2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

SpCoMapGAN: Spatial Concept Formation-based Semantic Mapping with Generative Adversarial Networks

Yuki Katsumata[1], Akira Taniguchi[1], Lotfi El Hafi[1], Yoshinobu Hagiwara[1], and Tadahiro Taniguchi[1]

***Abstract*— In semantic mapping, which connects semantic information to an environment map, it is a challenging task for robots to deal with both local and global information of environments. In addition, it is important to estimate semantic information of unobserved areas from already acquired partial observations in a newly visited environment. On the other hand, previous studies on spatial concept formation enabled a robot to relate multiple words to places from bottom-up observations even when the vocabulary was not provided beforehand. However, the robot could not transfer global information related to the room arrangement between semantic maps from other environments. In this paper, we propose SpCoMapGAN, which generates the semantic map in a newly visited environment by training an inference model using previously estimated semantic maps. SpCoMapGAN uses generative adversarial networks (GANs) to transfer semantic information based on room arrangements to a newly visited environment. Our proposed method assigns semantics to the map of an unknown environment using the prior distribution of the map trained in known environments and the multimodal observations made in the unknown environment. We experimentally show in simulation that SpCoMapGAN can use global information for estimating the semantic map and is superior to previous methods. Finally, we also demonstrate in a real environment that SpCoMapGAN can accurately 1) deal with local information, and 2) acquire the semantic information of real places.**

I. Introduction

In the field of autonomous mobile robots, such as cleaning robots that operate in human living environments by estimating the meaning of vocabulary related to places included in human utterances, semantic mapping connects semantic information to an environment map [1]. For example, for a cleaning robot to execute the command "Clean kitchen and John's room" given by a user, the robot needs to understand both where "kitchen" is, which is global information existing in many common environments, and where "John's room" is, which is local information existing only in specific environments. Therefore, accurate understanding of word meanings is important for robots to perform tasks triggered by communication with humans.

This study was partially supported by the Japan Science and Technology Agency (JST) Core Research for Evolutionary Science and Technology (CREST), grant number JPMJCR15E3, and by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research (B), grant number 18H03308, and Grant-in-Aid for Scientific Research on Innovative Areas, grant number 16H06569.

[1]Yuki Katsumata, Lotfi El Hafi, Akira Taniguchi, Yoshinobu Hagiwara, and Tadahiro Taniguchi are with Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan. {yuki.katsumata, lotfi.elhafi, a.taniguchi, yhagiwara, taniguchi}@em.ci.ritsumei.ac.jp

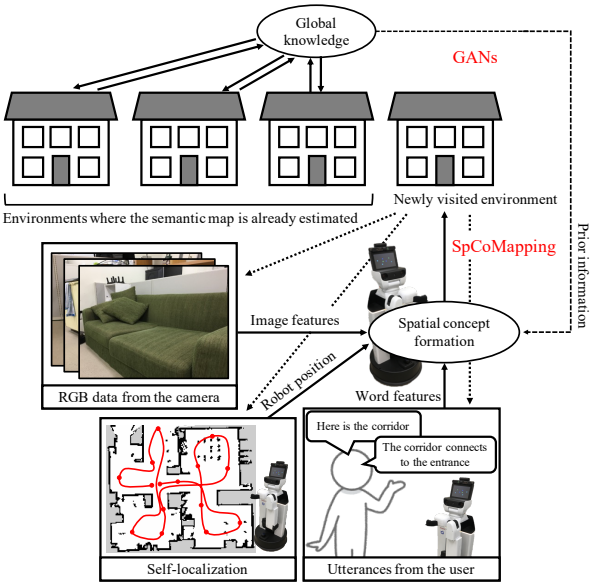


Fig. 1. Overview the proposed semantic mapping method: SpCoMapGAN.

When performing semantic mapping, it is therefore better for robots to deal with both of local and global information. SpCoMapping, proposed in [2], defined clusters of multimodal information that a robot acquires as spatial concepts, and was a spatial concept formation method that integrates a Markov random field (MRF) for estimating the semantic map and vocabulary representing places simultaneously. However, SpCoMapping could not use global information already acquired in other environments. In previous studies which use global information, methods were proposed for performing semantic mapping in newly visited environments using neural networks trained with a large amount of data related to places [3], [4], [5]. However, these above methods could not deal with local information and faced the problem that the vocabulary representing the places is determined by the labels included in the training dataset. In addition, previous studies could not deal with global information related to the room arrangement in semantic maps. Therefore, the robots could not estimate labels from the relationships among regions.

In this study, our proposed method models the joint distribution of the semantic map using previously estimated semantic maps by the robot, and generates the semantic map using that joint distribution in a newly visited environment. The joint distribution of the semantic map is the joint distribution of the class in each cell of the semantic map. However, this joint distribution is difficult to model because of the high number of dimensions. Therefore, we approximate it by an

inference model.

In this regard, generative adversarial networks (GANs) [6], which is a type of generative model, can approximate a real data distribution using adversarial learning. A related study on GANs, pix2pix [7], proposed a method that learns the relationship between a pair of images using U-net [8] in the generator part to enable line drawing coloring and image completion. The semantic mapping task can be similarly solved in the pix2pix framework by coloring the occupancy grid map as a line drawing.

Hence, we propose a novel spatial concept formation-based semantic mapping with GANs (SpCoMapGAN) which estimates the semantic map by training an inference model in the pix2pix framework using previously estimated semantic maps and integrating it with SpCoMapping. Fig. 1 shows an overview of our approach for semantic mapping combining bottom-up spatial concepts with global knowledge.

The main contributions of this work are as follows:

1) We propose SpCoMapGAN that uses already estimated semantic maps to model the joint distribution of the semantic map, and applies that model to improve the accuracy of semantic mapping in newly visited environments.
2) We show that the joint distribution of each cell class, which is difficult to model, can be approximated by an inference model trained in the pix2pix framework.

II. Related work

A. Spatial concept formation and semantic mapping

SpCoMapping is an extended method of spatial concept formation [9], [10], [11], [12], using a Markov random field (MRF) for semantic mapping [2]. SpCoMapping learns the vocabulary representing a place and a region simultaneously, taking into account the shapes of the environment and obstacles. Therefore, the robot associates multiple words to places using local information, even if the vocabulary is not provided beforehand. In this case, there is no need to provide the vocabulary to the robot in advance, and the robot can deal with words used by humans in the local environment. However, SpCoMapping cannot use global information already acquired in other environments. It means that SpCoMapping requires observation to estimate the vocabulary representing the places, and the estimation accuracy of the labels of regions without observation is lower than that of the regions with observation.

In previous studies which uses global information, methods were proposed for performing semantic mapping in newly visited environments using a network trained with a large amount of data related to places [3], [4], [13], [14], [15], [5]. These methods enable online semantic mapping in newly visited environments by the mean of transfer learning. Pal et al. proposed DEDUCE [5] which obtains the semantic labels of regions from image features using CNN and YOLO [16]. Because DEDUCE uses YOLO and a CNN trained with Places365 [17], which is a large dataset of place images, robots can obtain the semantic labels without environment-specific training even in unknown environments. However, these methods cannot deal with local information like "John's room" and face the problem that the vocabulary representing the place is determined by the labels of the dataset used for the training.

Moreover, the global knowledge included in the semantic maps previously estimated by the robot, such as the relationships among regions, cannot be transferred to a new environment with these approaches. Therefore, our proposed SpCoMapGAN method models the joint distribution of the semantic map using already estimated maps by the robot, and samples the semantic map using that joint distribution in a newly visited environment.

B. Coloring line drawing with GANs

When performing semantic mapping, we want the robot to maintain some features from the observation to the semantic map, such as the occupancy grid, unobserved areas, and semantic labels estimated from the observations in the occupancy grid map. Therefore, semantic mapping can be solved in the pix2pix [7] framework, which is one approach of GANs [6], by considering the similar task of coloring the occupancy grid map as a line drawing. Indeed pix2pix, which learns the relationship between a pair of images using U-net [8] in the generator part, enables line drawing coloring and image completion. The pix2pix generator has skip connections to extract the features of the original image at the encoder by using the image as input and adding these features at the decoder to the output of the generator.

More specifically, Mirza et al. proposed conditional GAN (CGAN), which is a GANs method using conditional information, to learn the relationships between the training and condition data [18]. pix2pix is a method based on the CGAN framework. Moreover, GANs face the problem of difficult convergence, because two networks are trained simultaneously [19]. In this regard, Miyata et al. proposed spectral normalization GAN (SNGAN) with improved performance using spectral normalization as the discriminator weight [20].

III. SpCoMapGAN: Spatial concept formation-based semantic mapping with GANs

We propose a novel spatial concept formation-based semantic mapping with GANs (SpCoMapGAN) which estimates the semantic map by training an inference model in the pix2pix framework using previously estimated semantic maps, and integrating it with SpCoMapping.

First, Section III-A introduces the overview of SpCoMapGAN. Next, Section III-B introduces an alternative simple model for the spatial concept formation. Section III-C describes the training of the semantic map inference model in the pix2pix framework. Finally, Section III-D describes the semantic map estimation by SpCoMapGAN.

A. Overview

Fig. 2 presents the overview of the approximation of the joint distribution of semantic map. SpCoMapGAN performs semantic mapping according to the following procedure:

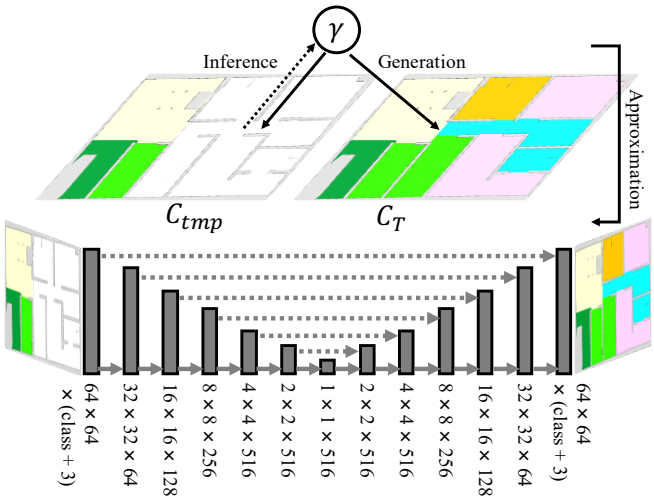


Fig. 2. Overview of the approximation of the joint distribution of each cell class in the semantic map. The joint distribution of each cell class of the semantic map is high-dimensional and difficult to model. Therefore, this joint distribution is approximated by an inference model trained in the pix2pix framework.

1) We prepare a large number of sets of the semantic maps C_T and the spatial concepts C_{tmp} previously estimated by robots.
2) An inference model of the semantic map is trained in the pix2pix framework using C_T as the training data and C_{tmp} as the condition.
3) The robot estimates the spatial concepts C_{tmp} using the observations made in a newly visited environment.
4) The semantic map of the newly visited environment is generated by SpCoMapGAN using the spatial concepts C_{tmp} and the inference model trained with the previously estimated semantic maps.

In this method, two types of probabilistic generative model of spatial concept formation are used according to the situation, namely SpCoMapGAN and an alternative simple model of the spatial concept formation. The alternative simple model of the spatial concept formation generates C_{tmp} which is used as the condition when training the inference model in the pix2pix framework. SpCoMapGAN trains the inference model in the pix2pix framework using already estimated semantic maps, and integrates it with the alternative simple model of the spatial concept formation.

The graphical model of SpCoMapGAN is presented in Fig. 3, and its variables are defined in Table I. Eq. (1)-(5) describe the generative model of the proposed method where $\mathrm{Dir}()$ represents the Dirichlet distribution and $\mathrm{Mult}()$ the multinomial distribution.

$$C \sim p(C \mid m, \gamma) \tag{1}$$
$$\theta_l \sim \mathrm{Dir}(\chi) \tag{2}$$
$$w_l \sim \mathrm{Dir}(\beta) \tag{3}$$
$$f_t \sim \mathrm{Mult}(x_t, \theta_{C_{i,j}}) \tag{4}$$
$$s_t \sim \mathrm{Mult}(x_t, w_{C_{i,j}}) \tag{5}$$

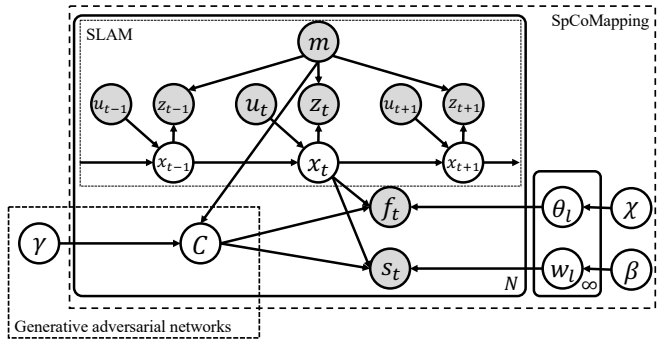


Fig. 3. Graphical model of SpCoMapGAN where gray nodes indicate observation variables, and white nodes unobserved variables.

TABLE I

DEFINITIONS OF THE VARIABLES OF THE GRAPHICAL MODEL.

Symbol	Definition
m	Environment map
x_t	Robot self-position
u_t	Control data
z_t	Distance data
C	Semantic map of the environment
γ	Parameter of the prior distribution of the semantic map
f_t	Image features
s_t	Word features (bag-of-words)
θ_l	Parameter of the multinomial distribution for f_t
w_l	Parameter of the multinomial distribution for s_t
α, β, χ	Hyperparameters of prior distributions

B. Alternative simple model for spatial concept formation

The alternative simple model of the spatial concept formation used in SpCoMapGAN is the method for forming the spatial concept C_{tmp} used as condition when training the inference model with the pix2pix framework in the newly visited environment. Any existing methods for spatial concept formation can be used for the alternative simple model. In this study, we use a simple spatial concept formation model excluding the dependencies between the cells of the semantic map.

The model parameters $c_{n,i}$, π, θ_l, and w_l are estimated by Gibbs sampling, where $c_{n,i}$ represents the semantic labels of the i-th cell on the semantic map of the n-th environment, and π the parameter of the multinomial distribution for $c_{n,i}$. The parameters $c_{n,i}$, π, θ_l, and w_l are sampled by the following equations:

$$t' = \{t \mid C_{\mathrm{convert}(x_{n,t})} = l, t \in (1:T)\}, \tag{6}$$
$$c_{n,i} \sim \prod_{t'} \mathrm{Mult}(f_{n,t'} \mid x_{n,t'}, \theta_{l=c_{n,i}}) \mathrm{Mult}(s_{n,t'} \mid x_{n,t'}, w_{l=c_{n,i}}) \mathrm{Mult}(c_{n,i} \mid \pi), \tag{7}$$
$$\pi \sim \prod_{i \in S_{\mathrm{free}}} \mathrm{Mult}(c_{n,i} \mid \pi) \mathrm{Dir}(\pi \mid \alpha), \tag{8}$$
$$\theta_l \sim \prod_{t'} \mathrm{Mult}(f_{n,t'} \mid \theta_{l=c_{n,i}}, x_{n,t'}) \mathrm{Dir}(\theta_l \mid \chi), \tag{9}$$
$$w_l \sim \prod_{t'} \mathrm{Mult}(s_{n,t'} \mid w_{l=c_{n,i}}, x_{t'}) \mathrm{Dir}(w_l \mid \beta), \tag{10}$$

where S_{free} represents the number of cells g_i of $p(g_i) > 0.5$.

In this case, the occupancy grid map is 2D, i.e., x_t is a variable with xy coordinates represented by $x_t = (x_{t,\mathrm{x}}, x_{t,\mathrm{y}})$.

The semantic map has semantic labels on the unoccupied area in the occupancy grid map. Both coordinate systems are transformed by the following equation:

$$\mathrm{convert}(x_t) = Ux_{t,\mathrm{x}} + x_{t,\mathrm{y}} \tag{11}$$

where U represents the width of the occupancy grid map.

C. Training the inference model

To train the inference model of the semantic map in the pix2pix framework, a set of the semantic maps C_T and the spatial concepts C_{tmp} is used. C_T and C_{tmp} are evaluated by the robot in environments where semantic maps are already estimated.

When the cell of the occupancy grid map with the index i is defined as g_i, the occupancy grid map m is represented by $m = \{g_i\}$, where $(i \in S)$ and S represents the number of cells in the occupancy grid map. Each g_i is assigned a binary occupancy value: $g_i = 1$ if the cell is occupied and $g_i = 0$ if not. The semantic map is the occupancy grid map with semantic labels in the unoccupied area, and can be expressed by $C = \{c_{n,i}\}$, where n is the index of the training environment, and i the index of S_{free}. Each $c_{n,i}$ has a semantic label.

When training the networks with pix2pix, the input is a semantic map in a 2D tensor of size $\mathrm{height} \times \mathrm{width}$ of the map. The image data of the semantic map used for the networks input are expressed as $C_{\mathrm{image}} = \{c_{\mathrm{image},i}\}$, where $c_{\mathrm{image},i}$ is a vector representation of the class labels of each cell in the semantic map. C_{image} is represented by $L + 3$ dimensions: the class type labels L plus three labels for the occupancy grid, unobserved area, and unoccupied area. In this representation, the occupancy grid, unobserved area, and unoccupied area are expressed as one-hot vectors, whereas the class labels are expressed as a categorical distribution. This implementation makes the inference model training easier to converge. The values of the occupancy grid and unobserved area are not updated by the inference model, but their information is used as an input to deal with the room shapes in the semantic map. The function for converting the spatial concept C into the input data format is defined by the following equation:

$$C_{\mathrm{image}} = I(m, \{x_t\}, C). \tag{12}$$

When training the inference model of the semantic map in the pix2pix framework, the objective function is expressed as follows:

$$V(G) \propto JSD(p_{\mathrm{data}}(C_{\mathrm{image}} \mid \sigma) \,\|\, p_g(C_{\mathrm{image}} \mid \sigma)), \tag{13}$$

where $JSD(A\|B)$ is the Jensen-Shannon divergence (JSD) between distributions A and B, σ the condition, $p_{\mathrm{data}}(x)$ the training data distribution, and $p_g(x)$ the distribution of the data generated by the generator. CGAN is trained to estimate the $p_g(C_{\mathrm{image}} \mid \sigma)$ that minimizes this objective function.

The output of the inference model is represented by $C'_{\mathrm{image}} = \{c'_{\mathrm{image},i}\}$. The values of the occupancy grid and unobserved area are overwritten by the values used for the inference model input because the environment map is known. In cells that are either unoccupied area or class labels in the input data, the index that is the largest in the output vector is used to estimate the semantic label.

D. Semantic map estimation using Gibbs sampling

SpCoMapGAN estimates C, θ_l, and w_l by Gibbs sampling. First, the sampling equations for θ_l and w_l are the same as the model presented in Section III-B.

Next, the equation for sampling C is:

$$C \sim p(C \mid m, \gamma, \{f_t\}, \{s_t\}, \{x_t\}, \Theta, W), \tag{14}$$

where

$$C_{\mathrm{tmp}} = I(m, \{x_t\}, \underset{C}{\mathrm{argmax}}\, p(C \mid \pi, \{f_t\}, \{s_t\}, \Theta, W)). \tag{15}$$

It is necessary to determine π for the alternative simple model. The number of cells is counted for each class from the sampling result of C in the previous cycle to obtain the multinomial distribution, and a pseudo π is calculated from the Dirichlet distribution. Eq. (14) used for sampling the spatial concept C is approximated by the inference model described in Section III-C and can be written as follows:

$$\begin{aligned} C &\sim p(C|\pi, m, \gamma, \{f_t\}, \{s_t\}, \{x_t\}, \Theta, W) \\ &\approx p(C|C_{\mathrm{tmp}}, \gamma). \end{aligned} \tag{16}$$

Finally, Eq. (16) is expressed by the generator trained in Eq. (13) approximately as follows:

$$p(C \mid C_{\mathrm{tmp}}, \gamma) \approx p_g(C_{\mathrm{image}} \mid \sigma = C_{\mathrm{tmp}}). \tag{17}$$

IV. Experiments

We verify the validity of the proposed method in simulation and real environments.

A. Experimental dataset

We used the HouseExpo dataset [21], which is a large-scale image dataset of 2D indoor layouts generated from the SUNCG dataset [22]. It includes 25 types of class labels. We used 5,000 images as training data, 200 images as model validation data, and 100 images as test data. We used the validation data to determine the parameters of the model. As the observation obtained by the robot could be mixed with noise, we prepared a noisy dataset to reproduce the observation noise in addition to the normal dataset.

In the dataset, as the spatial concepts C_{tmp} estimated by the robot could not be prepared, we randomly extracted labels from the dataset to use them as the spatial concepts C_{tmp} acquired by the robot. The percentage of labels extracted is the assignment rate. The assignment rate of the training data used for training was set to 10%. As pseudo observations, we used place images from Places365 [17] and Google image searches, and the word information from the place labels in the dataset. Moreover, considering the noise when the robot operates in the real environment, we prepared noisy data in which wrong labels were mixed with labels randomly extracted. These noisy data were used for both the training and test data, and contained 10% of wrong labels.

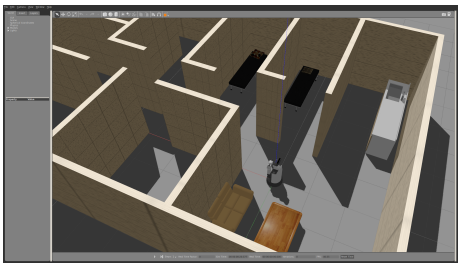


Fig. 4. A Gazebo simulation environment used in the experiment.

B. Training networks

We trained the semantic map inference model in the pix2pix framework. The network structure used U-net [8] for the generator, similarly to pix2pix [7]. For the semantic mapping, we want to maintain some features from the encoder to the decoder, such as the occupancy grid, unobserved areas, and semantic labels estimated from the observations in the occupancy grid map. Therefore, we used skip connections in the generator to retain the information.

We used SNGAN as the discriminator [20]. This spectral normalization approach helps the convergence of GANs that are difficult to train. Moreover, the spectral normalization can improve the generator accuracy [23]. Therefore, we also used spectral normalization in the generator.

The networks training time was 86.5 h for 20,000 epochs when using an Intel Core i9 7980XE CPU combined with a Nvidia Quadro GV100 GPU. The implementation was realized on Keras and TensorFlow, the optimizer was Adam, and the learning rate was 0.0002. The inference network structure is shown in Fig. 2.

C. Experiment I: Semantic mapping in simulation

We conducted an experiment with the robot performing semantic mapping in a simulation environment and compared the accuracy with the results of previous studies.

1) Conditions: We constructed 10 environments with Gazebo [24], randomly selected from the test data, to be used as the simulation environments. In addition, we used a virtual model of the Toyota Human Support Robot (HSR) [25]. Fig. 4 shows an example of our simulation environment. The comparison included the following:

(A) SpCoMapGAN (proposed)
(B) SpCoMapping [2]
(C) SpCoA [11]

In the experiment, a human first moved the robot in the simulation environment while it collected images and positions. The words were sent directly to the robot using Robot Operating System (ROS) [26] topics, assuming a state in which speech recognition could be performed reliably.

We used accuracy and adjusted Rand index (ARI) in the evaluation methods. We compared two types of accuracy: word accuracy, which was the probability that the correct place name would be obtained for each cell, and class accuracy, which was the probability that the correct region of the class was estimated for the place name. Moreover, the precision and recall were obtained for the class estimation results, and the f-measures were compared.

Furthermore, because we intended to demonstrate that it was possible to estimate the information of the unobserved region from the information of the observed region with the proposed method using transfer learning, we performed the experiment while deleting some of the label information of the test data. We defined the time at which humans provided the observation to the robot in the environment as T, and the missing experimental data consisted of the observation up to time $T/2$. T was the time took by the robot to visit all rooms and generate a map in the environment.

2) Results: Fig. 5 shows examples of the estimated semantic map for each method. Table II shows the evaluation results of the experiments, where Acc. means accuracy. Under both experimental conditions, the proposed method obtained superior results in terms of word accuracy and ARI compared to previous methods. As the word accuracy was the probability that the correct place name was obtained for each cell, the proposed method could accurately estimate the name of the robot self-position in the entire environment map, including pixels without observation, compared to previous methods. In addition, the ARI indicated that each pixel could be clustered correctly. Therefore, these results demonstrate that the proposed method can estimate the shape of spatial concepts more accurately than previous methods. Furthermore, when comparing the experimental conditions with and without missing observation, the differences in the ARI were remarkable with missing observation. This result indicates that SpCoMapGAN can improve the clustering accuracy in the unobserved regions by using the information of the observed regions.

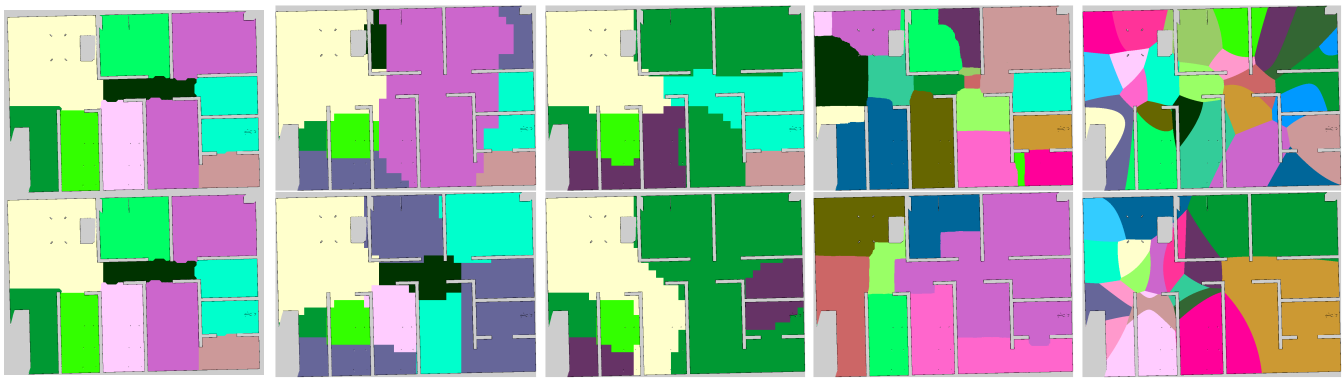
In the experimental condition with no missing observation, the previous methods exhibited superior class accuracy. However, the accuracy of the proposed method was superior to that of previous methods under the missing observation. This indicates that the proposed method can also estimate the unobserved regions by transfer learning from the observed environment. The accuracy of the proposed method was also effective in the f-measure results, and the usefulness of the proposed method was demonstrated in the accuracy rate and coverage of the estimated range.

Moreover, the accuracy was better for the models trained with noisy data than those trained without noise, whereas the ARI was better for the models trained without noise than those trained with noisy data. This demonstrates that the accuracy is improved when correcting the clustering error of the multimodal information by training with noisy data. However, as the irregular shapes of the environment were misunderstood as noise and erased, the ARI was decreased.

D. Experiment II: Semantic mapping in real space

We conducted an experiment to determine whether semantic mapping could be performed in a real environment.

1) Conditions: The inference model was trained with the HouseExpo dataset, without additional noise, using a 10% assignment rate. We used a laboratory room that replicates a home as the experimental environment. Moreover, we used the Toyota HSR as the robot. The place labels



(a) Ground truth (b) SpCoMapGAN trained using the training data (c) SpCoMapGAN trained using the noisy training data (d) SpCoMapping (e) SpCoA

Fig. 5. Experiment I: Examples of map completion using each method. First row: Estimation using all observation. Second row: Estimation using missing observation. In the map of (a), (b), and (c), the relationship among colors and labels is corresponding. In the map of (d) and (e), the relationship among colors and labels is not corresponding, but the same color in the same map has the same label.

TABLE II

EXPERIMENT I: SEMANTIC MAPPING RESULTS IN SIMULATION.

Methods	All observation				Missing observation			
	word Acc.	class Acc.	class f1	ARI	word Acc.	class Acc.	class f1	ARI
SpCoMapGAN	0.611	0.823	0.481	**0.491**	0.444	**0.791**	0.259	**0.405**
SpCoMapGAN (noise)	**0.635**	0.834	**0.484**	0.446	**0.456**	0.773	**0.294**	0.334
SpCoMapping	0.299	**0.845**	0.404	0.408	0.284	0.751	0.237	0.117
SpCoA	0.391	0.822	0.329	0.261	0.248	0.774	0.164	0.150

were `living_room`, `office`, `kitchen`, `dining_room`, `entryway`, and `meeting_room`. We used three sentences extracted from the sentence collection of each place label found in an English dictionary site[1] as human utterance data. Here, the proper nouns included in the human utterance data were changed to use the same ones overall. The total number of teachings was 124. The image information was extracted using the AlexNet [27] CNN trained with the Places205 dataset [28], and the word information was weighted using tf-idf [29] after converting the utterance data into bag-of-words representations.

2) Results: Fig. 6 shows the semantic map estimated in the real environment and the five words obtained in each region in order of decreasing probability. The experimental results demonstrate that SpCoMapGAN is useful for semantic mapping, even in a real environment.

Regarding the acquired words, the robot associates multiple words to the place regions. For example, in the upper right table of Fig. 6 (blue area), in addition to the `meeting_room` originally given as a place label, words strongly related to the meeting room, such as *groups*, *teamwork*, and *brainstorming*, were estimated at the top. This means that a particular place can be indicated using words other than the pre-prepared place labels by using local information. Even when tf-idf was used, general words such as *she* and *was* were mapped to regions, but as the number of sentences in the utterance data increased, the weight and probability of the general words decreased.

Focusing on the map segmentation, `living_room` and `dining_room` were considered to be classified into the same category because the image information was similar, as indicated in the lower left of Fig. 6(a).

[1]https://www.yourdictionary.com/

V. CONCLUSIONS

We proposed SpCoMapGAN which performs semantic mapping in a newly visited environment using a network that approximates the joint distribution of the classes of all cells trained by the semantic maps of a large number of known environments in the pix2pix framework. Experiments in a simulation environment indicated that the proposed method could transfer features from the already estimated semantic maps, such as the relationships between rooms, to a newly visited environment. Moreover, the accuracy was remarkably improved when estimating the unobserved regions from information gathered in observed regions by using the joint distribution of the classes of all cells.

However, although SpCoMapGAN was able to deal with room shapes to a certain extent, regions that straddle walls and obstacles were sometimes inaccurately estimated because occupancy grids and unobserved areas are input into the network as images to model the dependence on all cells. This problem could be solved by integrating MRF similarly to SpCoMapping.

As future work, we will improve the model by integrating SpCoMapGAN with SLAM so that semantic mapping can be performed even when the map of the environment is unknown.

APPENDIX

A. Experiment to compare the inference models to a rule-based algorithm

We compared the accuracy of the semantic mapping with a rule-based algorithm using several datasets to demonstrate the validity of each inference model.

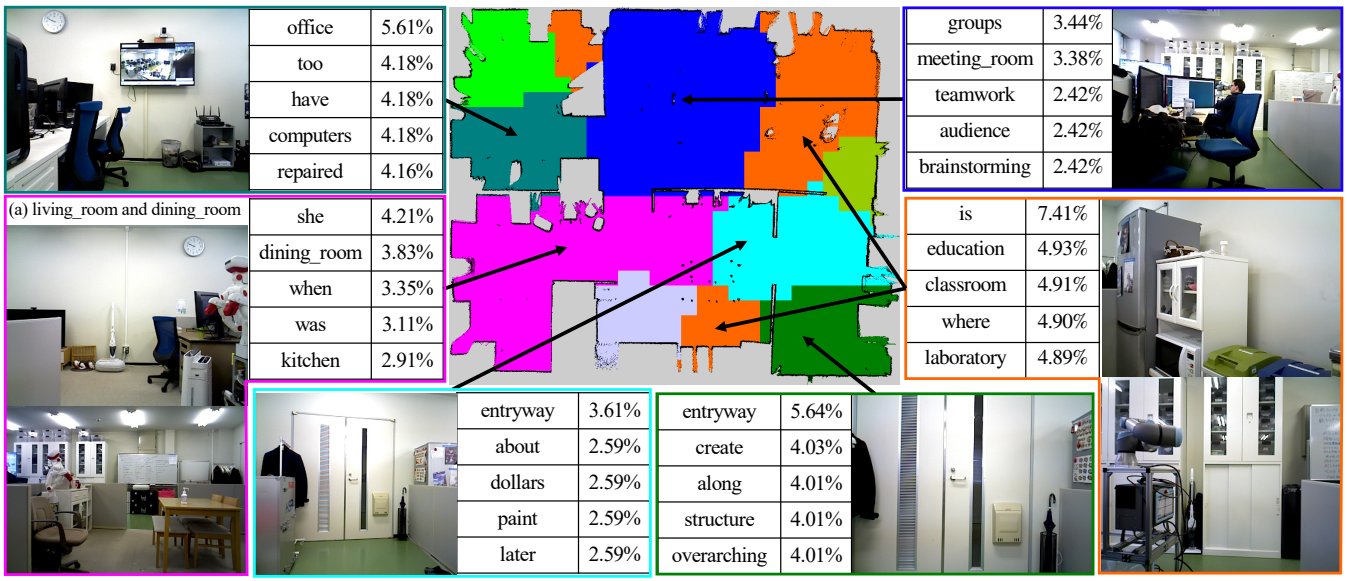


Fig. 6. Experiment II: Semantic map estimated in a real environment. The tables show the five words obtained in each region in decreasing probability.

1) Conditions: We prepared the HouseExpo dataset [21], described in Section IV-A, and the HOME'S dataset [30]. The HOME'S dataset includes floor plan images of apartments in Japan. It also includes six types of class labels. We used 3,500 images as training data, 200 images as model validation data, and 100 images as test data.

The comparison included the following: (i) SNGAN, (ii) U-net, and (iii) nearest neighbor (NN). SNGAN and U-net used weights trained on 10,000 iterations with the training data. NN used the closest semantic label in the condition data provided to the cells as the estimation result for each cell.

The evaluation method included both accuracy and ARI. For the test data, the assignment rates of the HouseExpo were 2.5%, 5%, and 10%, whereas the assignment rates of the HOME'S dataset were 2.5% and 10%. Both of these used 100 data with and without noise.

In this experiment, we also compared each result with and without missing observation. In the information-missing dataset, half of the labels for each environment were randomly deleted.

Fig. 7 shows an example of the correct labels and condition data under each experimental condition.

2) Results: Fig. 8 shows an example of each result of the semantic mapping. Table III shows the evaluation results of accuracy and ARI.

Comparing the results of SNGAN and U-net, which were trained as inference models, and NN, which is a rule-based algorithm, NN showed higher values when the assignment rate in the HouseExpo dataset was lower. However, when the assignment rate was high, the result of the inference model improves the accuracy and ARI. Because the network was trained using data with an assignment rate of 10%, the accuracy of the inference model was increased as the difference between the training and test data decreased. With the HOME'S dataset, the performance when using the inference model was superior than when using the rule-based

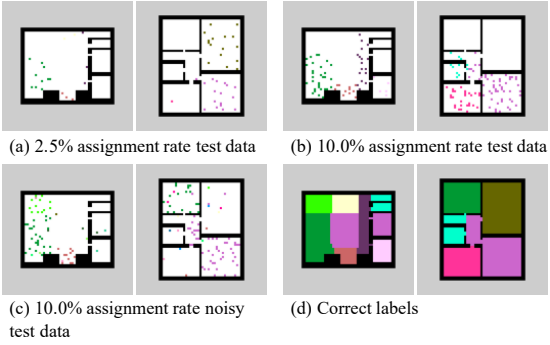

(a) 2.5% assignment rate test data
(b) 10.0% assignment rate test data
(c) 10.0% assignment rate noisy test data
(d) Correct labels

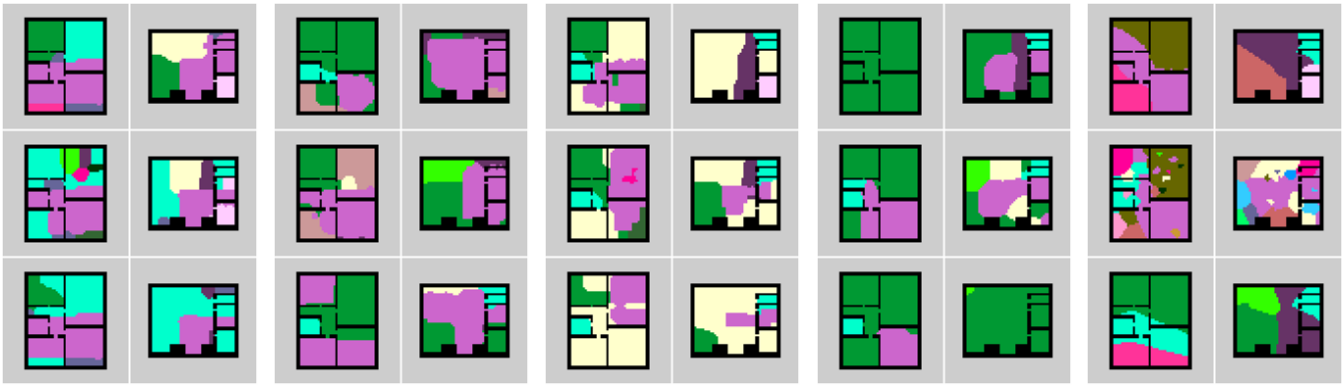
Fig. 7. Test data under each condition and their correct labels.

algorithm at 2.5% and 10%. Because the HOME'S dataset is made of real floor plans of existing houses, the network is effective when performing semantic mapping in a real environment.

For the two methods using the inference model, the results of the SNGAN and U-net exhibited no significant quantitative difference overall. However, when investigating the data that were actually generated, U-net often supplemented information in the missing data area with labels that existed in the environment, whereas SNGAN used the information of labels that did not exist in the environment from the surroundings. This can be confirmed from the fact that the accuracy of SNGAN was superior to that of U-net in the experimental results using HouseExpo at 10%, when the assignment rate of the condition data was the biggest and the estimation error of the missing data area was reduced. As GANs have a discriminator that estimates the distance between the distribution of the data generated by the generator and that of the training data, we consider that the performance of modeling the dependence between each region in the environment, as if it was real, was high.

References

[1] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Robotics and Autonomous Systems*, vol. 66, pp. 86–103, 2015.



(a) SNGAN trained using the training data (b) SNGAN trained using the noisy training data (c) U-net trained using the training data (d) U-net trained using the noisy training data (e) NN

Fig. 8. Examples of map completion using each method. From upper row: Test data of 2.5% assignment rate, test data of 10% assignment rate, and noisy test data of 10% assignment rate.

TABLE III

EXPERIMENTAL RESULTS FOR EACH INFERENCE MODEL.

Methods	HouseExpo						HouseExpo with noise						HOME'S				HOME'S with noise			
	2.5%		5%		10%		2.5%		5%		10%		2.5%		10%		2.5%		10%	
	Acc.	ARI	Acc.	ARI	Acc.	ARI	Acc.	ARI	Acc.	ARI	Acc.	ARI	Acc.	ARI	Acc.	ARI	Acc.	ARI	Acc.	ARI
SNGAN	0.497	**0.375**	0.576	**0.523**	**0.578**	0.481	0.376	**0.301**	0.508	**0.428**	0.534	**0.478**	0.339	0.168	0.338	0.161	0.310	0.173	0.359	0.199
SNGAN (noise)	0.412	0.250	**0.588**	0.458	0.552	0.416	0.371	0.246	**0.530**	0.404	**0.536**	0.398	**0.681**	0.485	0.737	0.550	**0.629**	0.459	**0.687**	0.529
U-net	0.403	0.244	0.536	0.479	0.573	0.466	0.309	0.257	0.495	0.348	0.503	0.420	0.235	0.169	0.298	0.201	0.247	0.159	0.341	0.239
U-net (noise)	0.388	0.202	0.515	0.434	0.551	**0.509**	0.329	0.148	0.485	0.353	0.503	0.434	0.666	**0.513**	**0.741**	**0.576**	0.628	**0.510**	0.649	**0.532**
NN	**0.532**	0.273	0.556	0.364	0.552	0.343	**0.378**	0.264	0.369	0.319	0.472	0.390	0.528	0.315	0.494	0.305	0.406	0.208	0.496	0.344

[2] Y. Katsumata, A. Taniguchi, Y. Hagiwara, and T. Taniguchi, "Semantic mapping based on spatial concepts for grounding words related to places in daily environments," *Frontiers in Robotics and AI*, vol. 6, p. 31, 2019.

[3] R. Goeddel and E. Olson, "Learning semantic place labels from occupancy grids using CNNs," in *IEEE/RSJ IROS*, 2016, pp. 3999–4004.

[4] N. Sunderhauf, F. Dayoub, S. McMahon, B. Talbot, R. Schulz, P. Corke, G. Wyeth, B. Upcroft, and M. Milford, "Place categorization and semantic mapping on a mobile robot," in *IEEE ICRA*, 2016, pp. 5729–5736.

[5] A. Pal, C. Nieto-Granda, and H. I. Christensen, "DEDUCE: Diverse scene detection methods in unseen challenging environments," in *IEEE/RSJ IROS*, 2019, pp. 4198–4204.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE CVPR*, 2017, pp. 5967–5976.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[9] A. Taniguchi, T. Taniguchi, and T. Inamura, "Simultaneous estimation of self-position and word from noisy utterances and sensory information," *IFAC-PapersOnLine*, vol. 49, no. 19, pp. 221–226, 2016.

[10] S. Isobe, A. Taniguchi, Y. Hagiwara, and T. Taniguchi, "Learning relationships between objects and places by multimodal spatial concept with bag of objects," in *ICSR*, 2017, pp. 115–125.

[11] A. Taniguchi, T. Taniguchi, and T. Inamura, "Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 4, pp. 285–297, 2016.

[12] ——, "Unsupervised spatial lexical acquisition by updating a language model with place clues," *Robotics and Autonomous Systems*, vol. 99, pp. 166–180, 2018.

[13] L. F. Posada, A. Velasquez-Lopez, F. Hoffmann, and T. Bertram, "Semantic mapping with omnidirectional vision," in *IEEE ICRA*, May 2018, pp. 1901–1907.

[14] M. Brucker, M. Durner, R. Ambrus, Z. C. Marton, A. Wendt, P. Jensfelt, K. O. Arras, and R. Triebel, "Semantic labeling of indoor environments from 3D RGB maps," in *IEEE ICRA*, 2018, pp. 1871–1878.

[15] J. C. Rangel, M. Cazorla, I. Garcia-Varea, C. Romero-Gonzalez, and J. Martinez-Gomez, "Automatic semantic maps generation from lexical annotations," *Autonomous Robots*, 2018.

[16] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv*, 2018.

[17] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[18] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[19] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training GANs," in *NeurIPS*, 2016, pp. 2234–2242.

[20] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *ICLR*, 2018.

[21] L. Tingguang, H. Danny, L. Chenming, Z. Delong, W. Chaoqun, and M. Q.-H. Meng, "HouseExpo: A large-scale 2D indoor layout dataset for learning-based algorithms," *IEEE/RSJ IROS*, p. 882, 2019.

[22] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," *IEEE CVPR*, 2017.

[23] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *ICLR*, 2019.

[24] N. Koenig and A. Howard, "Design and use paradigms for Gazebo, an open-source multi-robot simulator," in *IEEE/RSJ IROS*, vol. 3, 2004, pp. 2149–2154.

[25] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of human support robot as the research platform of a domestic mobile manipulator," *Robomech*, vol. 6, no. 4, 2019.

[26] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: An open-source robot operating system," in *ICRA Workshop on Open Source Software*, 2009.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[28] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014, pp. 487–495.

[29] G. Salton and M. J. Mcgill, *Introduction to modern information retrieval*. McGraw-Hill, 1986.

[30] National Institute of Informatics and LIFULL Co., Ltd., "LIFULL HOME'S dataset," 2015.