Abnormal Occupancy Grid Map Recognition using Attention Network

Fuqin Deng[1,3,4*], Hua Feng[1*], Mingjian Liang[1], Qi Feng[1], Ningbo Yi[1],
Yong Yang[4], Yuan Gao[3], Junfeng Chen[3], and Tin Lun Lam[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 network with multiple novel attention mechanism modules. We propose an effective channel and spatial Residual Squeeze-and-Excitation (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 experiments 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 in fusing raw sensor data into one environment representation [2]. In narrow indoor environments or spacious outdoor environments, an occupancy grid map can be used for the autonomous positioning and navigation by collecting the position information of obstacles. In recent years, 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].

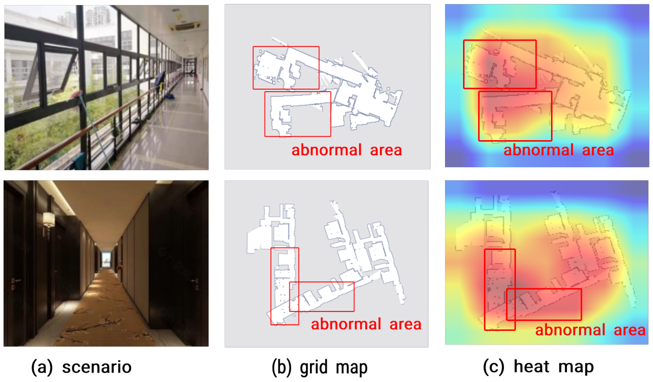


Fig. 1. Problems with occupancy grid map in existing computer vision tasks. Column (a) represent domain-specific scenarios, (b) show the corresponding abnormal occupancy grid maps, and (c) represent attention heat maps. In column (b), the regions marked in the red frame represent the abnormal areas. In column (c), our attention network method can detect these abnormal areas and then report these failure cases. More examples are in Fig. 5

On the path to autonomous navigation, an occupancy grid map is essential for the 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] build 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 with 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. I,

This work was supported in part by the National Key R&D Program of China (2020YFB1313300), 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 (503170060259) 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

[1] School of Intelligent Manufacturing, the Wuyi University, Jiangmen, China.

[2] School of Science and Engineering, the Chinese University of Hong Kong, Shenzhen, China.

[3] The Shenzhen Institute of Artificial Intelligence and Robotics for Society, the Chinese University of Hong Kong, Shenzhen, China.

[4] The 3irobotix Co.,Ltd, Shenzhen, China.

[†] Corresponding author is Tin Lun Lam `tllam@cuhk.edu.cn`

e. g. ResNet32). Furthermore, based on attention mechanism [16], Squeeze-and-Excitation (SE) block [17] and convolutional block attention module (CBAM) [18] have been embedded in the input feature map for adaptive feature refinement from channel-wise and spatial axes. Specially, these attention networks (with SE and CBAM) can achieve an accuracy result of approximately 94% (as shown in Tab. I, e. g. ResNet32 + SE). However, these methods are inefficient in terms of global feature and multi-scale feature extraction.

To improve the representation ability of feature extraction, an effective network with the attention module is designed to aggregate specific features of occupancy grid map. Different from the SE and CBAM, we design a channel and spatial Residual Squeeze-and-Excitation (csRSE) attention module for the global context extraction on the different aggregation strategies. The proposed csRSE module is a global context modeling module which aggregates the features by using residual block and hybrid attention mechanisms. In csRSE module, the residual block can generate the hierarchical features and benefit gradient propagation. Besides, the cSE block and sSE block automatically contain sufficient channel-wise and spatial information of the intermediate layers, then assign different attention weights to the abnormal regions. In our attention network, csRSE attention modules are embedded to generally focus on the multi-scale features and suppress unnecessary information. As in Fig. 1 (c), the visualizing results of our csRSE module show the important feature in occupancy grid maps.

Main contributions of this work are listed as follows:

- We introduce a csRSE attention module for global contextual extraction.
- We contribute an occupancy grid map dataset (OGMD) for SLAM occupancy grid map recognition task.
- We propose an attention network with multiple csRSE modules for SOTA abnormal occupancy grid map recognition.

II. RELATED WORK

A. Occupancy Grid Map

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 the 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 was proposed in [13], where the radar sensor measurement model was trained and this method outperformed 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 [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 the 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 channel-wise and spatial information extraction and the 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 to simulate future research and our dataset is available online[1] .

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×50m. To further increase the number of training examples, we applied random rotation and offset to cropped areas of 34m×34m used as training examples. These occupancy grid

[1] The available online website: https://github.com/ThomerShen/OGMD

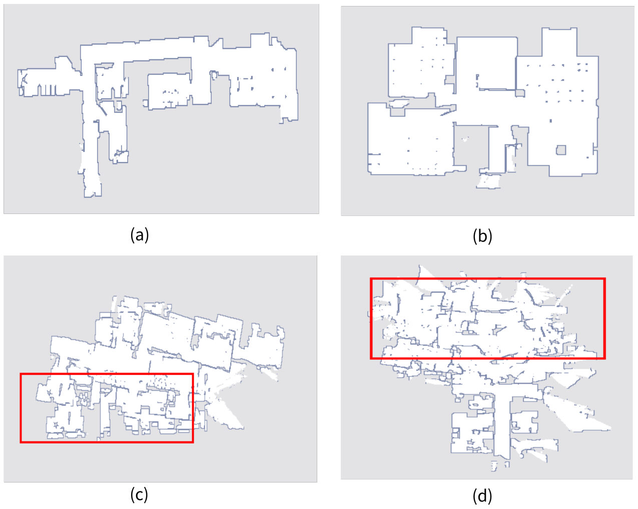


Fig. 2. Example 2D occupancy grid maps in our OGMD. (a) and (b) represent normal occupancy grid maps, (c) and (d) represent abnormal occupancy grid maps. The regions marked in the red frame represent the abnormalities.

maps are labeled as two categories, including 3210 normal 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 a real indoor 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 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 overlap region of grid map, 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, after the residual block calculation, we get the feature I_1, then the refined feature map transformation $I_1 \rightarrow O$ can be defined as

$$O = I_1 + I_1 \otimes I_3, \tag{1}$$

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 sSE block, so the attention map is computed as

$$I_2 = I_1 \otimes F_c(I_1), \tag{2}$$

$$I_3 = I_2 \otimes F_s\left(I_2\right), \tag{3}$$

where the I_2 is the 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:

$$\begin{aligned} F_c(I_1) &= MLP(GAP(I_1)) \\ &= \sigma\left(W_2\delta\left(W_1 GAP(I_1)\right)\right), \end{aligned} \tag{4}$$

where $W_1 \in \mathbb{R}^{\bar{C}\times C}, W_2 \in \mathbb{R}^{C\times\bar{C}}$ are the weights of the FC layers, $\bar{C} = \frac{C}{r}, r$ is the reduction ratio (r is set to 16). δ refers to the ReLU function, σ 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

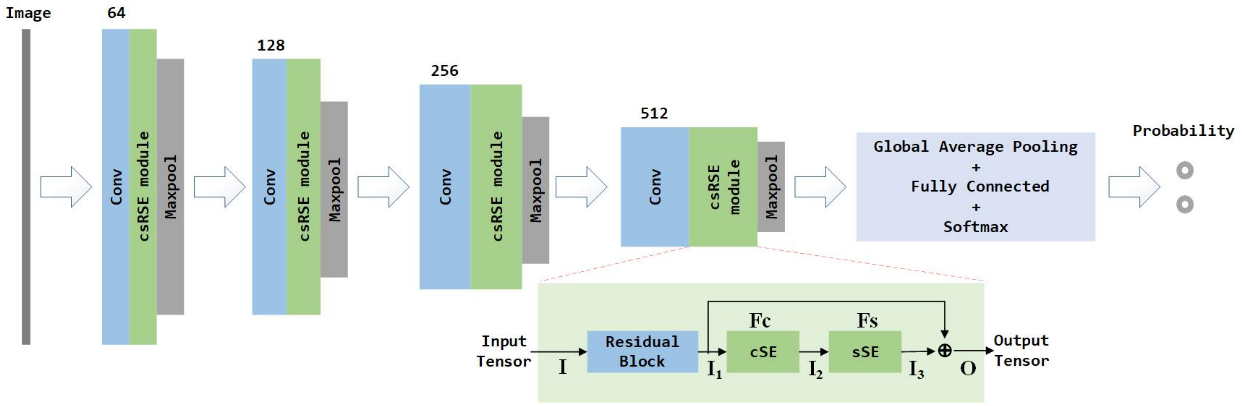


Fig. 3. The csRSE-integrated network (ResNet32 + csRSE) architecture for occupancy grid map recognition. The Conv, and csRSE module represent the convolution layer, the channel and spatial Residual Squeeze-and-Excitation (csRSE) module. Given the input feature map $I \in \mathbb{R}^{C\times H\times W}$, through the residual block calculation, we get the feature I_1. Then through the cSE block and sSE block, it can compute corresponding channel attention map I_2 and final attention map I_3. After the element-wise multiplication, we get the refined feature map O.

sigmoid function $\sigma\left(F_s(I_2)\right)$ 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

$$\begin{aligned} I_3 &= \sigma\left(F_s(I_2)\right) \\ &= \sigma\left(W * I_2\right) \\ &= \sigma\left(f_4^{1\times1}\left(f_3^{3\times3}\left(f_2^{3\times3}\left(f_1^{1\times1}\left(I_2\right)\right)\right)\right)\right), \end{aligned} \tag{5}$$

where σ refers to the sigmoid function, $*$ denotes the convolution operation. The $f^{1\times1}$ and $f^{3\times3}$ represents a convolution operation with the filter size of 1×1 and 3×3, 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 the 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 operations. 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 area, the detailed heat maps of the OGMD dataset are 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-wise 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-wise and spatial axes, respectively. As a result, the convolution layers and csRSE modules efficiently help 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 the 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 uses dilated convolutions [31]. The stacked block is defined 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 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 layers to compute probabilities of the 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 × 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 the 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:

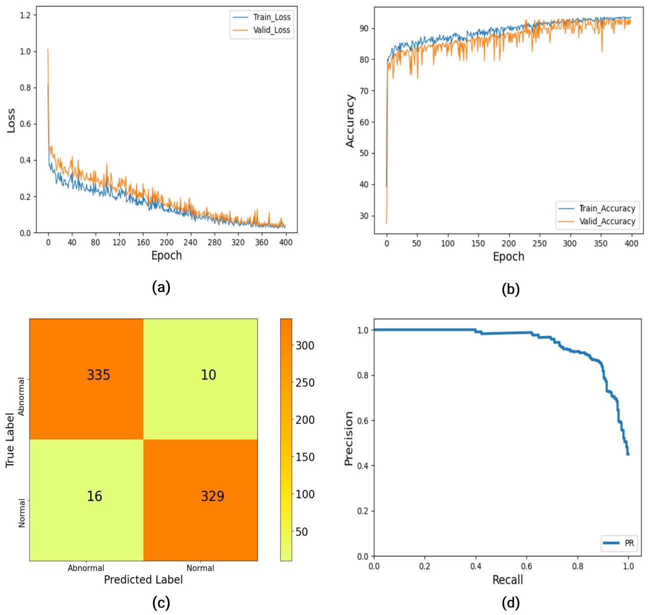
$$Loss = -(y_i \cdot \log(\hat{y_i}) + (1 - y_i) \cdot \log(1 - \hat{y_i})), \quad (6)$$

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 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 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.



(a) (b) (c) (d)

Fig. 4. (a) Loss curve, (b) Accuracy curve, (c) Confusion matrix, (d) Precision-Recall curve of experiment results on the test dataset.

C. Experimental Analysis

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. In 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 the OGMD test dataset. The y-axis is the true label and the x-axis is the predicted label. 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 years, Grad-CAM is a visualization method that utilizes gradients to calculate the spatial locations of convolutional layers. Since the gradients are computed for a unique class, Grad-CAM clearly show the image regions that have an impact on the final prediction.

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

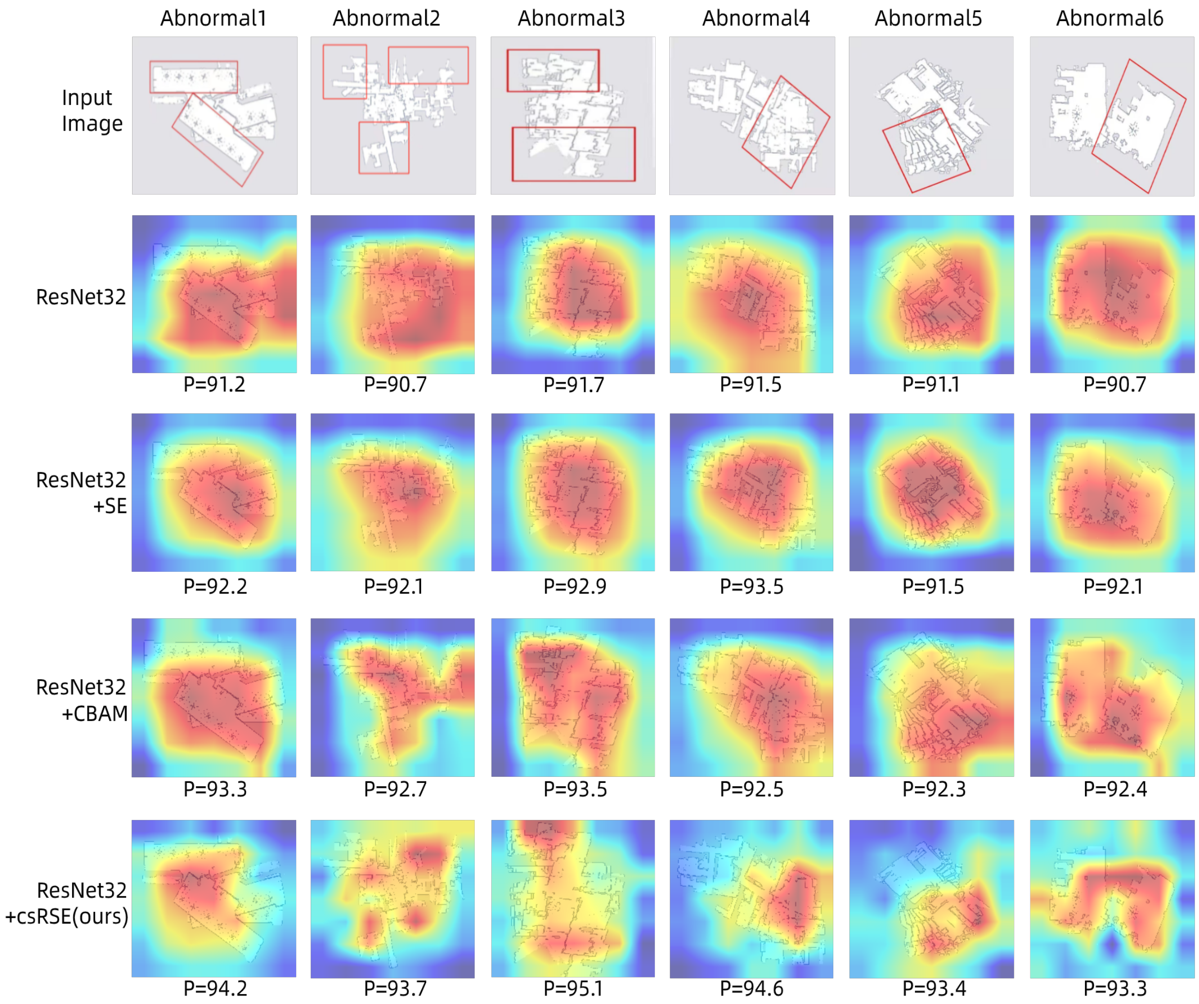


Fig. 5. Visualization results of attention maps (heat maps) on the occupancy grid map recognition. We compare the visualization results of baseline (ResNet32), SE-integrated network (ResNet32 + SE), CBAM-integrated network (ResNet32 + CBAM) and csRSE-integrated network (ResNet32 + csRSE). The regions 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 be clearly seen that the Grad-CAM masks cover attended regions that marked in the red frame represent the abnormal regions. And 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 maps. The P denotes the 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 and detect the abnormal regions of occupancy grid maps. 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.

References

[1] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, 1989.

[2] O. Hachour, "Path planning of autonomous mobile robot," *International journal of systems applications, engineering & development*, vol. 2, no. 4, pp. 178–190, 2008.

[3] J. Borenstein and Y. Koren, "The vector field histogram-fast obstacle avoidance for mobile robots," *IEEE Transactions on Robotics and Automation*, vol. 7, no. 3, pp. 278–288, 1991.

[4] S. T. A. W. DieterFox and T. D. T. M. TimoSchmidt, "Map learning and high-speed navigation in rhino," 1998.

[5] C.-C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte, "Simultaneous localization, mapping and moving object tracking," *The International Journal of Robotics Research*, vol. 26, no. 9, pp. 889–916, 2007.

[6] W. Hess, D. Kohler, H. Rapp, and D. Andor, "Real-time loop closure in 2d lidar slam," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1271–1278.

[7] X. Guo, J. Hu, J. Chen, F. Deng, and T. L. Lam, "Semantic histogram based graph matching for real-time multi-robot global localization in large scale environment," *in 2021 IEEE Robotics and Automation Letters (RAL)*, 2021.

[8] T. Ikeda, M. Kawamoto, A. Sashima, and J. Tsuji, "An indoor autonomous positioning and navigation service to support activities in large-scale commercial facilities," in *International Conference on Serviceology*, 2013, pp. 191–201.

[9] P. Sodhi, B.-J. Ho, and M. Kaess, "Online and consistent occupancy grid mapping for planning in unknown environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 7879–7886.

[10] S. Tzeng, "Loop detection for a network device," *Google Patents*, Dec. 21 2006.

[11] G. Tanzmeister and D. Wollherr, "Evidential grid-based tracking and mapping," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1454–1467, 2017.

[12] F. Roos, D. Kellner, J. Dickmann, and C. Waldschmidt, "Reliable orientation estimation of vehicles in high-resolution radar images," *in 2016 IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 9, pp. 2986–2993, 2016.

[13] A. Scheel and K. Dietmayer, "Tracking multiple vehicles using a variational radar model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3721–3736, 2019.

[14] F. Piewak, T. Rehfeld, M. Weber, and J. M. Zöllner, "Fully convolutional neural networks for dynamic object detection in grid maps," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 392–398.

[15] M. A. Nielsen, "Neural networks and deep learning," *Determination press San Francisco, CA*, vol. 25, 2015.

[16] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *Journal of Neuroscience*, vol. 13, no. 11, pp. 4700–4719, 1993.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *2018 European Conference on Computer Vision (ECCV)*, September 2018.

[19] L. Matikainen, H. Kaartinen, and J. Hyyppä, "Classification tree based building detection from laser scanner and aerial image data," *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 36, no. Part 3, p. W52, 2007.

[20] M. Bertozzi, L. Bombini, P. Cerri, P. Medici, P. C. Antonello, and M. Miglietta, "Obstacle detection and classification fusing radar and vision," in *2008 IEEE Intelligent Vehicles Symposium*, 2008, pp. 608–613.

[21] M. Munz, K. C. J. Dietmayer, and M. Mahlisch, "A sensor independent probabilistic fusion system for driver assistance systems," in *2009 12th International IEEE Conference on Intelligent Transportation Systems*, 2009, pp. 1–6.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] C. Szegedy, W. Liu, Y. Jia, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[25] A. G. Howard, M. Zhu, B. Chen, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[26] G. Philipp, D. Song, and J. G. Carbonell, "Gradients explode-deep networks are shallow-resnet explained," *International Conference on Learning Representations (ICLR)*, February 2018.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[28] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[29] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *arXiv preprint arXiv:1506.02025*, 2015.

[30] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *2019 IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[31] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[33] Z. Wang, Z. Xie, and P.-C. Qiu, "Comparison of data standardization method in semantic relation similarity calculation," *Computer Engineering*, vol. 38, no. 10, pp. 38–40, 2012.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *2015 International Conference on Computer Vision (ICCV)*, December 2015.

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.