of our approach.**

I. Introduction

An understanding of ground terrain in open-world environments using a camera is a popular computer vision research area because of its widespread applications in robotics and automatic vehicular control [1]–[6]. In the field of autonomous driving [3], ground terrain classification is very important because certain types of terrain may negatively affect a robot's movement. Similarly, information about the surrounding terrain may help a robot modify its course of action during autonomous navigation [2], [6]. In the field of assistive robotics, the robot can warn a visually impaired person of potential danger concerning the ground type [7], [8].

Gaining an understanding of terrain types from visual input is a highly challenging task because the terrain of the same type can vary in appearance, while different terrain types may be very similar in appearance. To address the inherent ambiguity of vision-based terrain classification, other modality-based classification methods, such as audio-based [9]–[12], tactile-based [13]–[15] or vibration-based [16], [17], methods have been proposed. For the audio-based terrain classification method, audio sensors measure terrain properties through the interaction between the robot's wheel with its and. While these conventional studies have proved that each modality is effective for identifying terrain type, methods using only a single modality remain ambiguous because they may include noise and may not be able to keep up with changes in buildings and scenes. Recent works on multi-modal learning have demonstrated robust complementary features that yield superior performance in many learning tasks [18]–[20]. We follow the paradigm of multi-modal learning, leveraging two diverse modalities, sound and vision, to learn features for identifying robust terrain.

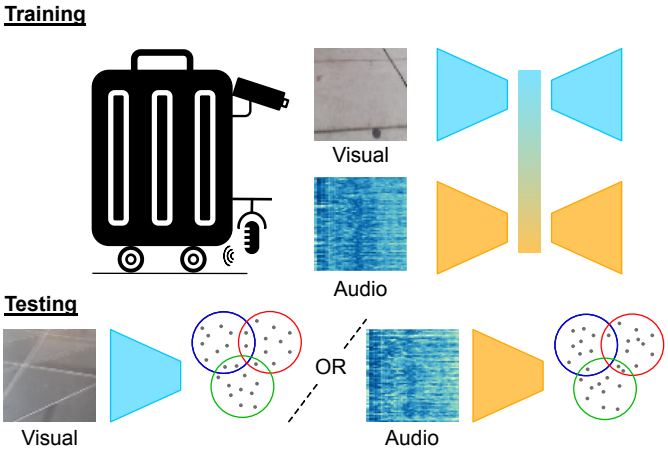

Fig. 1: Overview of our terrain clustering framework. We train the model to extract the features from audio-visual data in a self-supervised manner. At the testing, we assume that only a single modality (either audio or visual) can be accessed due to the extreme conditions, the obtained data is incrementally clustered into terrain types.

Several methods have been proposed for multi-modal terrain type classification [5], [6], [21]. Otsu *et al*. [5] proposed a method that classifies the terrain type based on image and vibration data. Zurn *et al*. [6] presented a self-supervised visual terrain discovery method learned from the audio data, which semantically labels the terrain type to facilitate robot navigation. Kurobe *et al*. [21] proposed a multi-modal self-supervised learning scheme that extracts audio and image features to cluster terrain types. Then the terrain cluster labels are used to train image-based convolutional neural networks (CNNs) to predict the terrain types. However, several problems in deploying the method as the terrain type discovery application the real-world scenarios.

First, the data from multiple modal sensors are not always useful. For example, it is difficult to capture informative features from input data signals in extreme conditions. In the case of visual sensors, extreme conditions are those with low illumination conditions, such as those during the night [22],

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

as they make it difficult to capture the textures of terrain in images. In the case of audio sensors, crowding is an extreme condition, as it is difficult to capture the sounds of vehicle—terrain interaction sound. The prediction model should be capable of handling these extreme conditions in real-world scenarios. However, the conventional multi-modal prediction method [5], [6], [21] allows the model to predict only a single input source, such as from only images [6], [21]. In this paper, we aim to tackle the problem that one of the modalities of gathering data is non-informative while also aiming to solve the terrain type identification task. Specifically, when training the multi-modal feature learning model, the audio and visual data are captured under normal conditions. When testing the model, only one modality (either image or audio) is available for prediction. Note that we do not aim to detect whether the data are captured in extreme conditions, but we assume that one input modality is missing at the time of the test.

The second problem is that when the model is applied to robotics, the robot captures the data sequentially. As the robot encounters data, the clustering model should not only predict the clustered index of the data but also progressively update the prediction model to achieve an accurate prediction. Moreover, if the clustering model is updated while it is running, it gains the potential to cluster terrain types that were not seen during its training.

Moreover, most state-of-the-art classification methods require a significant number of data samples, which can be arduous to obtain in supervised learning settings, where labels must be manually assigned to data samples [23], [24]. In contrast, self-supervised learning enables the automatic labeling of training data by exploiting the correlations between different input signals, thereby greatly reducing the amount of manual labeling work [25].

In this paper, we present a multi-modal self-supervised learning framework for terrain type clustering, which can access a single modality at the test time and which incrementally updates the clustering model. The overview of this work is depicted in Fig. 1. To the best of our knowledge, no work has been proposed that addresses these two important problems for deploying such models in the real-world applications. To handle the missing modality at test time, we employ a multi-modal variational autoencoder (MVAE) [26]. This method has a new training paradigm that learns using joint distribution and is robust against missing data. As the latent variables are sampled from multivariate Gaussian distribution, we employ incremental Gaussian mixture models (IGMM) for the incremental clustering. Moreover, we present a novel input data pre-processing method to effectively extract the features from audio and image data that are informative for terrain type clustering. We transformed audio data into time—frequency representations of cochleograms, and we extracted the edges of image data and encoded both edge images and raw images. We evaluated our terrain type clustering framework with the dataset collected by Kurobe *et al*. [21]. We also evaluated the clustering accuracy with the conventional multi-modal terrain type clustering model, and we conducted extensive ablation studies to show the effectiveness of our approach.

In summary, these are the major contributions of this work:

- To the best of our knowledge, we are the first to present a novel single-modal incremental terrain clustering framework learned in a self-supervised manner from multi-modal audio-visual data.
- Our method combines an MVAE and an IGMM for terrain type clustering. Using the IGMM clustering algorithm allows for the incremental clustering of terrains and updates the Gaussian mixture model during test-time.
- We present an input data pre-processing method for generating the informative latent variables for terrain type clustering.
- We also evaluated the clustering accuracy using conventional multi-modal terrain type clustering, and we conducted extensive ablation studies to show the effectiveness of our approach.

II. Related Work

Our work is uniquely positioned in the context of research using the MVAE model and IGMM for self-supervised multi-modal learning on terrain clustering. Our proposed model, shown in Fig. 1, undergoes multi-modal training with both image and audio, but the MVAE architecture enables the single modal inference of either image data or audio data upon testing.

A. Terrain type understanding

An understanding of terrain types is essential in the path planning of system for autonomous robots' navigation systems [2], [6], [27], [28] because the condition of the terrain can affect the robot's stable and safe running. In this subsection, we introduce the conventional single-modal and multi-modal methods of identifying terrain types.

1) Single-modal based: Many papers have proposed methods of understanding terrain type, including audio [9]–[12], tactile [13]–[15], vibration [16], [17], vision [2], [29]–[31].

Some classification approaches are vision-based: stereo-based [2], [29], feature-based [31], and spectral-based [30] approaches. However, vision-based classification is very sensitive to brightness or reflections. In recent studies, it is more common to use visual data alongside other data in self-supervised or multi-modal methods, as we show in the following subsection. As audio-based clustering is not affected by light conditions, it has been highly investigated in the field of robotics, especially legged robots [9] and vehicle robots [10]–[12].

2) Multi-modal based: Multi-modal methods of terrain type recognition have been actively investigated for the last few years [5], [6], [21], [32]. Zürn *et al*. [6] proposed an audio-visual-based, self-supervised terrain type semantic segmentation method. This method relates image features and audio features, which are extracted from different models, by reflecting the results of classifying image features onto the triplet loss of the audio data. This model is self-supervised, but, because the image feature extractor and audio feature

extractor are completely independent, they do not adapt to situations lacking either image data or audio data at test-time. In contrast, our method assumes such extreme conditions will occur and uses one strongly combined architecture.

Kurobe *et al*. [21] performed feature extraction from images and audio with two independent variational autoencoders. Their method and our method differ in terms of their purpose settings. The final goal of Kurobe *et al*.was terrain type identification from only images using a convolutional neural network (CNN) which is trained by both image and audio data. Therefore, the number of categories is fixed in the training process. In contrast, our goal is not the identification but the clustering of terrain features through an MVAE: relatively similar terrain types are assigned to the same class and different ones are assigned to different classes. Because of this fundamental difference in purpose, our method can potentially be applied to unknown terrains at test-time. Besides, the advantage of using both image and audio in [21] is limited to improving the accuracy of test-time clustering. In our method, all sensing data is available for use at test-time in a single-modal way or combined with any other sensing data in a multi-modal way, depending on the testing environment.

Natalia *et al*. [32] proposed a multi-modal fusion strategy for the gesture and near-range action recognition. Their random dropping of separate channels (ModDrop) shares some similarities with ours in allowing the model to obtain arbitrary combinations of the modalities.

Similarly to those two methods [6], [21], our method is trained from both image and audio data to leverage the multi-modal information. These two methods aim to learn the informative features from audio-visual data at the training time, and they predict the terrain types from only the image data. Our method also aims to extract the informative feature from the audio-visual data. However, at the test-time, our method predicts the terrain types from *either image or audio* data to handle the extreme capturing conditions in the real environment.

B. Self-supervised learning

Many approaches make use of multi-modal data, using one modality to supervise the training of another modality. Wellhausen *et al*. [33] proposed a navigation system using RGB-based semantic segmentation in a fully automated, self-supervised way. Books *et al*. [34] adopted vibration-based classification using a support vector machine (SVM) classifier to supervise vision-based classification. Zürn *et al*.and Kurobe *et al*. [6] [21] successfully related image data and audio data to identify limited numbers of terrain types in a self-supervised framework. In this paper, we consider our method to be self-supervised in that the result of a single-modal test is improved by another modality through multi-modal training.

C. Input preprocessing for audio data

As 44100 Hz or 48000 Hz are major sample rates of digital audio data, and a certain length is required for robustness against noises, the number of dimensions of audio input can increase enormously. To handle this problem, some studies have preprocessed audio data. For instance, Libby *et al*. [10] performed both time- and frequency-domain analyses as the preprocessing of the raw audio signal. Ojeda *et al*. [11] computed the discrete Fourier transforms (DFT) of sound data and use them as the input of neural networks (NNs). Valada *et al*. [12] used an short-time Fourier transform (STFT) based log scale spectrogram for audio pre-processing. At present, it is more common to transfer audio into frequency-domain feature extraction, but it is difficult to capture terrain features in just one small window size; we therefore adopted time-frequency domain approach by splitting a sound segment of 2.8 seconds into 64 time-domain pieces and then splitting each of those pieces into 64 frequency-domain pieces.

III. Method

In this section we explain the composition of our learning framework in Fig. 2: preprocessing of input audio and image data in Sec.III-A, self-supervised feature learning using a MVAE [26] in Sec.III-C, feature clustering using an IGMM in Sec.III-D. Sec.III-B provides an explanation of the VAE that is preliminary to the MVAE. At test-time, a single modality (either image or audio) can be accessed, and the feature vector is extracted using the VAE's encoder. Then, we incrementally update the GMM model.

A. Input preprocessing

1) Image: To encode the image into the latent vector using VAE, we take edge images in addition to RGB images to include edge information explicitly in the latent vector. Autoencoders and VAEs trained with Euclidean distance loss are known to produce blurry images [35]–[37]. This is a crucial problem for the terrain type clustering task because edge information is an important cue for clustering the terrain types. For instance, it is difficult to distinguish a gray **T**ile image from a gray **C**arpet image without the edge information. Instead of using a generative adversarial networks (GANs) scheme [37], [38] to reconstruct sharp images, we explicitly encoded edge images separately from the RGB images to encode edge information into a latent vector. Specifically, we applied a Laplacian filter to generate the denoised edge images. We first applied Gaussian filtering with a 5×5 kernel and then a four-neighbor Laplacian filter with a 5×5 kernel to the raw RGB image.

2) Audio: A one-dimensional raw audio signal is transformed into two-dimensional data before being input into the NNs in audio-based methods of understanding terrain [39], [40]. Many audio transformation methods have been proposed by [41]. Mel-frequency cepstral coefficients (MFCCs) have been used predominantly as one of the most effective parameterizations of acoustic features. MFCCs are a cepstrum-based feature representation method that mimics human hearing features: a difference on the Mel scale can be felt like the same difference in pitch.

The most distinctive point of the Mel spectrogram is that it is specialized for acoustic representation, so it omits pitch

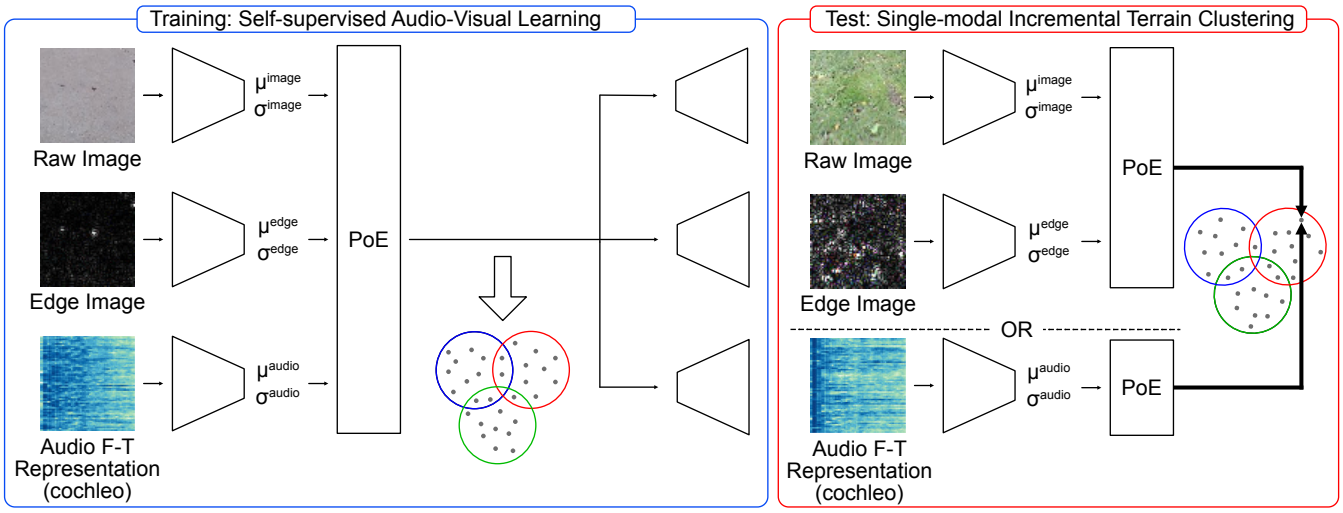

Fig. 2: Overview of our proposed terrain clustering framework. The training scheme consists of two modules, feature learning using MVAE from audio-visual data and feature clustering using GMM. The testing is consists of feature extraction using the encoder of trained MVAE and cluster prediction using IGMM. Our method predicts the terrain cluster from a single modality data (either image or audio) even though the training is conducted in a multi-modal manner to extract the informative feature.

information. However, using the sound of rolling wheels as an example, we assumed that in the terrain clustering task, pitch information can be an important cue because it depends on the material of the floor on which the robot runs. We, therefore, investigated another processing method, the cochleogram.

The cochleogram is a gammatone-based filtering method. Cochleogram images are obtained from a detailed biophysical cochlear model, with the assumption that modeling the principles of the human auditory system in greater detail could increase the system's performance [42]. Its gammatone filter banks are a parallel sequence of bandpass filters defined as follows [41]:

$$h_i(t) = at^{n-1}e^{-2\pi B_i t}\cos(2\pi\omega_i t + \phi)\ (t \geq 0), \qquad (1)$$

where a is a constant factor; t is the timestep; n denotes the order of the filter; B_i is the bandwidth; ω_i is the center frequency in rad/sec; and ϕ indicates the phase. In this paper, we used equivalent rectangular bandwidth (ERB) for B_i.

B. Variational autoencoder

VAE is a powerful method that extracts low-dimensional features in the bottle-neck layer between the decoder, p parameterized with θ and encoder, and q parameterized with ϕ in the spherical Gaussian form. A key point of this algorithm is that to combat the intractability of the marginal likelihood of the data that we want to maximize in training-time, VAE optimizes the evidence lower bound (ELBO). The definition of the ELBO can be defined as [43]:

$$\begin{aligned} ELBO(x) \equiv\ & \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] \\ & - \beta D_{KL}(q_\phi(z \mid x) \parallel p(z)), \end{aligned} \qquad (2)$$

where $D_{KL}(q \parallel p)$ is the Kullback-Leibler (KL) divergence between two distributions, p and q and β is an annealing factor [43] to alleviate the effect of D_{KL}'s sharp increase in the early epoch, which disturbs the decrease of losses. Another key point of VAE is that it enables backward propagation via the reparameterization trick. Given the autoencoder's outputs $\mu(x)$ and $\sigma(x)$, where x is the input, we sampling the latent vector z from $\mathcal{N}(\mu(x), \sigma(x))$ by sampling $\varepsilon \sim \mathcal{N}(0, I)$ instead of directly sample z from $q(z|x)$.

C. Multi-modal feature learning

We encode three modalities—raw RGB image x^{raw}, edge image x^{edge}, and two-dimensional audio data x^{audio}, which are generated by combining MFCCs and a cochleogram— into a single latent vector. To handle the modality missing at test-time (either image or audio data), we employ the paradigm of MVAE [26].

Wu and Goodman proposed [26] MVAE as an extension of the VAE, such that its input is not one x_1 but extended to N-modalities $x_1, x_2, ..., x_N$ [26]. Because this model is established by assuming that each modality is conditionally independent, the reconstructed modality can be decoded from one shared latent vector z. Under this assumption, the product-of-experts (PoE) structure [44], which originally aimed to reduce the number of inference networks, is applied to the MVAE by combining variational parameters in individual expert models. Theoretically, the PoE maximizes the log-likelihood of data. Also, because of MVAE's sub-sampled training paradigm, we can force every subset of modality to have close latent variables in the inference network while avoiding numerical problems.

Under this condition, we can consider set X, which includes arbitrary subsets of present modalities. The multi-modal ver-

sion of ELBO is [26]:

$$ELBO(X) \equiv \mathbb{E}_{q_\phi(z|x)}[\sum_{x_i \in X} \lambda_i \log p_\theta(x_i \mid z)] - \beta D_{KL}(q_\phi(z \mid x_i) \parallel p(z)), \quad (3)$$

where λ_i is the balancing factor that regulates the effects on loss of each modality. Then, the target loss function of MVAE can be written as:

$$\mathcal{L} \equiv ELBO(x_1, x_2, ..., x_N) + \sum_{i=1}^{N} ELBO(x_i) + \sum_{j=1}^{k} ELBO(X_j) + \beta D_{KL}, \quad (4)$$

where N is the number of modalities, and M denotes the number of randomly chosen subsets, X_j.

Though our method uses images and edges as independent input for MVAE and reconstructs them respectively, input x^{image} and x^{edge} are fundamentally the same, so we always regard image and edge as a single unit. A byproduct of this unity is the reduction of computational costs. In summary, the target loss in our training time can be rewritten as:

$$\mathcal{L} \equiv ELBO(x^{image}, x^{edge}, x^{audio}) + ELBO(x^{image}, x^{edge}) + ELBO(x^{audio}) + \beta D_{KL}. \quad (5)$$

To calculate the first term of Eq.3, which indicates the reconstruction error, we used mean squared error (MSE) defined as the following:

$$MSE = \frac{1}{N} \sum_{i}^{N} \sum_{k}^{D} (y_i^{(k)} - x_i^{(k)})^2, \quad (6)$$

where y is the reconstructed data, x is the input data, N is the size of the mini-batch used in training, and D is the dimension of the feature.

D. Clustering of the latent vectors

If multiple items of data have the same characteristics, the latent variables tend to be stochastically located close to each other within a Gaussian distribution in latent space. On the other hand, if data have different characteristics, they tend to be located away from each other. What is significant is that even if the classification of the material of the terrain is the same (*e.g.*, Grass), latent variables can separate from each other. Conversely, if different materials (Carpet and Grass) have similar colors and make similar sounds when robots are running over them, they are clustered closely. Sometimes, materials are mixed up or clusters overlap in latent space. This is the very reason that soft clustering is more suitable than hard clustering for our aims, which is why we determined the stochastic assignment of the GMM to be suitable. We also demonstrated the clustering using the k-means algorithm: the results of GMM clustering outperformed the results of k-means clustering in terms of accuracy and stability.

Another essential and powerful advantage of using GMM is that we do not have to determine the number of clusters, unlike with k-means. This characteristic is very important for unsupervised learning, in which there are no labels to help count the true numbers of clusters, and for terrain clustering, in which the number of terrain types endlessly increases. Here we use the IGMM algorithm proposed by Engel and Heinen [45], which enables us to deal flexibly with a new point that arrives at test-time and that does not seem to belong to any existing cluster. The algorithm uses a minimum likelihood criterion to assign the point to one cluster. If a new latent variable arrives, the data point x is to belong to a new cluster or an existing cluster j, where the likelihood $p(x|j)$ is the minimum likelihood and is lower than the threshold of each cluster. Then the parameters (*i.e.*, the means μ^j, the covariances σ^j, and the mixing parameters $p(j)$) accumulate as the summation of the posterior probability $p(j|x)$.

IV. Experiment

A. Dataset

We used a dataset introduced by Kurobe *et al*. [21]. This dataset uses a super-directional microphone to capture mono-stereo audio data and uses an RGB camera to capture image data with about 24 fps. Further information about the capturing sensor setup is explained in the previous paper [21].

However, in this paper, we modified the original dataset. The Kurobe *et al*. [21] dataset aims to predict terrain class based only on the RGB image at the test time, audio data are not included in the original test set. Moreover, Kurobe *et al*. [21] experimented only with scene-specific evaluation. That is to say, the trained model can be used only for the single scene, and Kurobe *et al*.do not evaluate scene-generic performance. In this paper, we aim to generalize for multiple different scenes.

The dataset includes 21 independent movies with RGB frames and audio data that include seven classes: *Pavement*, *Grass*, *Rough concrete*, *Concrete flooring*, *Carpet*, *Tile*, and *Linoleum*. To generate an the RGB frame and its corresponding audio clip, we used 2.8 seconds of the audio clip for each image frame.

There are some temporal gaps between a temporal audio clip and the corresponding temporal image frame recorded by a camera due to the dataset's setup [21] because the super-directional microphone picks up the sound of wheels and the RGB camera captures the terrain ahead. To avoid ambiguity in the borders of terrain due to this gap, we excluded 35 segments in which the camera was recording a border. Finally, we separated the dataset into $41,315$ data pairs for the training set and $7,734$ for the test set. Note that we annotated the terrain types for the data pair, but this ground-truth label was used only for evaluation, and we did not use the labels in the training or clustering processes.

B. Network Architecture

The architecture of the proposed model is similar to the image network employed in MVAE [26]. The encoder of the image and the audio is composed of two-dimensional convolutional layers with a kernel size of 4×4, a batch-normalization layer, and Swish activation function [46]. The decoder of the

image and audio is composed of transposed two-dimensional convolutional layers with a kernel size of 4×4, a batch-normalization layer, and Swish activation function. We replace fully connected layers with convolutional layers with 5×5 of kernel size to reduce the number of parameters in the network.

C. Data processing

1) Image: We cropped the images in the dataset to 720×720 from the lower center of the images, then we re-scale them to 68×68. These images were also randomly cropped to 64×64 size for data augmentation.

2) Audio: Every audio frame was processed as we explained in the experiment section. The number of feature dimensions of the cochleogram was set to 48 and those of MFCCs was set to 16 to obtain a final input shape of 64×64 with one channel. To process the MFCCs, we limited the frequency levels for the calculations by cutting off frequencies higher than $1500Hz$. We used MFCCs with 64 triangular filters on the Mel scale, and 2^{16}-point fast Fourier transform [21]. For the cochleogram, we used the following values: 9.265 as the asymptotic filter quality factor and 24.7 as the minimum bandwidth.

D. Training details

We used the Adam optimizer [47] with a learning rate of 10^{-4} and a minibatch size of 300. The annealing factor β was set to anneal from 0.0 to 1.0 in the first 20 epochs.

E. Evaluation metrics

In this paper, we employ two evaluation metrics: normalized mutual information (NMI) and clustering accuracy (ACC), followed by the conventional clustering method [6].

For NMI, the formula is written as:

$$NMI(Y, C) = \frac{2 \times I(Y; C)}{H(Y) + H(C)}, \tag{7}$$

where Y is the ground-truth class labels; C is the cluster labels predicted by GMM or IGMM; $H(x)$ is entropy across x; and $I(Y; C)$ is the mutual information, given as:

$$I(Y; C) = H(Y) - H(Y \mid C), \tag{8}$$

where $H(Y \mid C)$ is the conditional entropy of C. A higher value is better for both the NMI and accuracy metrics.

The accuracy of the clustering is defined as:

$$ACC(Y, C) = \max_{m} \frac{\sum_{i=1}^{N} 1\{l_i = m(c_i)\}}{N}, \tag{9}$$

where N is the total number of clusters created; m_i is the assigned label in the clustering process; and l_i is the ground-truth label for each frame.

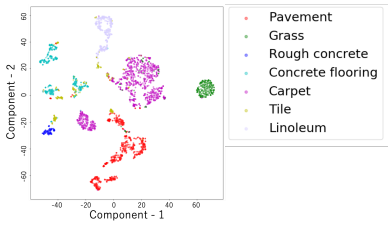

Fig. 3: Visualization of the latent vectors using t-SNE. The ground-truth cluster labels are colorized.

F. Comparison with other method

No conventional method has been proposed for terrain type clustering based on audio-visual feature learning and single-modal terrain type prediction. Therefore, we compared this study's clustering results with those of the conventional method [21] using our dataset. We do not compare our results with those of the method proposed by Zurn *et al*. [6] because they aimed to segment images into terrain types, which differs from our aim.

We compared the clustering results using two methods as follows:

- Kurobe *et al*. [21] without CNN:
  Kurobe *et al*. trained VAEs for audio and image input and then created pseudo labels with the trained data using clustering. To conduct a fair comparison with our method, we train the image and audio VAEs, and the features are extracted, then the features extracted using GMM. We calculated the values of NMI and ACC based on the clustering labels.
- Kurobe *et al*. [21] with CNN:
  Kurobe *et al*. predicted cluster labels using CNN trained in a self-supervised manner using the pseudo labels from the audio and image data. Regarding the output labels of the CNN that resulted from clustering, we calculated the values of NMI and ACC using the ground-truth labels and the output labels.

V. Results

A. Visualization of the latent space

We visualized the latent space generated by the PoE structure using the t-SNE algorithm [48] against the training data. The result of t-SNE on the latent variables is shown in Fig. 3. In the figure, the ground-truth terrain type is colorized: the embeddings labeled by the same color share similar traits. We observe that the clusters are well separable and highly correlate with the ground truth classes.

B. Qualitative evaluation

Fig. 4 shows an example of the qualitative results of our method. The category (the name of the terrain type) of each cluster is unknown, but the categories are shown as the reference. In each box, the most right data are the training data samples that belong to the same cluster. Our method takes an input of either image or audio at the test time and predicts

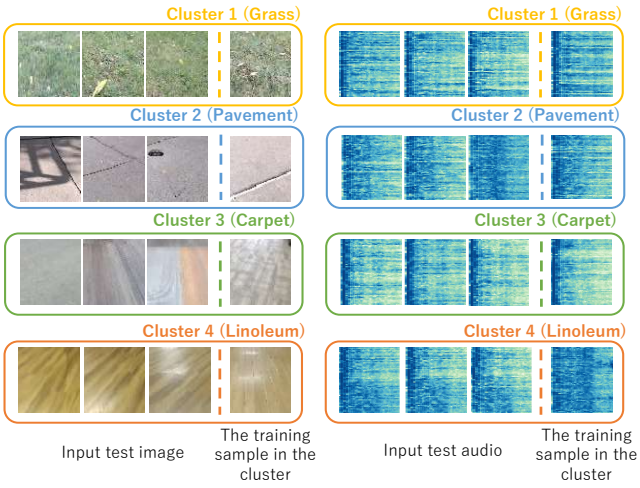

Fig. 4: The qualitative results of our method. The figure left shows the results with images, and the right shows the results with audio data. The data with the same predicted clusters are covered with the same colorized boxes.

TABLE I: Comparison with Kurobe *et al*. [21] (*w/* is the abbreviation of *with*, and *w/o* is the abbreviation of *without*).

Method	Input	NMI ↑	ACC (%) ↑
[21] w/o CNN	Audio+image	0.589	58.12
[21] w/ CNN	Image	0.001	23.18
Ours w/o update	Image	**0.401**	48.90
Ours w/ update	Image	0.377	**50.63**
Ours w/o update	Audio	0.353	50.30
Ours w/ update	Audio	**0.500**	**74.39**

the terrain cluster index. From the figure, we can observe that our method successfully predict the correct terrain cluster index from a single modality data (either image or audio) even though the feature learning is conducted in a multi-modal manner.

C. Quantitative evaluation

We summarize the quantitative evaluation of our method in Table. I. Since Kurobe *et al*. [21] did not assume a large scale of data with many types of terrain, the results show a fatal decline of the NMI and the accuracy when using only image data, though its results when using an input of image and audio outperforms our method. Also, Table. I shows the evaluation of the clustering with and without an incremental update. By comparing the accuracy of non-updated clustering and updated clustering, we can observe that the incremental update at the test time improves the clustering accuracy.

D. Ablation study of the visual input

We evaluated the importance of using edge images as input. Comparing the methods using edge images and the methods not using the edge shows that the method which takes the input of both RGB and the edge outperform the method with only RGB by 10% in terms of the accuracy. This shows the effectiveness of using the edge information for the terrain type clustering.

TABLE II: The comparison of the method (using RGB and edge image, and using only RGB image as an input of visual information). We note that for training, we use both image data and audio data at the training stage.

Input	w/o update		w/ update	
	NMI ↑	ACC(%) ↑	NMI ↑	ACC(%) ↑
RGB	0.272	40.16	0.389	57.53
Ours (RGB+Edge)	**0.353**	**50.30**	**0.500**	**74.39**

TABLE III: Comparison of the results with different audio processing methods: MFCCs, MFCCs + cochleogram, and ours (cochleogram) to show the effectiveness of using cochleogram for terrain type clustering.

Method	Input	w/o update		w/ update	
		NMI	ACC(%)	NMI	ACC(%)
MFCCs	Audio	**0.559**	**55.92**	0.235	34.73
MFCCs + cochleogram	Audio	0.389	49.57	**0.443**	47.98
Ours (cochleogram)	Audio	0.401	48.90	0.377	**50.63**
MFCCs	Image	0.295	47.26	0.389	61.02
MFCCs + cochleogram	Image	0.318	45.72	0.423	67.71
Ours (cochleogram)	Image	**0.353**	**50.30**	**0.500**	**74.39**

E. Ablation study on sound input

We also performed an experiment to confirm the effectiveness of our method, comparing the encoding using only MFCCs and using the combination of cochleogram and MFCCs [40]. To make combined audio inputs, we concatenated the MFCCs' features $x^{mfcc} \in \mathbb{R}^{16\times64}$ and cochleogram features $x^{coch} \in \mathbb{R}^{48\times64}$. The final input is $x^{audio} = [x^{mfcc}; x^{coch}] \in \mathbb{R}^{64\times64}$.

The results of comparing methods of audio processing are shown in Table. III. Though the MFCCs method outperformed the cochleogram method when the GMM was not updated during prediction, its NMI and ACC drop significantly below those of the cochleogram method after updating the GMM and reassigning the test data. Besides, the cochleogram method outperformed the combined method for ACC. For the case of the incremental clustering, it can be said that the cochleogram method is suitable for preprocessing among the three preprocessing methods.

VI. Conclusion

We presented a novel framework to cluster terrain types from the latent variables of the MVAE using IGMM. Because of subset training and the product-of-experts architecture of the MVAE, our method works even though either visual data or audio data is not available at the test time. The result demonstrated that even if one of the modality is missing due to the extreme conditions, our model can receive partial input and predict the correct terrain type. This framework also enables us to conduct incremental clustering to achieve high-performance clustering. We also demonstrated that our input preprocessing method, which uses edge and RGB information for visual data and cochleogram for sound data, can be used to extract the informative feature for terrain type clustering.

ACKNOWLEDGMENT

This research is supported by JST (JPMJMI19B2).

REFERENCES

[1] A. Angelova, L. Matthies, D. Helmick, and P. Perona, "Fast terrain classification using variable-length representation for autonomous navigation," in *Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[2] R. Manduchi, A. Castano, A. Talukder, and L. Matthies, "Obstacle detection and terrain classification for autonomous off-road navigation," *Autonomous Robots*, vol. 18, pp. 81–102, 01 2005.

[3] D. F. Wolf, G. S. Sukhatme, D. Fox, and W. Burgard, "Autonomous terrain mapping and classification using hidden markov models," in *International Conference on Robotics and Automation*, 2005, pp. 2026–2031.

[4] M. Hebert and N. Vandapel, "terrain classification techniques from ladar data for autonomous navigation," in *Collaborative Technology Alliances conference*, May 2003.

[5] K. Otsu, M. Ono, T. J. Fuchs, I. Baldwin, and T. Kubota, "Autonomous terrain classification with co- and self-training approach," *Robotics and Automation Letters*, vol. 1, no. 2, pp. 814–819, 2016.

[6] J. Zürn, W. Burgard, and A. Valada, "Self-supervised visual terrain classification from unsupervised acoustic feature learning," *CoRR*, vol. abs/1912.03227, 2019.

[7] M. Z. Hashmi, Q. Riaz, M. Hussain, and M. Shahzad, "What lies beneath one's feet? terrain classification using inertial data of human walk," *Applied Sciences*, vol. 9, p. 3099, Jul. 2019.

[8] K. Yang, L. M. Bergasa, E. Romera, R. Cheng, T. Chen, and K. Wang, "Unifying terrain awareness through real-time semantic segmentation," in *Intelligent Vehicles Symposium*, 2018, pp. 1033–1038.

[9] J. Christie and N. Kottege, "Acoustics based terrain classification for legged robots," in *International Conference on Robotics and Automation*, 2016, pp. 3596–3603.

[10] J. Libby and A. J. Stentz, "Using sound to classify vehicle-terrain interactions in outdoor environments," in *International Conference on Robotics and Automation*, 2012, pp. 3559–3566.

[11] L. Ojeda, J. Borenstein, G. Witus, and R. Karlsen, "Terrain characterization and classification with a mobile robot," *Journal of Field Robotics*, vol. 23, Feb. 2006.

[12] A. Valada, L. Spinello, and W. Burgard, *Deep Feature Learning for Acoustics-Based Terrain Classification*. Cham: Springer International Publishing, 2018, pp. 21–37.

[13] S. S. Baishya and B. Bäuml, "Robust material classification with a tactile skin using deep learning," in *International Conference on Intelligent Robots and Systems*, 2016, pp. 8–15.

[14] K. Takahashi and J. Tan, "Deep visuo-tactile learning: Estimation of material properties from images," *CoRR*, vol. abs/1803.03435, 2018.

[15] B. Park, J. Kim, and J. Lee, "Terrain feature extraction and classification for mobile robots utilizing contact sensors on rough terrain," *Procedia Engineering*, vol. 41, pp. 846 – 853, 2012.

[16] C. Bai, J. Guo, and H. Zheng, "Three-dimensional vibration-based terrain classification for mobile robots," *IEEE Access*, vol. 7, pp. 63 485–63 492, 2019.

[17] C. C. Ward and K. Iagnemma, "Speed-independent vibration-based terrain classification for passenger vehicles," *Vehicle System Dynamics*, vol. 47, no. 9, pp. 1095–1113, 2009.

[18] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation," in *Conference on Robot Learning*, 2019.

[19] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *International Conference on Computer Vision*, Oct. 2019.

[20] H. Blum, A. Gawel, R. Siegwart, and C. Cadena, "Modular sensor fusion for semantic segmentation," in *International Conference on Intelligent Robots and Systems*, Oct. 2018, pp. 3670–3677.

[21] A. Kurobe, Y. Nakajima, H. Saito, and K. Kitani, "Audio-visual self-supervised terrain type discovery for mobile platforms," *CoRR*, vol. abs/2010.06318, 2020.

[22] D. Hu, L. Mou, Q. Wang, J. Gao, Y. Hua, D. Dou, and X. X. Zhu, "Ambient sound helps: Audiovisual crowd counting in extreme conditions," *CoRR*, vol. abs/2005.07097, 2020.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[24] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *CoRR*, vol. abs/1611.05431, 2016.

[25] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," *CoRR*, vol. abs/1812.05069, 2018.

[26] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *International Conference on Neural Information Processing Systems*, 2018, pp. 5580—-5590.

[27] M. Ono, T. J. Fuchs, A. Steffy, M. Maimone, and J. Yen, "Risk-aware planetary rover operation: Autonomous terrain classification and path planning," in *Aerospace Conference*, 2015, pp. 1–10.

[28] A. Chilian and H. Hirschmüller, "Stereo camera based navigation of mobile robots on rough terrain," in *International Conference on Intelligent Robots and Systems*, 2009, pp. 4571–4576.

[29] P. Moghadam and W. Wijesoma, "Online, self-supervised vision-based terrain classification in unstructured environments," in *International Conference on Systems, Man and Cybernetics*, 2009, pp. 3100–3105.

[30] Xiuwen Liu and DeLiang Wang, "Texture classification using spectral histograms," *IEEE Transactions on Image Processing*, vol. 12, no. 6, pp. 661–670, 2003.

[31] P. Filitchkin and K. Byl, "Feature-based terrain classification for littledog," in *International Conference on Intelligent Robots and Systems*, 2012, pp. 1387–1392.

[32] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: adaptive multi-modal gesture recognition," *CoRR*, vol. abs/1501.00102, 2015. [Online]. Available: http://arxiv.org/abs/1501.00102

[33] 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 Robotics and Automation Letters*, vol. 4, no. 2, pp. 1509–1516, 2019.

[34] C. Brooks and K. Iagnemma, "Self-supervised classification for planetary rover terrain sensing," in *Aerospace Conference*, 2007, pp. 1–9.

[35] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: Feature learning by inpainting," in *Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[36] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conference on Computer Vision*, 2016, pp. 649–666.

[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

[38] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images," *CoRR*, vol. abs/1804.04488, 2018.

[39] D. Hu, L. Mou, Q. Wang, J. Gao, Y. Hua, D. Dou, and X. xiang Zhu, "Ambient sound helps: Audiovisual crowd counting in extreme conditions," *CoRR*, vol. abs/2005.07097, 2020.

[40] A. Tjandra, S. Sakti, G. Neubig, T. Toda, M. Adriani, and S. Nakamura, "Combination of two-dimensional cochleogram and spectrogram features for deep learning-based asr," in *International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4525–4529.

[41] H. Chaurasiya, "Time-frequency representations: Spectrogram, cochleogram and correlogram," *Procedia Computer Science*, vol. 167, pp. 1901–1910, Jan. 2020.

[42] M. Russo, L. Kraljević, M. Stella, and M. Sikora, "Cochleogram-based approach for detecting perceived emotions in music," *Information Processing & Management*, vol. 57, no. 5, p. 102270, 2020.

[43] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2016.

[44] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[45] P. Engel and M. Heinen, "Incremental learning of multivariate gaussian mixture models," in *Advances in Artificial Intelligence*, 2010, pp. 82–91.

[46] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *CoRR*, vol. abs/1710.05941, 2017.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[48] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, pp. 2579–2605, 2008.