2021 The 21st International Conference on Control, Automation and Systems (ICCAS 2021)
Ramada Plaza Hotel, Jeju, Korea, Oct. 12∼15, 2021

Semantic Mapping Based on Image Feature Fusion in Indoor Environments

Cong Jin[1,2], Armagan Elibol[2], Pengfei Zhu[1], and Nak Young Chong[2*]

[1]College of Intelligence and Computing, Tianjin University
Jinnan, Tianjin 300350, China (jciwqh@163.com, zhupengfei@tju.edu.cn)
[2]School of Information Science, Japan Advanced Institute of Science and Technology
Nomi, Ishikawa 923-1292, Japan ({s2010233, aelibol, nakyoung}@jaist.ac.jp) * Corresponding author

Abstract: It is of the utmost importance for the robot to understand human semantic instructions in human-robot interaction. Combining semantic information with SLAM-based maps leads to a semantic map. Deep neural networks are able to extract useful information from the robot's visual information. In this paper, we integrate the RGB feature information extracted by the classification network and the detection network to improve the robot's scene recognition ability and make the acquired semantic information more accurate. The image segmentation algorithm labels the areas of interest in the metric map. Furthermore, the fusion algorithm is incorporated to obtain the semantic information of each area, and the detection algorithm recognizes the key objects in the area. We have demonstrated an efficient combination of semantic information with the occupancy grid map toward accurate semantic mapping.

Keywords: Semantic mapping, Deep learning, Scene recognition, Image feature fusion

1. INTRODUCTION

Over the past decades, robots have been introduced in our daily lives with different roles and tasks. Recently, humanoids are being seen as a key role player in assisted living facilities. In order for the robot to work well in the indoor environment, in addition to the navigation ability integrated with the (metric) map building ability, it is necessary to recognize the semantic information of the relevant scene. For example, when we issue a command to the robot "please help me get a cup in the kitchen", the robot needs to understand where the kitchen is and be able to understand what a cup is to complete the task. This requires the robot to obtain the semantic information of the environment through scene recognition or object detection, and pass the label to the metric map, which is known as semantic mapping. Scene classification plays a vital role in the generation of semantic maps. There have been several methods proposed on scene classification for generating semantic maps such as traditional image feature point matching or deep learning algorithms. In our work, we propose a deep-learning based novel feature fusion method, which achieves the state-of-the-art performance in indoor scene classification combining object detection to enrich semantic information.

The contributions of this work can be summarized as follows: (1) The proposed method has good performance in indoor scene classification, the accuracy rate exceeds the baseline, and the model parameters are small, which can be well transferred to embedded devices. (2) We added the task of object detection in the scene recognition process, which enriched the semantic information.

The remainder of this paper is organized as follows. After discussing related work in the following section, Section III proposes our feature fusion method based on deep learning and details the semantic mapping process. Section IV describes the results and analysis of the experiments. Section V draws a conclusion with future work.

2. RELATED WORK

Using visual information to infer semantic location classification has become an important field in robotic applications. Rottmann *et al.* [1] trained the classifier by the features extracted from the visual and laser ranging data, and enhanced the robustness of the scene classifier using the Markov model. Visual features of histogram were used in [2].

With the advances in deep learning, a variety of methods using neural networks are increasingly playing a significant role in scene recognition. Sünderhauf [3] used the AlexNet [4] to train the Place205 dataset for scene classification. Embedding the classification system into a Bayesian filter framework, previous domain knowledge can be merged, and the framework can ensure the consistency of time. Place205 annotates 205 kinds of scene data, and uses deep neural network for classification training. This paper uses SLAM to generate 2D grid map, and then input the image provided by the vision system into the classification network to get the environment, and then add the environment semantic information into the grid map. Brucker *et al.* [5] proposed a method to obtain 3D-RGB from rooms for map reconstruction and semantic label assignment. Using the deep learning technology, according to the automatically generated virtual RGB view and the geometric analysis of the 3D structure of the map, the scene is classified and the object is detected, so as to get the room type. Rangel *et al.* [6] used the information provided by vocabulary annotation to generate semantic images from RGB images acquired. Pal *et al.* [7] combined the object detection and scene recognition algorithm, and designed 5 models for experimental comparison. They used the Place365 [8] dataset to train resnet18 [9] network for scene classification of 7 indoor scenes. In addition, YOLOv3 [10] was used to train the COCO dataset [11] for object detection to assist in detecting scene classification model. When the scene

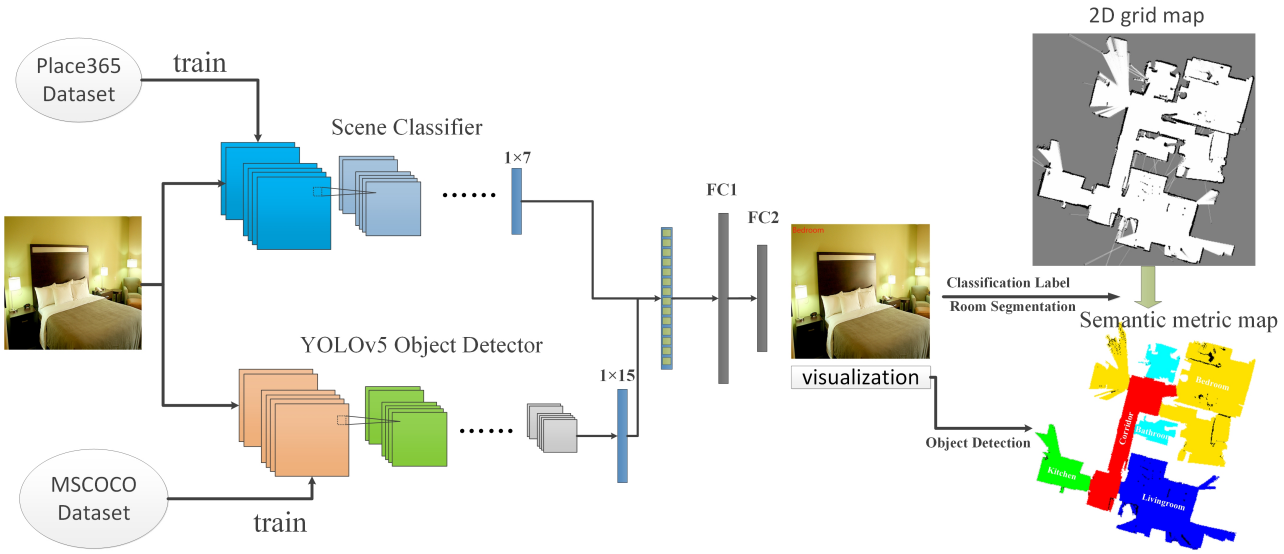

Fig. 1. System architecture with the Yolov5 object detector, multiple classification networks, and the EfficientNet-B2 network. We segment the grid map, obtain the scene information of each area using the scene classifier, and generate a semantic metric map. We use the detector to obtain the object information of each area to enrich the semantic map.

confidence is lower than the threshold, the object detection model can be used to identify the scene. Using these large datasets, label information can be detected even in unknown environment, without retraining the model according to the specific environment.

3. METHODOLOGY

The robot can create occupancy grid maps through SLAM. Our task is to use the RGB images to create a semantic map based on the grid map. The proposed semantic mapping combines two parts: a portable indoor scene classification model and an indoor object detection model. In the classification model, we compared the SOTA networks in ImageNet competitions, and chose a model with precision and speed trade-off. In the detector, we use the YOLOv5 model, which has proven to be accurate and fast. We fused the image features of the classifier and the detector as the input layer to output the classification results under the action of the two fully connected hidden layers, thereby obtaining the semantic label of the current image. Finally, when making a semantic map, we can use the robot's pose estimation and laser scanning data to create a semantic map. We chose the MAORIS [12] algorithm of map segmentation to segment the metric map first. After the map segmentation, we fill in the corresponding segmentation area according to the semantic information of the image. In addition, we use 35 object categories as part of the map information to enrich semantic information. The specific process is illustrated in Fig.1.

3.1 Selection of scene classification extractor

Regarding the classification of indoor scenes, we chose the same 7 indoor scenes as in [7], namely bathroom, bedroom, corridor, kitchen, living room, dining room, and office. The scene dataset uses the scene pictures of Place365 [8], which contains 5,000 training sets for each category. We originally planned to use all the data for training to ensure that the robot can recognize the meaning of the scene in most environments, but the accuracy of this detection is very low on some data sets, and most of the scenes are outdoors. So we only kept 7 main indoor scenes. In the selection of the basic architecture, we trained ResNext50 [9], SE-ResNeXt, MobileNet V3 [13], MixNet [14] and EfficientNet [15] using the Place365 dataset. In order to achieve real-time scene recognition of mobile robot embedded devices such as NVIDIA JETSON TX2, we made a trade-off between accuracy and speed. In the end, we chose EfficientNet-B2 as the basic architecture showing good performance with small number of parameters.

3.2 Selection of object detector

Compared with image classification, object detection not only requires the identification of the physical category in the image, but also the output of the object's position parameters [10], [16]. We used the YOLOv5 model. The test speed of this model on NVIDIA JETSON TX2 is about 60ms per image, which meets our speed requirements with good detection accuracy. YOLOv5 has well inherited and optimized the structure of YOLOv4, using CSPDarkNet53 as the backbone, and Neck using FPN and PAN modules. The prediction and output of the results are completed in the head. We believe that every indoor scene has a specific object category, and we can assist in judging the current scene category by detecting specific objects. For specific correspondence, we refer to the paper [7]. However, among the 80 object categories in the MSCOCO dataset, there are many invalid categories that cannot be used in our scenario. Therefore we extracted the 15 object data in Table 1 as the training set, where the numbers 1-6 respectively represent the key objects in the bathroom, bedroom, dining room, kitchen,

living room, and office. If there are no objects, it belongs to the corridor. We specially trained a detector model for these 15 objects to extract object information as shown in Fig. 2.

Table 1. 15 categories of scene-specific objects

1) toilet	1) sink	2) bed	3) dining table
3) wine glass	3) bowl	4) oven	4) microwave
4) refrigerator	5) sofa	5) vase	6) TV
6) laptop	6) keyboard	6) mouse	

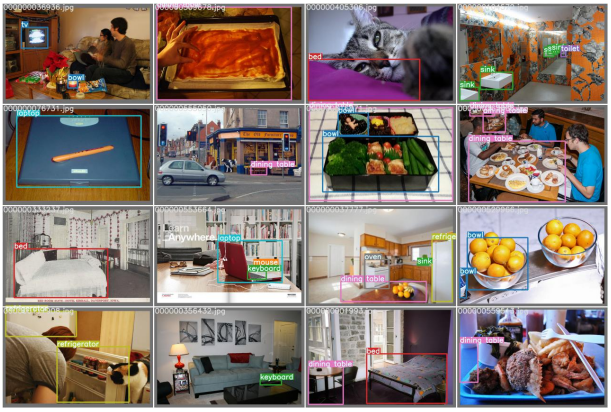

Fig. 2. Visualized images of the detector training process

3.3 Scene classifier based on feature fusion

Since the key object detection of the detector can promote the result of scene classification, we fuse the results extracted by the object detector with the features extracted by the classification network to obtain more accurate results. Extracting the features of the last layer of the classifier, the 7D logits feature vector with the dimension of the scene value is obtained. In the extraction process of the detector, the information of $< x, y, w, h >$ in the $N \times 6$ vector is removed, where N is the number of scenes whose detected confidence is greater than the set threshold, and the data in N is the index of 15 types of objects to be supervised. $< x, y, w, h >$ means to obtain the location information of the object. Therefore, we can create a 15D vector. For each type of object, no matter how many times it appears, we only take the data with the largest confidence and store it in the 15D index to create a vector containing the object's confidence. Then we fuse the two vectors to obtain a 22D vector as the new neural network input layer. We redefine the two fully connected layers, and output the 7D scene recognition result, using the softmax function and the cross entropy loss function given by

$$Loss = -\sum_{i=0}^{C-1} y_i log(p_i) = -log(p_c), \qquad (1)$$

where $p = [p_0, ..., p_{c-1}]$ corresponds to the probability distribution of the vector output network, and p_i represents the probability that the sample belongs to the i-th category. $y = [y_0, ..., y_{c-1}]$ is the one hot representation of the sample label, and C corresponds to the index of the sample.

3.4 Semantic map creation

For the creation of the semantic metric map, we plan to use some map segmentation algorithms to divide the occupied grid map into different regions firstly, and then determine the specific semantic information of the region based on the RGB images obtained in the region. There are many map segmentation algorithms such as MAORIS segmentation [12], and spectral clustering and quadtree [17]. We use the MAORIS algorithm to complete the map segmentation task in the blank area. This method first calculates the distance between image and free space image of the map. Then the method groups adjacent pixels of same value in regions, immediately remove ripples, then merge regions with similar values and removes regions created by thick walls, and finally straightens boundaries. A large number of regions can be generated by this algorithm. Specifically, it can be expressed by the following formula:

$$M = \{R_1, R_2, ...R_n\}, \qquad (2)$$

$$I_{i1}, I_{i2}, ...I_{in} \in R_i, \qquad (3)$$

where M represents the input map, $R = [R_1, ...R_n]$ corresponds to the set of divided areas, $I_i = [I_{i1}, ...I_{in}]$ corresponds to the set of RGB images obtained by the robot in the area R_i. We need to determine which area the divided region belongs to using the formula given by:

$$Class(R_t) = Index(Max \sum_{i=0}^{n} p(y_j|I_{ti})), \qquad (4)$$

where $p(y_i|I_{ti})$ corresponds to the 7D probability distribution of the output through our model, y_j is the known local scene class, and I_{ti} is the i-th image in the t-th region. We add up all the probability distributions to get the index of the maximum value in the 7D vector. The index of the maximum value is the category of the region R_t. In this way, we get the categories of all regions to generate a semantic map.

In addition, the semantic map can contain some key objects which can promote the robot's understanding of the objects and enrich the semantic information. We use YOLOv5 as the detector, and select a large number of indoor objects as the training set in the ImageNet and MSCOCO datasets. The 35 types of objects selected are shown in Table 2.

Table 2. Object categories in ImageNet and MSCOCO

chair	table	bowl	mug	lamp
display	stove	flowerpot	bed	piano
laptop	sofa	coffee maker	keyboard	wine bottle
bookcase	mouse	water bottle	washer	microwave
refrigerator	guacamole	dishwasher	milk can	blow dryer
file cabinet	soap dispenser	toaster	printer	ladle
can opener	ewer	toilet	oven	cell phone

4. EXPERIMENTS

We evaluated our proposal in multiple datasets, compared with traditional methods.

4.1 Training and evaluation on Place365 dataset

Place365 is a classic scene dataset, consisting of Places365-Standard and Places365-Challenge. The training set of Places365-Standard has about 1.8 million images from 365 scene categories, and each category has a maximum of 5,000 images. We selected 7 categories of bathroom, bedroom, corridor, kitchen, living room, dining room, and office for training and testing. We selected 5 types of networks as the basic architecture for training. In addition, in EfficientNet, we used B0, B2, and B3 networks according to the model's width, depth, and resolution scaling bases α, β, and γ, respectively. The training parameters of EfficientNet are as follows: RMSProp optimizer with decay 0.9 and momentum 0.9; batch norm momentum 0.99; weight decay 1e-5; initial learning rate 0.256 that decays by 0.97 every 2.4 epochs. We train each model for a maximum of 450 epochs, and select the model with the best performance. The python version we used is 3.7.10, the PyTorch version is 1.7.0, cuda is 10.1, opencv is 4.5.1, and torchvision is 0.8.1. The operating system is Ubuntu 20.04, and the GPU is GeForce RTX 2080 Ti. Fig. 3 shows the convergence curve of the function during the training process. Since our method has acquired prior knowledge of the classification and detection model, only two fully connected layers and activation functions are needed to train, yielding very fast convergence.

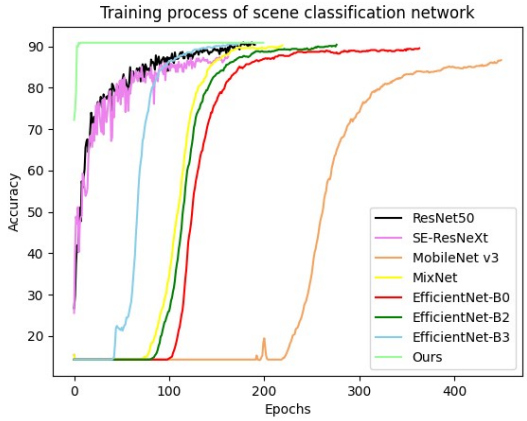

Fig. 3. Models training process and convergence curve

Table 3 shows the experimental results of each model tested on Place365. It can be seen that our method can achieve the best results when using a small amount of parameters. Note that EfficientNet is a very good network, showing very good generalization ability with fewer parameters. Therefore, our method is based on the feature fusion of EfficientNet-B2 and Yolov5s model, which has shown to be more effective than the two used alone.

4.2 Evaluation on Robot@Home dataset

Robot@Home [18] is a collection of raw and processed data from 5 domestic settings compiled by a mobile robot equipped with 4 RGB-D cameras and a 2D laser scanner. Its main purpose is to serve as a testbed for semantic mapping algorithms through the categorization of objects and/or rooms. The RGB images of 5 houses collected by the robot are used as input to various models to evaluate accuracy. Among them, the dataset contains 7,242, 8,597, 10,422, 3,886, and 4,228 RGB images from Home1 to Home5, and we use all these images as the testset. We compare the basic architecture EfficientNet-B2 with our method. Table 4 shows the experimental results.

The experimental results in 5 different houses showed that the average performance of our method is higher than that of the EfficientNet-B2 network. The results show that we add object information to the network and perform information fusion to help the classification results, which can increase by about 2% on average. When the category itself has a high accuracy rate, such as the bathroom, the performance of our method is not improved, because EfficientNet itself can fit this part of the data very well. But for the original low-accuracy categories such as bedroom and living room, EfficientNet cannot fit these test data well, but our method can improve greatly on the original basis, and even increase by 7%. This is due to the information of the object detection network we added. It makes up for the features that are difficult to extract from the classification network. For the kitchen category, our approach performs worse than EfficientNet. This may be because the characteristics of the internal object objects in the kitchen were not well learned during the training of the detector. The target domain of the test image is too far from the original domain. We can solve the problem by re-dividing the object category and improving the quality of the object dataset. We compared the image difference between Place365 and this dataset. The domain of the image set is very different, but the average accuracy of the model is higher than 60%, indicating that our method has good generalization ability and can be transferred to various scenarios.

4.3 Evaluation on VPC dataset

The Visual Place Categorization dataset (VPC Dataset) [2] contains 6 home information with various rooms. The RGB image is the frame data obtained by the camera sensor, which has a strong sequence. We selected 6 scene categories for testing. Among them, there are about 7,200 RGB images in the bathroom, more than 10,000 in the bedroom, about 1,000 in the corridor, more than 2,700 in the dining room, about 3,000 in the kitchen

Table 3. Results of classification in the Place365 dataset

Models	Accuracy(%)	Parameter(M)
ResNet50	90.571	22.99
SE-ResNeXt	87.571	14.77
MobileNet v3	86.714	4.21
MixNet	90.280	10.37
EfficientNet-B0	89.571	**4.02**
EfficientNet-B3	90.714	10.71
EfficientNet-B2	90.429	7.71
Yolov5s	74.857	7.30
Ours	**90.857**	15.02

Table 4. Results in the Robot@Home dataset

	Alma		Anto		Pare		Rx		Sarmis	
Scene	EfficientNet	Ours	EfficientNet	Ours	EfficientNet	Ours	EfficientNet	Ours	EfficientNet	Ours
Bathroom	0.9137	**0.9180**	0.9117	**0.9175**	**0.8855**	0.8820	0.9723	**0.9754**	0.9414	**0.9497**
Bedroom	0.4025	**0.4780**	0.4063	**0.4180**	0.3446	**0.3658**	0.6334	**0.6991**	0.3425	**0.3679**
Corridor	**0.8128**	0.8115	**0.5826**	0.5809	**0.5544**	0.5519	**0.5423**	0.5373	**0.624**	0.6200
Living Room	0.4004	**0.4082**	0.5612	**0.5826**	0.3039	**0.3124**	0.1539	**0.2211**	0.4777	**0.4936**
Kitchen	-	-	0.5858	**0.6423**	0.3998	**0.4123**	-	-	0.3707	**0.3922**
Avg.	0.6324	**0.6539**	0.6095	**0.6283**	0.4976	**0.5049**	0.5755	**0.6082**	0.5513	**0.5647**

and living room. Due to the large difference in the distribution of the number of images in various scenes, we test according to the scene category instead of the room. Table 5 shows the experimental results.

Table 5. Results in the VPC dataset

Networks	Bathroom	Bedroom	Corridor	Dining Room	Kitchen	Living Room	Avg.
SE-ResNext	66.33	41.36	71.24	43.71	56.92	52.80	55.39
MobileNetV3-L	63.80	40.14	71.54	46.17	57.83	55.91	55.89
MixNet	73.47	47.24	63.57	44.20	67.32	58.02	58.97
ResNet50	72.02	50.92	65.39	47.88	65.17	61.42	60.47
EfficientNet-B0	68.95	48.50	72.86	55.36	62.83	64.27	62.13
EfficientNet-B2	79.99	51.98	65.99	55.14	67.77	72.08	65.49
EfficientNet-B3	73.97	**55.22**	**73.66**	49.35	**68.32**	73.02	65.59
Ours	**80.25**	53.03	66.79	**56.47**	67.61	**73.18**	**66.22**

It can be seen that our approach has the best average performance, and it surpasses other architectures greatly in the bathroom scene. The average accuracy of our model in the VPC dataset exceeds 66%, showing that it can be generalized for various application scenarios. We removed the corridor scene and kept the variables consistent with the paper [7]. Table 6 shows experimental results in comparison with baseline methods. The first baseline [2] uses the SIFT algorithm of feature point matching, the CENTRIST (CE) descriptor, and the Bayesian filter (BF) method for scene recognition. The next method [19] implements context-based scene recognition by introducing the Histogram of Oriented Uniform Patterns (HOUP). [20] presented a system recognizing places utilizing both global configurations observation and local objects information. It combines global configuration with a Bayesian filtering framework (G+BF) and object feature matching (G+SIFT+BF). We also included classification networks based on deep learning, such as AlexNet [4] and ResNet, along with Batch Normalization (BN) to improve accuracy and speed up convergence. The paper [7] uses ResNet18 (Scene-only) as the basic network and YOLOv3 (Combined) as the detector for feature fusion to improve the accuracy of scene recognition. Finally, we tested the performance of EfficientNet and our approach. According to the network scaling, we tested the experimental results of B0, B2, B3 with BN. In this paper, the accuracy of scene recognition in each room of our method is 62.40%, 62.59%, 69.63%, 70.42%, 64.94%, 66.02%, and the final average accuracy rate is 66.00%, higher than all other approaches.

4.4 Semantic map in Robot@Home

We use our deep learning-based image feature fusion method as a classifier for scene recognition. This classifier can achieve state-of-the-art effects in indoor environment recognition. Through the MAORIS segmentation algorithm, we get the metric map. Then we take all the images in the region as input, and get a probability distribution through our classifier. We set a 50% threshold as the accuracy of classification. If the maximum probability distribution is greater than 50%, we think this scene category is the category we ultimately need. If it is less than the threshold, we consider this to be an unknown area. In addition, these images will obtain the objects in them through the object detection algorithm YOLOv5, which can greatly improve the semantic information of the map. We tested the 2D metric map in the Robot@Home dataset. Fig. 4 shows the visualization result of the semantic map generated by Pare Home. It can be seen that the semantic map not only retains the scene information, but also retains the object information. Compared with the ground truth, our semantic map is very consistent with the real scene distribution. We will further increase the number of detection categories of the network so that it can be applied to more scenarios.

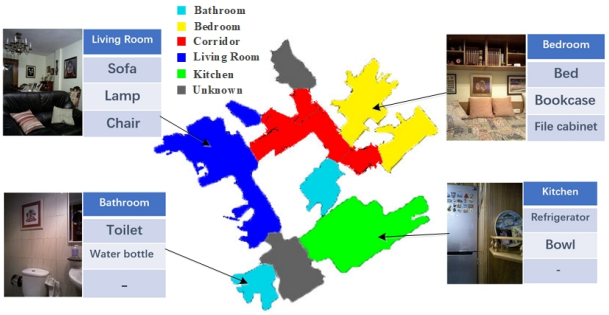

Fig. 4. Results of semantic mapping in the Pare Home

5. CONCLUSION

In this paper, we fused the features of the scene recognition algorithm and the key object detection algorithm, and conducted experiments on three real indoor datasets. The results confirmed that our approach outperformed the existing state-of-the-art approach, and showed a good generalization ability, which can transfer the model to a variety of different indoor scenes. The advantages of our model can be more reflected especially where the scene recognition does not work well. This makes the proposed semantic map closer to ground truth, laying the foundation for the robot to perform more tasks. As future work, we will add more scene information and work on topological mapping having many advantages compared to metric maps, such as small storage memory and convenient use of graph algorithms for the shortest path planning tasks. Topological maps build upon the proposed

Table 6. Comparison with Baseline in the VPC dataset

Approach	[2]				[19]	[20]		AlexNet		ResNet18	[7]		EfficientNet			Ours
Config.	SIFT	SIFT+BF	CE	CE+BF	HOUP	G+BF	G+O+BF	Base	BN	BN	Sce.	Comb.	B0	B2	B3	Comb.
Accuracy	35.0	38.6	41.9	45.6	45.9	47.9	50.0	50.2	53.7	55.0	63.7	65.7	58.6	64.2	63.7	**66.0**

semantic metric maps and RGB-D image detection and segmentation.

REFERENCES

[1] A. Rottmann, O. M. Mozos, C. Stachniss, W. Burgard. Semantic place classification of indoor environments with mobile robots using boosting. *AAAI Nat'l Conf. on Artificial Intelligence*, 5:1306–1311, 2005.

[2] J. Wu, H. I. Christensen, J. M. Rehg. Visual place categorization: Problem, dataset, and algorithm. *IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, 4763–4770, 2009.

[3] N. Sünderhauf, F. Dayoub, S. McMahon, B. Talbot, R. Schulz, P. Corke, G. Wyeth, B. Upcroft, M. Milford. Place categorization and semantic mapping on a mobile robot. *IEEE Int'l Conf. on Robotics and Automation*, 5729–5736, 2016.

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

[5] M. Brucker, M. Durner, R. Ambruş, Z. C. Márton, A. Wendt, P. Jensfelt, K. O. Arras, R. Triebel. Semantic labeling of indoor environments from 3d rgb maps. *IEEE Int'l Conf. on Robotics and Automation*, 1871–1878, 2018.

[6] J. C. Rangel, M. Cazorla, I. García-Varea, C. Romero-González, J. Martínez-Gómez. Automatic semantic maps generation from lexical annotations. *Autonomous Robots*, 43(3):697–712, 2019.

[7] A. Pal, C. Nieto-Granda, H. I. Christensen. Deduce: Diverse scene detection methods in unseen challenging environments. *arXiv preprint arXiv:1908.00191*, 2019.

[8] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

[9] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. *IEEE Conf. on Computer Vision and Pattern Recognition*, 770–778, 2016.

[10] J. Redmon, A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick. Microsoft coco: Common objects in context. *European Conf. on Computer Vision*, 740–755, 2014.

[12] M. Mielle, M. Magnusson, A. J. Lilienthal. A method to segment maps from different modalities using free space layout maoris: map of ripples segmentation. *IEEE Int'l Conf. on Robotics and Automation*, 4993–4999, 2018.

[13] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.* Searching for mobilenetv3. *IEEE/CVF Int'l Conf. on Computer Vision*, 1314–1324, 2019.

[14] Mingxing Tan and Quoc V Le. Mixconv: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595*, 2019.

[15] M. Tan, Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *Int'l Conf. on Machine Learning*, 6105–6114, 2019.

[16] M. Tan, R. Pang, Q. V. Le. Efficientdet: Scalable and efficient object detection. *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 10781–10790, 2020.

[17] Y. Tian, K. Wang, R. Li, L. Zhao. A fast incremental map segmentation algorithm based on spectral clustering and quadtree. *Advances in Mechanical Engineering*, 10(2):1687814018761296, 2018.

[18] J. R. Ruiz-Sarmiento, C. Galindo, J. González-Jiménez. Robot@ home, a robotic dataset for semantic mapping of home environments. *Int'l Journal of Robotics Research*, 36(2):131–141, 2017.

[19] E. Fazl-Ersi, J. K. Tsotsos. Histogram of oriented uniform patterns for robust place recognition and categorization. *Int'l Journal of Robotics Research*, 31(4):468–483, 2012.

[20] H. Yang, J. Wu. Object templates for visual place categorization. *Asian Conf. on Computer Vision*, 470–483, 2012.