一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

Unknown Object Segmentation by View Independent Scene Change Detection

Jiaxin LI[†]　Yasutomo KAWANISHI[‡ †]　Daisuke DEGUCHI[†]　and　Hiroshi MURASE[†]

† Graduate School of Informatics, Nagoya University　Furo-cho, Chikusa-ku, Nagoya, 464-8601 Japan
‡ RIKEN Guardian Robot Project　2-2-2 Hikaridai Seika-cho, Sorakugun, Kyoto 619-0288 Japan
E-mail:　† jiaxinl@vislab.is.i.nagoya-u.ac.jp {deguchi, murase}@nagoya-u.jp
‡ yasutomo.kawanishi@riken.jp

Abstract　Exploring the indoor environment and finding unknown objects that appeared in a scene is important for scene understanding by a robot. While background subtraction is traditionally used for segmenting unknown object regions, it cannot be directly used for a moving camera on the robot. In this paper, we address a task called view-independent panoptic scene change detection, which is the task of segmenting unknown object regions by comparing two images from different viewpoints before and after the objects appear. We propose a method for segmenting unknown object regions by modeling each segmented known instance region as a background. For the background modeling, we propose a deep metric-learning-based method. In addition, we create a new panoptic scene change detection dataset consisting of images taken from different camera viewpoints. Through experiments, we confirm that the proposed method can segment regions of unknown class objects; the deep metric-learning-based method performs more accurately than a simple histogram-based method, achieving good performance on the change detection dataset.

Keywords　Scene change detection, Unknown object, Background modeling, Deep metric learning

1. INTRODUCTION

In 2020 David Hall et al. proposed The Robotic Vision Scene Understanding Challenge [1], which is dedicated to evaluating how well a robotic vision system understands the semantic and geometric aspects of the indoor environment. They proposed one task called Scene Change Detection. In this task, the robot is required to enter a scene and establish a semantic background building of that scene first. Then, the robot is moved to a different, new location in the same environmental scene. They mention that there are possibilities for the addition or removal of objects from the new scene. Finally, a description of the changes that appeared before and after this scene was required.

We assume that RGB image data is captured using a camera mounted on a robot for the Scene Change Detection task. The semantics of the environment will be constructed by performing some processing on the acquired RGB images. For image processing, with the development of computer technology, in recent years, deep neural networks have been actively studied in many image-processing tasks, such as image recognition, image segmentation, and others. Among them, based on deep learning, image segmentation tasks have been widely researched. However, the existing segmentation methods can only identify the labeled categories that appear in training and cannot identify unknown categories not included in the training dataset. In the actual situation, there are many unknown objects that are not included in the training set; that is called an open-set setup. The open-set setup requires segmenting instances of unknown categories [2].

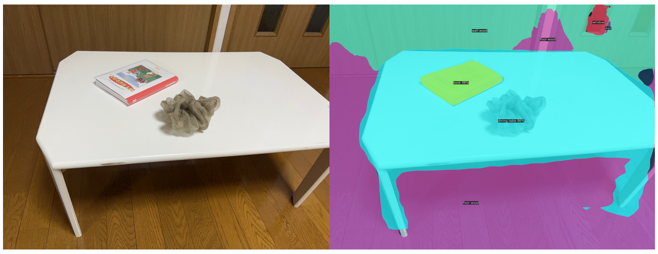


Figure1. Result of panoptic segmentation and original image.

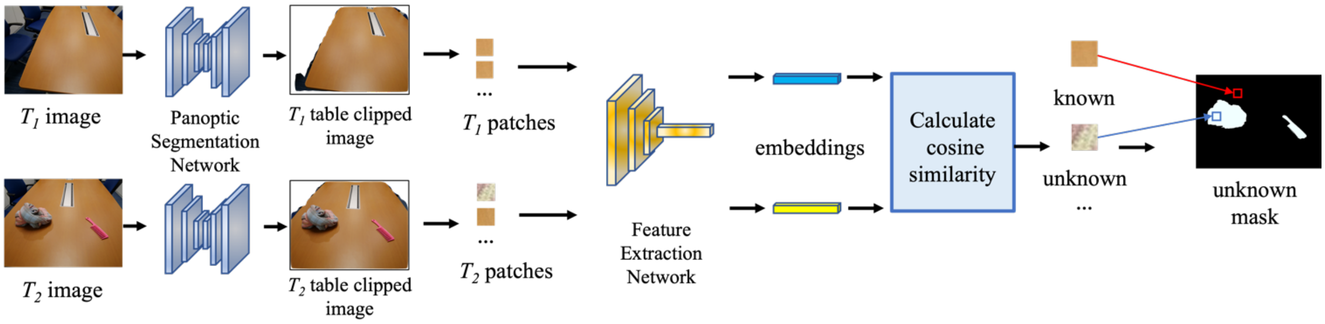
In this paper, we propose a method that detects these unknown instances based on change detection between two images before and after unknown class instances appear.

2. RELATED WORKS

2.1. Open-set Segmentation

Open-set exists relative to close-set. Chuanxing Geng et al. [3] refer to a close-set by describing that under a common closed-set, there is the assumption that the training and test data come from the same label and feature space. This means that the kinds of training and test data sets overlap exactly. All classes in the test data will have been observed during training.

In recent years, the open-set segmentation problem has been widely studied. Jaedong Hwang et al. proposed a novel exemplar-based open-set panoptic segmentation



T_1 image
Panoptic Segmentation Network
T_1 table clipped image
T_1 patches
T_2 image
T_2 table clipped image
T_2 patches
Feature Extraction Network
embeddings
Calculate cosine similarity
known
unknown
unknown mask

Figure2. The processing of deep metric learning-based method.

network (EOPSN) [4] to solve the open-set panoptic segmentation problem. However, this network is not suitable for the task mentioned earlier because it needs a large number of images to mine more unknown objects belonging to the same class, which is difficult to achieve under the task conditions mentioned. Because there are a large number of unknown object classes in real life, it is impossible to extract features from the mined unknown objects and cluster them by only one environment image after the change has appeared.

2.2. Change Detection

As described in the previous section, there is the problem of too many classes of unknown objects. Therefore, we follow the task of change detection, which detects changed objects in the scene from two images captured at different times in different angles. Change detection refers to the process of identifying changes in an image over time. This can include changes in the appearance, position, or number of objects in a scene [5].

However, since traditional methods compare two images to detect changes between them, they expect that the inputs are well aligned, and they cannot be directly applied to the scene change detection task for a moving robot. That's because remote sensing is easy to align since the distance to the ground is far, and the targets are flat, while the distance is close to the camera and the objects are 3D in our problem setting. Therefore, the difference in viewpoint's angle has a large impact.

2.3. Panoptic Segmentation

Our proposed method is based on instance-wise segmentation. To segment the regions that obtain the same attributes before and after the change appears, it is feasible to use panoptic segmentation methods.

Panoptic segmentation is a task that combines both instance segmentation and semantic segmentation in a model. Panoptic segmentation not only detects and segments individual objects but also labels each pixel with a class label [6]. However, there is one problem, if an unknown class object which has a smaller size is placed on a known class object, its pixels will be assigned the known class labels. For example, in Fig.1, when a book and an object of an unknown class appear on the dining table at the same time, the book is segmented, but the pixels of the unknown class object are assigned the semantic of the dining table.

3. PROPOSED METHOD

3.1. Overview

Based on the scene change detection task, we propose a new task called the view-independent panoptic scene change detection task, which requires segmenting small unknown object regions by comparing two images from different viewpoints before and after the objects appear. Meanwhile, we propose a method to solve this task called the deep metric learning-based method. We further segment small unknown object regions by comparing two images even from different viewpoints before and after the objects appear in the proposed task. In this research, we assume these unknown objects are much smaller in size than the known objects that placed them already included in the data used to train the panoptic segmentation network. This ensures that when segmentation obtains regions with the same attributes, the unknown objects that appear after the change are included in the region that is desired to be observed.

Overall, consider the image before the appearance of the unknown objects as T_1 image, and the one after the appearance as T_2 image. As mentioned before, if a small unknown object appears in T_2 image, through the panoptic segmentation network, it will be segmented in the known instance region. This known instance region will be compared with the same known instance region in T_1 image.

The difference is detected as an unknown region.

3.2. Deep Metric Learning

The method uses features extracted by a feature extraction network trained by deep metric learning. The feature extraction network is trained to make patches from the same object instance region closer and patches from different instance regions apart. The detailed steps are as follows:

First, small patch regions of the specified size are extracted from an instance clipped image of the T_I image, and the same class label is assigned to the patches. For an instance-clipped image of size $H \times W$, we crop it into patches of size $l_{patch} \times l_{patch}$. In total, N patches are obtained. The calculation formula is as follows:

$$H_N = \left\lfloor \frac{H}{l_{patch}} \right\rfloor, \quad W_N = \left\lfloor \frac{W}{l_{patch}} \right\rfloor$$

$$N = H_N \times W_N$$

Then we check all patches and delete some patches which have the void pixel above 50%. It is noted here that even for two instances of the same class (table, sofa, etc.), we classify the patches of these two instances into different classes. In deep metric learning, we only treat pairs of two patches from the same instance within the same scene as samples of the same instance. All other patches are considered dissimilar samples.

The structure of the feature extraction network refers to LeNet, proposed by Yann LeCun [7]. We remove the last softmax layer and leave only the fully connected layer to get a feature embedding. For training, we use patches of all instances as input, and they have the labels of their corresponding instances. With the feature extraction network, we will get the embedding of each patch.

3.3. Loss Function

We perform triplet mining on all the embeddings we obtained. As shown in Fig.2, we set a margin, which represents the difference between the distance of anchor-positive and the distance of anchor-negative. The negative is further from the anchor than the positive is from the anchor. After mining all the triples, we update the network using the triplet loss as the loss function.

The loss function is given by the following equation:

$$L_{triplet} = \left\lfloor d_{ap} - d_{an} + m \right\rfloor_{+}$$

where d_{ap} is the distance of anchor-positive, and d_{an} is the distance of anchor-negative. m is the value of the margin. It represents the desired difference between the distance of anchor-positive and the distance of anchor-negative. All distances are cosine similarity between anchor-positive and anchor-negative.

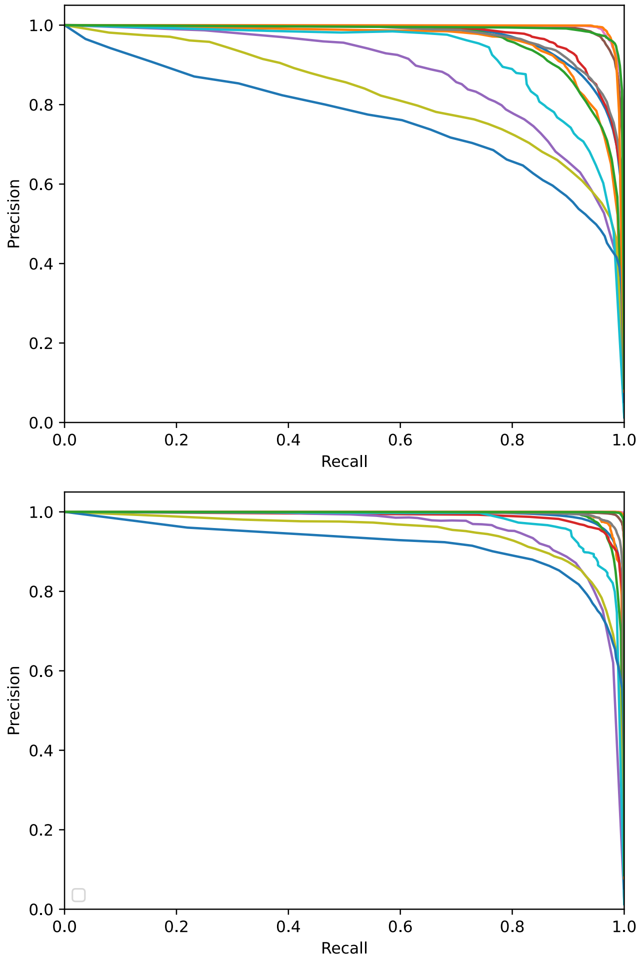


Figure3. The PR curves of different deep metric learning-based method. The different color curves represent different kinds of instances. The curve on the top is higher overall than the curve in the bottom. **Top:** without random patches. **Bottom:** with random patches.

3.4. Random Patches

Besides of input one patch image at a time to the network for training, we also consider relationships between three patches by inputting them into the network for training. For all the instance clipped images used for training, the number of instances is n. All the patches obtained will be divided into n different classes $\{C_1, C_2 \cdots C_n\}$. For patches belonging to class C_i, two additional patches will be randomly selected from the same class C_i and these three patches will be concatenated over the channel dimension, and we consider them as positive patches. Besides, we take one patch from class C_i and the other two patches from other class C_j, concatenate these three patches over the channel dimension and consider them as negative patches. For class C_i and class C_j, the generated patches are the

Table1. Result of Unknown Object Segmentation

Scene	S02		S03	
Metric	PA	mIoU	PA	mIoU
Hist	75.5	43.8	69.9	36.7
DML	92.4	72.6	82.6	71.2
DML-ran	92.5	74.9	85.3	74.7
Scene	S04		S07	
Metric	PA	mIoU	PA	mIoU
Hist	78.9	42.1	71.8	49.8
DML	94.7	64.8	90.3	78.1
DML-ran	94.0	68.5	92.1	79.6
Scene	All Scenes			
Metric	PA		mIoU	
Hist	85.3		54.4	
DML	89.9		69.2	
DML-ran	90.4		73.7	

negative patches of both of them.

We further extend the proposed method to the structure of the network according to the above additions. When training, we use the newly generated patches as input to the network, split the input tensor into three chunks over the channel dimension first and let them each pass through the first convolutional layer sharing parameters with each other. We will concatenate them again over the channel dimension before the second convolutional layer. Doing this processing during training to eliminate the effect of color changes on the surface of the instance on the calculation of cosine similarity values.

4. EXPERIMENT

4.1. Dataset

We collected a new dataset for evaluating the proposed task. In the dataset, 4,884 images were taken in 12 mimic daily life scenarios, including a cabinet where a microwave oven is placed, a table in a dining room, a sofa in a living room, and so on. In each scene, we took 3 to 4 images of the scene without the unknown objects in their original state as T_1 images. We also took a large number of images of the scene with the unknown objects from other angles as T_2 images. Meanwhile, the Panoptic-DeepLab network [8] will be trained by the MSCOCO dataset [9].

4.2. Experimental Setting

We trained the feature extraction network at first. In the training of the feature extraction network, we will use only three to four T_1 images of each scene without the unknown objects. The T_1 image will be obtained as a clipped image of various instances by the panoptic segmentation network that has been trained with the MSCOCO dataset. These images will be segmented into 30×30-sized patches. For the deep metric learning-based method, first, we pass the T_1 images without the unknown object placed and the T_2 images through the panoptic segmentation network as well. For the known instance which we want to observe, we cut it into 30×30-sized patches and store it. These patches from the T_2 image will be used to compute cosine similarity with a random one of the patches from the T_1 image in turn to determine the unknown object region.

4.3. Comparative Method

As a comparative method, we also constructed a simple histogram-based method to compare with the proposed deep metric learning-based method. The histogram-based method calculates an RGB histogram of the instance region in the T_1 image and calculates an RGB histogram for a patch in the corresponding region in the T_2 image. Then, these histograms are compared based on the Pearson correlation coefficient [10] between them. When the value of Pearson Correlation is close to 1, this patch will be considered a known instance region. In contrast, if the value is close to 0, the patch will be considered an unknown object region.

4.4. Result of Feature Extraction Network

We use the K-Nearest Neighbor algorithm [11] to classify an embedding into all known instance classes. We set the neighbor to 32 and then use the $embeddings_{test}$ obtained from the test data after passing the feature extraction network to make predictions. The precision-recall curve (PR curve) is plotted to see if the classification performance of the feature network improves after adding random patches during training.

As Fig.3 shows, it can be seen that after adding random patches, the overall PR curve of embeddings obtained by the feature extraction network is higher than that before adding. This demonstrates that the performance of the network is better in solving the classification problem if the embedding obtained from multiple random patches is used in training the feature extraction network followed by deep metric learning.

4.5. Result of Unknown Object Segmentation

The experiment results obtained under some scenes are listed in Table 1 and Fig.4. From Fig. 4, it can be observed that the two methods based on deep metric learning visually have better performance than the histogram-based method. For the parts of the T_2 clipped image that are mis-

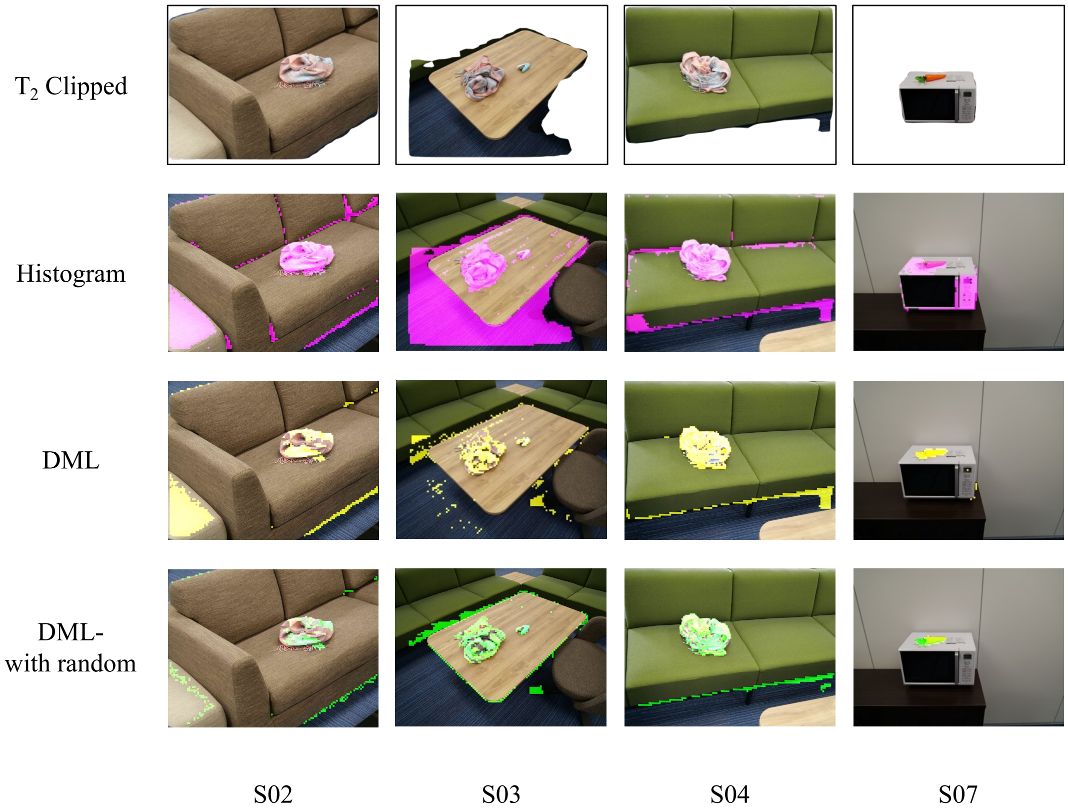


Figure4. From top to bottom, they are T_2 clipped images, results of histogram-based method, results of deep metric learning-based method and results of deep metric learning-based method adding random patches.

segmented regions around the observation area due to the results of the existing panoptic segmentation network, the histogram-based method treats them as unknown object regions. Meanwhile, the seams of the sofa, the patterns of the table, and the surface of the microwave oven are mis-segmented more than the other two methods. By comparing the two methods based on deep metric learning, we can see a significant reduction in the number of regions incorrectly segmented as unknown object regions in the seams, patterns, and edges affected by the segmentation network after adding random patches to the training.

In this research, we also use two semantic segmentation evaluation metrics, Pixel Accuracy (PA) and mIoU, to evaluate the proposed method. From Table 1, it can be seen that the PA and mIoU of the two methods based on deep metric learning are higher than those of the histogram-based method. Meanwhile, the PA and mIoU are slightly higher after adding random patches.

5. CONCLUSION

In this paper, we focus on the task called view-independent panoptic scene change detection. We expect to segment unknown objects not included in the training data in this task. We proposed the method of segmenting small unknown object regions by comparing two images from different viewpoints before and after the unknown objects appear. In this paper, we propose a deep metric learning-based method for segmenting unknown object regions by modeling each segmented known instance region as a background. Also, a further improved method that adds random patches for training based on the deep metric learning-based method is also proposed.

For the experiment, we built our own dataset and used the PR curve to observe and evaluate the proposed methods. For a comparative method, we implemented a histogram-based method. Meanwhile, we use the commonly used semantic segmentation evaluation metrics PA and mIoU to evaluate the method and confirm that the deep metric learning-based method can effectively segment the unknown objects by comparing two images from different viewpoints before and after the unknown objects appear. Besides, the performance is better than the histogram-based method.

For future work, we will consider using some other deeper feature networks to extract more advanced, abstract features in the deeper layers of the network. Besides, we consider retraining the existing panoptic segmentation network using the newly captured dataset, hoping to reduce the error edge of the panoptic segmentation network in segmenting the instances. In this paper, we assume that the object instances are identified and matched across frames; however, it is not a trivial problem. We will tackle re-identification for generic objects.

ACKNOWLEDGMENT

Parts of this work was supported by JST, CREST Grant Number JPMJCR22D1, Japan. Parts of this research were supported by Grant-in-Aid for Scientific Research JP21H03519.

REFERENCES

[1] D. Hall, B. Talbot, S.R. Bista, H. Zhang, R. Smith, F. Dayoub, and N. S¨underhauf, "The robotic vision scene understanding challenge," arXiv preprint arXiv:2009.05246, 2020.

[2] W. Wan, M. Feiszli, H. Wang, J. Malik and D. Tran. "Open-World instance segmentation: exploiting pseudo ground truth from learned pairwise affinity," Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4412-4422, 2022.

[3] C. Geng, S.-j. Huang, and S. Chen, "Recent advances in open set recognition: A survey," IEEE transactions on pattern analysis and machine intelligence, vol.43, no.10, pp.3614–3631, 2020.

[4] J. Hwang, S.W. Oh, J.-Y. Lee, and B. Han, "Exemplar-based open-set panoptic segmentation network," Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.1175–1184, 2021.

[5] L. Ma, M. Li, T. Blaschke, X. Ma, D. Tiede, L. Cheng, Z. Chen, and D. Chen, "Object-based change detection in urban areas: The effects of segmentation strategy, scale, and feature space on unsupervised methods," Remote Sensing, vol.8, no.9, p.761, 2016.

[6] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Doll´ar, "Panoptic segmentation," Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9404–9413, 2019.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol.86, no.11, pp.2278– 2324, 1998.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C.L. Zitnick, "Microsoft COCO: Common objects in context," Computer Vision – ECCV 2014, part V, pp.740–755 2014.

[9] B. Cheng, M. D Collins, Y. Zhu, T. Liu, T. S Huang, H. Adam, and L. Chen. "Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12472-12482, 2020.

[10] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," Noise reduction in speech processing, pp.1–4, 2009.

[11] J.M. Keller, M.R. Gray, and J.A. Givens, "A fuzzy k-nearest neighbor algorithm," IEEE transactions on systems, man, and cybernetics, vol.SMC-15, no.4, pp.580– 585, 1985.