Abnormal Occupancy Grid Map Recognition using Attention Network

Fuqin Deng\textsuperscript{1,3,4*}, Hua Feng\textsuperscript{1*}, Mingjian Liang\textsuperscript{1}, Qi Feng\textsuperscript{1}, Ningbo Yi\textsuperscript{1}, Yong Yang\textsuperscript{4}, Yuan Gao\textsuperscript{3}, Junfeng Chen\textsuperscript{3}, and Tin Lun Lam\textsuperscript{2,3,†}

Abstract—The occupancy grid map is a critical component of autonomous positioning and navigation in the mobile robotic system, as many other systems’ performance depends heavily on it. To guarantee the quality of the occupancy grid maps, researchers previously had to perform tedious manual recognition for a long time. This work focuses on automatic abnormal occupancy grid map recognition using the residual neural networks and a novel attention mechanism module. We propose an effective channel and spatial Residual SE(csRSE) attention module, which contains a residual block for producing hierarchical features, followed by both channel SE (cSE) block and spatial SE (sSE) block for the sufficient information extraction along the channel and spatial pathways. To further summarize the occupancy grid map characteristics and experiment with our csRSE attention modules, we constructed a dataset called occupancy grid map dataset (OGMD) for our experiments. On this OGMD test dataset, we tested few variants of our proposed structure and compared them with other attention mechanisms. Our experimental results show that the proposed attention network can infer the abnormal map with state-of-the-art (SOTA) accuracy of 96.23% for abnormal occupancy grid map recognition.

I. INTRODUCTION

The occupancy grid map was first introduced for surface point positions with two-dimensional (2D) planar grids [1], which had gained great success fusing raw sensor data in one environment representation [2]. In the narrow indoor environments or spacious outdoor environments, occupancy grid map can be used for the autonomous positioning and navigation by collecting the position information of obstacles. In recent year, occupancy grid map has applications in obstacle avoidance [3], multi-sensor data fusion [4], object tracking [5], simultaneous localization and mapping (SLAM) [6], and multi-robot global localization [7].

This work was supported in part by the National Key R&D Program of China (2020YFB13113000), the funding (AC01202101025, AC01202101026) from the Shenzhen Institute of Artificial Intelligence and Robotics for Society, the special projects in key fields of Guangdong Provincial Department of Education (2019KZDZX1025), Guangdong Science and Technology Major Special Fund (No.2019-252), Innovative Program for Graduate Education (50317006259) from the Wuyi University, and Shenzhen Peacock Plan of Shenzhen Science and Technology Program (KQTD2016113010470345). We would also like to thank Dr. Li Nan Nan from Macau University of Science and Technology for his valuable discussion on this paper.

*Authors contributed equally
1School of Intelligent Manufacturing, the Wuyi University, Jiangmen, China.
2School of Science and Engineering, the Chinese University of Hong Kong, Shenzhen, China.
3The Shenzhen Institute of Artificial Intelligence and Robotics for Society, the Chinese University of Hong Kong, Shenzhen, China.
4The 3irobotix Co.,Ltd, Shenzhen, China.
†Corresponding author is Tin Lun Lam tllam@cuhk.edu.cn

On the path to autonomous navigation, an occupancy grid map is essential for trajectory planning module [8]. In the SLAM process, the occupancy grid maps aim at estimating the occupancy probability and recording discretized object’s locations [9]. In practice, many factors may affect the quality of the obtained occupancy grid maps. As shown in the 1st row of Fig. 1 (a), when the mobile robot meets the transparent glass walls, the laser sensor cannot receive the laser light penetrating through the transparent glass, making it hard to determine the specific location of walls. Moreover, if the mobile robots without a loop detection module [10] builds an occupancy grid map on a long corridor in the 2nd row of Fig. 1 (a), they may lose some critical scan measurement data. Due to these overlapped or incomplete occupancy grid maps (e. g. Fig. 1(b)), it would make mobile robots hard to plan the path and avoid obstacles during navigation.

For abnormal occupancy grid map recognition, there are mainly three types of methods, including shape estimation [11], [12], edge detection [13] and data-driven methods [14]. The shape-based and edge-based methods are designed to associate measurements to object’s feature, while some abnormal occupancy grid maps are falsely detected as positive. For data-driven method, especially with the convolutional neural networks (CNN) [15], the abnormal occupancy grid map can be detected in terms of low-level features (e. g. blur, burr, and overlap). The recognition experiments with mainstream CNN-based method can get the approximately 91% accuracy result (as shown in Tab. 1).
The research on automatic map construction using robots has always been a key point for researchers [19], [20]. Elfes et al. [21] introduced a grid-based algorithm for 2D environment modeling. The continuous spaces of the environment are discretized by these evenly-spaced grids. According to the probabilistic formulation, each grid cell of constructed map represents the probability of the corresponding region that may be occupied, free or even not yet explored. With the help of occupancy grid map, the robot can identify the presence or absence of an obstacle in the space of the environment.

However, the occupancy grid map suffers from corrupted obstacle silhouettes, occlusions, and false distance estimates in static environments. To address these problems, a highly engineered method [11] was proposed to extract information for the environment modeling. A data-driven method for map recognition is proposed in [13], where the radar sensor measurement model was trained and this method outperforms the manually designed model. Piewak et al. [14] trained a neural network to reduce false distance estimation in a dynamic environment. Their approach referred to a pixel-wise classification task to determine whether a cell in the actual environment is occupied or free.

B. Deep Architectures and Attention Mechanisms

In the last years, many attempts have been made to improve the original CNN architecture to achieve better accuracy. In image classification and object detection, recent approaches like AlexNet [22], VGG-16 [23], InceptionNet [24], or MobileNet [25] are based on the plain CNN architecture. However, CNN-based networks have problems in gradient propagation and convergence with the amount of data increases [26]. To address this issue, ResNet [27] proposes a simple identity skip-connection to ease the optimization problem of deep networks. Our attention network is a backbone architecture based on ResNet, which can enhance the input features but also well benefit gradient propagation on occupancy grid map data training process.

Recently, various attention mechanisms had been proposed for image recognition tasks. Hu et al. [17] had proposed a compact SE block to extract the inter-channel information. Residual Attention Network (RAN) [28] modified ResNet by stacking identical soft attention modules which is beneficial to refine the feature maps. Spatial attention [29] had been made to recalibrate the channel dependency as an effective extraction module. Then, Woo et al. [18] introduced a CBAM module that sequentially recalibrates channel and spatial attention to refine intermediate feature maps. However, these methods miss the global spatial information, which is an important factor for feature fusion and accurate attention map generation. Inspired by the global context block [30], our csRSE attention network, which contains a residual block for producing hierarchical features, followed by both cSE block and sSE block for the sufficient spatial information extraction and the resulted hierarchical features enhancement.

III. A NEW SLAM DATASET

To facilitate the research of the occupancy grid map recognition problem, we contribute a new SLAM dataset containing 6916 occupancy grid maps through an indoor robot vacuum cleaner. To the best of our knowledge, OGMD is a benchmark specifically for the occupancy grid map recognition. We contribute the source code and our dataset public available to simulate future research.

A. Dataset Construction

The constructed OGMD covers diverse daily-life indoor environment scenarios, such as residential houses, supermarkets, office buildings, hotels, and schools. These occupancy grid maps are created with an initial size of 50m x 50m. To further increase the number of training examples, we applied random rotation and offset to cropped areas of 34m x 34m used as training examples. These occupancy grid maps are labeled as two categories, including 3210 normal

1The available online website: https://github.com/ThomerShen/OGMD
occupancy grid maps and 3706 abnormal occupancy grid maps respectively. In the following experiments, 6916 occupancy grid maps are randomly divided into 4150 training examples, 2079 validation examples, and 690 test examples, and the ratio is roughly 6:3:1.

B. Dataset Analysis

In OGMD, the occupancy grid maps are generated by the scan data of the robot laser sensor. For detail, each cell of occupancy grid map is obtained by the scan measurement data. In real collection scene, the occupancy grid maps are created by using either one scan or an accumulation of multiple sensor scans. The occupancy grid maps created by mobile robot are depicted in Fig. 2.

Through the scans of the laser sensor, all obstacles are displayed as a line segment in the occupancy grid map. The aggregation of the scans represents the feature parameters of obstacles in the indoor environment. As shown in Fig. 2 (a) and (b), the bright white region of the occupancy grid map indicates a flat and open space. In contrast, the dark line segment represents the obstacles in the indoor environment.

C. Dataset Evaluation

Each grid cell of occupancy grid map contains the distributed probability of obstacles and free-space in the actual indoor environment. Fig. 2 (a)–(d) show the created occupancy grid maps in different indoor scenarios. The criteria for evaluating the abnormal occupancy grid maps are as follows:

- The region in occupancy grid map is inconsistent with the actual environment, such as the repeated mapping and the overlap region, like Fig. 2 (c).
- The edges of obstacles on the map are unclear, like Fig. 2 (d).
- The meaningful information is missing in the scanning region of mobile robot, like Fig. 2 (c) and (d).

IV. METHOD

The detailed structure of our attention network is illustrated in Fig. 3. In the figure, for the input feature map \( I \in \mathbb{R}^{C \times H \times W} \), our attention network computes the channel and spatial attention map \( O \in \mathbb{R}^{C \times H \times W} \). Assuming the input and output dimensions are the same, through channel and spatial Residual SE (csRSE) module, the refined feature map transformation \( I_3 \rightarrow O \) can be defined as

\[
O = I_1 + I_1 \otimes I_3,
\]

where \( \otimes \) denotes element-wise multiplication. Through the residual block calculation and the attention mechanism recalibration, the network can extract global informative features.

In our module, after the residual block calculation, we get the feature \( I_1 \), then we compute the channel attention \( F_c \in \mathbb{R}^{C \times 1 \times 1} \) and the spatial attention \( F_s \in \mathbb{R}^{1 \times H \times W} \) through the cSE block and the sSE block, so the attention map is computed as

\[
I_2 = I_1 \otimes F_c(I_1),
\]

\[
I_3 = I_2 \otimes F_s(I_2),
\]

where the \( I_2 \) is output feature map of channel attention. The \( I_3 \) is the final refined output feature map. The following subsection will describe the details of cSE Block and sSE block.

A. Channel SE Block

In the cSE block, we recalibrate the inter-channel relationship for the feature response, which involves two steps, spatial squeezation and channel excitation. In the first step, for the feature map \( I_1 \in \mathbb{R}^{C \times H \times W} \), we use the Global Average Pooling (GAP) to squeeze the global information. Then a unique channel vector \( V_c \in \mathbb{R}^{C \times 1 \times 1} \) of each channel is produced by the mean of GAP.

In the second step, to estimate attention from the \( c \)-th element of statistic channel vector, we use the Multi-layer Perceptron (MLP) which contains two Fully Connected (FC) layers, the Rectified Linear Units (ReLU) function, and the sigmoid function to capture channel-wise dependencies. The purpose of this MLP is to emphasize the channels with the meaningful information. In short, the output of channel attention is computed as:

\[
F_c(I_1) = MLP(GAP(I_1)) = \sigma(W_2\delta(W_1GAP(I_1))),
\]

where \( W_1 \in \mathbb{R}^{C \times C}, W_2 \in \mathbb{R}^{C \times C} \) are the weights of the FC layers, \( C = \frac{C}{r}, r \) is the reduction ratio (\( r \) is set to 16), \( \delta \) refers to the Rectified Linear Units (ReLU) function, \( \sigma \) refers to the sigmoid function.

B. Spatial SE Block

In the sSE block, we use the convolutional layers to implement channel squeeze and spatial excitation. Here, it is assumed an alternative representation of the input tensor as \( I_2 \). Four standard convolution layers \( W \) are concatenated...
to produce the spatial attention map $F_s(I_2) = W * I_2$. The sigmoid function $\sigma(F_s(I_2))$ determines the importance of the specific location across the feature map. Like the previous block, this recalibration process indicates which locations are more meaningful during the training procedure. As a result, the output of spatial attention block can be expressed as:

$$I_3 = \sigma(F_s(I_2)) = \sigma(W * I_2) = \sigma(f_1^{1x1} (f_3^{3x3} (f_2^{1x1} (I_2)))),$$

where $\sigma$ refers to the sigmoid function, * denotes the convolution operation. The $f_1^{1x1}$ and $f_3^{3x3}$ represents a convolution operation with the filter size of $1 \times 1$ and $3 \times 3$, respectively.

**C. Network for Grid Map Recognition**

For the occupancy grid map recognition task, the goal is to improve representation ability of network architecture by using attention mechanism, which can focus on meaningful features and suppress unnecessary ones. As illustrated in Fig. 3, csRSE module, the residual block is designed for aggregating hierarchical features at multiple levels and accelerating convergence. To contain sufficient spatial information of the intermediate layers, the sSE block are stacked after the cSE block for spatial feature extracting. The sSE block extracts the hierarchical features and improves the representation of interests via four convolution operation. All in all, cSE block and sSE block can be regarded as the extracting core, which can aggregate specific global context features of grid maps. These global features are weighted averaged from all areas via a specific attention map to each specific map areas. As attention maps are computed for each areas, the detail heat map of the OGMD dataset is displayed for the specific areas (as shown in Fig. 5).

Secondly, to extract multi-scale features, the convolution layers and csRSE modules are employed as the core of feature extraction in the proposed network architecture. Through the convolution operations, it can extract multi-scale features by different kernel size of convolution layer. After the convolution layer, we adopt a csRSE module to emphasize meaningful grid map features along two principal dimensions (channel and spatial axes). To achieve the abnormality in occupancy grid maps, we sequentially apply cSE block and sSE block, so that the csRSE modules can learn ‘what’ and ‘where’ to attend in the channel and spatial axes, respectively. As a result, the convolution layers and csRSE modules efficiently helps the information flow within the network by learning which map feature information to emphasize or suppress.

As illustrated in Fig. 3, there are four convolutional layers, four csRSE modules, one GAP, and one FC layer in the proposed csRSE-integrated network (ResNet32 + csRSE) architecture. At the beginning of proposed network, the input 3-channel occupancy grid map image is transformed into a 64-channel feature map. The ResNet in the proposed csRSE-integrated network is similar to [27] but use dilated convolutions [31]. The stacked block is designed as one consecutive convolution layer, one residual block, one cSE block, and one sSE block. The outputs of two attention blocks are fused to generate high-level contextual features, which will be used to guide the low-level contextual features extracted by the residual block. The residual block and two attention blocks are stacked for blending cross-channel information and spatial hierarchical features together. After each stacked block, the max-pooling is performed and each the filter size gets doubled, which is designed to double the spatial size of feature maps and halve channels. The bottom block is stacked with GAP, FC and softmax layer to compute probabilities of predicted category.

**V. EXPERIMENTS AND RESULTS**

To evaluate the proposed method, we carry out experiments on our OGMD dataset for occupancy grid map recognition. Experimental results demonstrate that the proposed attention network generally outperforms the networks with the SE block and CBAM module, respectively.

**A. Experimental Setups**

We implement occupancy grid map recognition experiments on our OGMD dataset using the PyTorch framework
TABLE I
COMPARISONS WITH METHODS THAT HAVE THREE DIFFERENT
ATTENTION MODULES ON OUR OGMD
TEST DATASET.

| Methods          | Param. (M) | GFLOPs | Accuracy (%) |
|------------------|------------|--------|--------------|
| ResNet20         | 0.25       | 2.019  | 91.27        |
| ResNet20 + SE    | 0.27       | 2.019  | 93.62        |
| ResNet20 + CBAM  | 0.27       | 2.021  | 93.86        |
| ResNet20 + csRSE(ours) | 0.27     | 2.03   | 94.83        |
| ResNet32         | 0.45       | 3.418  | 91.43        |
| ResNet32 + SE    | 0.47       | 3.418  | 94.15        |
| ResNet32 + CBAM  | 0.47       | 3.422  | 94.21        |
| ResNet32 + csRSE(ours) | 0.47   | 3.431  | 96.23        |

[32]. For training, the image resolution is converted to 224 x 224 and normalized with zero-mean normalization [33]. The entire network is trained end-to-end for 400 epochs by Stochastic Gradient Descent (SGD) with the momentum of 0.9 and a weight decay of 1e-4. The mini-batch size is set to 16. It takes about 8 hours for network to converge on an NVIDIA GTX 2080Ti GPU. We initialize the weights according to the method in [34] and use binary cross-entropy (BCE) loss function. The formulation of BCE loss is as follows:

\[
\text{Loss} = - (y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)),
\]

where \(y_i\) and \(\hat{y}_i\) denote the actual label category and the probability of prediction, respectively.

B. Experimental Results

We conduct experiments to validate the effectiveness of csRSE module by comparing with two popular attention modules (SE, CBAM). In experiments, we evaluate these methods on strong backbones, by replacing two different baseline network (ResNet20, ResNet32) and adding three attention modules (SE, CBAM, and csRSE). From Tab. I, all baseline network methods have achieved measurable competitive results through few calculation parameters. The networks with attention modules get the better accuracy than those without any attention methods.

Compared with the csRSE-integrated network (ResNet32 + csRSE) and CBAM-integrated network (ResNet32 + CBAM), we can find that the networks with two attention modules get the better performance than the methods with only one channel-wise attention block (ResNet32 + SE). Through the result of csRSE-integrated networks (ResNet32 + csRSE), it is seen that the attention network shows a clear advantage against the other as it gets the better accuracy (96.23%) on the test dataset. By comparison results on SE block and CBAM, the proposed csRSE attention modules capture the global context information as well as preserve more structural information.

Fig. 4 (a) shows the training and validation loss curve of the csRSE-integrated network. We can see that both the training loss and the validation loss are high at the beginning epochs. As the training epoch increases, both the losses remain quite stable. Fig. 4 (b) shows the curve of the training accuracy and validation accuracy based on the OGMD dataset. At the beginning, both the training accuracy and validation accuracy are getting stable as the learning rate gradually shrinks. Finally, we can get the best accuracy as the losses decrease.

Fig. 4 (c) shows the confusion matrix for our csRSE-integrated network on OGMD test dataset. The y-axis is the true labels and the x-axis is the predicted labels. Overall, the model has achieved a good prediction accuracy of 96.23%. Fig. 4 (d) shows the precision-recall curve of our OGMD test dataset. The y-axis shows the precision value and the x-axis shows the recall value. The experiment results show the actual prediction situation in the occupancy grid map recognition.

D. Experimental Visualization

To intuitively verify the effectiveness of the csRSE module, we visualize the attention maps (heat maps) by the Grad-CAM [35] on OGMD test dataset. In recent year, Grad-CAM is a visualization method which shows the attended regions of spatial locations in convolutional layers. The Grad-CAM results show attended regions of image clearly. By observing the attended regions, the networks can extract the important feature of image for predicting a class.

In Fig. 5, we randomly test six abnormal images with different attention networks, including the baseline network (ResNet32), SE-integrated network (ResNet32 + SE), CBAM-integrated network (ResNet32 + CBAM) and our
Fig. 5. Visualization results of attention maps (heat maps) on the occupancy grid map recognition. We compare the visualization results of baseline (ResNet32), SE-integrated network (ResNet32 + SE), CBAM-integrated network (ResNet32 + CBAM) and csRSE-integrated network (ResNet32 + csRSE). The regions that marked in the red frame represent the abnormal regions. P denotes the softmax score of each network for the ground-truth class. The higher P means the better classification.

csRSE-integrated network. It can clearly see that the GradCAM masks cover attended regions that marked in the red frame represent the abnormal regions. It can be seen that our method is capable of accurately detecting both duplicate maps (e.g., the abnormal1, abnormal2, and abnormal6 input images) and overlapping maps (e.g., the abnormal3, abnormal4, and abnormal5 input images). This is mainly because the global contextual features extracted by the csRSE module can help the network better locate abnormal regions and detect abnormal occupancy grid map. The P denotes final softmax score of networks for the ground-truth class. The higher P has the better classification result, meaning that how this network is making good use of features.

As can be observed, our method can successfully eliminate such occupancy grid maps and detect the abnormal regions. This is mainly contributed by the proposed csRSE attention network, which exploits contextual information in target abnormal regions and provides more texture details for abnormal region localization. Based on the visualization, we conjecture that the proposed csRSE attention network can more effectively detect the abnormal areas from occupancy grid maps and store more global texture details for the occupancy grid map recognition tasks.

VI. CONCLUSION

In this paper we present a residual neural network using attention mechanism for the abnormal occupancy grid map recognition task. We contribute an OGMD dataset that covers various occupancy grid maps. Then we propose a csRSE attention module, which contains a residual block for producing hierarchical features, followed by both cSE block and sSE block for the sufficient global information from channel-wise and spatial axes. Our attention network is constructed via applying multiple convolution layers and csRSE modules to multi-scale features extraction, which achieves a SOTA prediction accuracy of 96.23% for the occupancy grid map recognition. In future works, we are going to detect and predict the moving obstacles in the outdoor environment. Moreover, we will apply this attention network to study the problem of global localization for vision based multi-robot SLAM.
