Recognizing Outdoor Scenes by Convolutional Features of Omni-directional LiDAR Scans

Kazuto Nakashima, Seungwoo Nham, Hojung Jung, Yumi Iwashita, Ryo Kurazume and Oscar M. Mozos

***Abstract*— We present a novel method for the outdoor scene categorization using 2D convolutional neural networks (CNNs) which take panoramic depth images obtained by a 3D laser scanner as input. We evaluate our approach in two outdoor scene datasets including six categories: coast, forest, indoor parking, outdoor parking, residential area, and urban area. Our results on both datasets (over 94%) outperform previous approaches and show the effectiveness of this approach for outdoor scene categorization using depth images. To analyze our trained networks we visualize the learned features by using two visualization methods.**

I. Introduction

Understanding the surrounding environment is an important capability for autonomous robots and vehicles that allows them to identify the type of their location and make better decisions accordingly. Environment understanding is a challenging task that requires high level interpretation of the sensor data and generalization capabilities to reason about a variety of environments including unseen previous ones.

In this paper we address the problem of place categorization in which a robot should determine the type of the place where it is located. Information about the place greatly improves communication between robots and humans [1], [2]. It also allows autonomous robots to make context-based decisions to complete high-level tasks [3]–[7]. Moreover, information about the type of the place can be used to build semantic maps of environments [8], [9], and high level conceptual representations of a space [10], [11]. Finally, an autonomous vehicle able to determine the type of its location can make better context-based decisions [12]. As an example, a vehicle can lower the speed when driving through a residential area.

This paper focuses on semantic categorization of places in outdoor scenarios using depth panoramic images obtained by 3D laser sensors. Fig 1 shows an example of depth panoramic images in Sparse MPO dataset [13]. Depth images are more robust to changes in illumination, which is a big advantage when navigating in outdoor environments. The main novelty of this paper is the use of deep learning to learn the different outdoor categories. In particular we use a Convolutional Neural Network (CNN) to classify the panoramic depth images. In addition, we compare our categorization results using CNNs with previous approaches [13] obtaining always better performances.

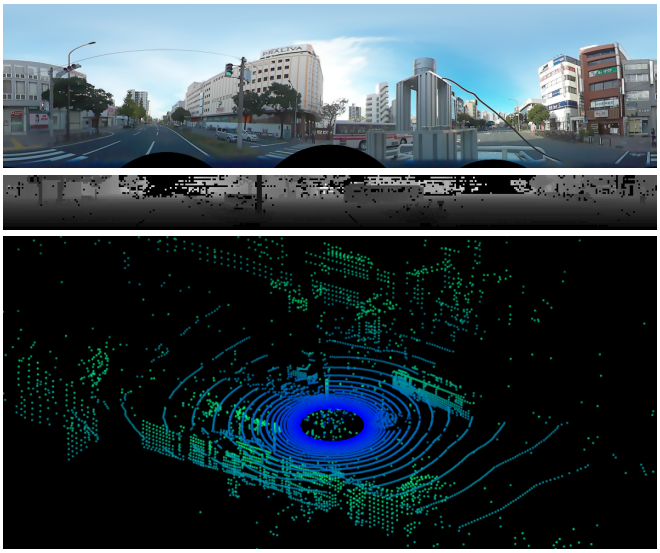

Fig. 1. Samples from the Sparse MPO dataset [13]. **Top:** Panoramic color image of the urban scene. **Middle:** Corresponding panoramic depth image directly obtained by LiDAR scanning. **Bottom:** 3D point cloud converted from the image. Panoramic depth image represents the surrounding geometrical information compactly.

II. Related Work

Outdoor place recognition has been studied in computer vision and robotics, and it is now highly motivated because of its application to outdoor autonomous robots and vehicles.

Color images have been used for the recognition tasks at the instance and category level. As an example of instance-level recognition Torii et al. [14] densely extracts SIFT [15] features from the query image and retrieves the most similar image from a view-synthesized database. Conversely, our aim is the category-level recognition, that targets unknown places and predicts their semantic categories. In this sense, Lazebnik et al. [16] proposed a spatial pyramid image representation extending the bag-of-features approach [17] and applied it to 15-scene categories. In addition, Xiao et al. [18] applied several local and global feature techniques for image classification in the SUN dataset. Moreover, a histogram of oriented uniform patterns of images is applied for place recognition and categorization in outdoor environments in [19].

Kazuto Nakashima Seungwoo Nham, and Hojung Jung are with the Graduate School of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, Japan. {k_nakashima, nham, hojung}@irvs.ait.kyushu-u.ac.jp

Yumi Iwashita is with the Jet Propulsion Laboratory, California Institute of Technology, M/S 198-235 4800, Oak Grove Drive Pasadena, 91109 CA, USA. Yumi.Iwashita@jpl.nasa.gov

Ryo Kurazume is with the Faculty of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, Japan. kurazume@ait.kyushu-u.ac.jp

Oscar M. Mozos is with Technical University of Cartagena (UPCT), Spain. oscar.mozos@upct.es

Alternatively, depth information has been used to categorize places. The work in [8] extracts geometrical features from 2D laser scanners to categorize indoor environments. The approach in [20] transforms the laser depth readings into images that are classified using CNNs. Also, 3D depth information is processed to categorize indoor places using RGB-D sensors [21], [22] and 3D laser scans [23].

In addition, different sensor modalities have been combined for the categorization of places. In [24] a support vector machine is used to combined classifications from camera images and 2D laser readings, and in [23] depth information is combined with reflectance data to extend the feature vector.

A common problem in the image-based techniques is how to select efficient visual features. Recently, convolutional neural networks have gained a great deal of attention as a powerful method to automatically extract the image features in visual recognition tasks such as object recognition [25], object detection [26], and semantic segmentation [27].

For color-based place recognition, Arandjelović et al. [28] proposed a CNN architecture to recognize instances of places by treating the problem as an image retrieval task. Gomez-Ojeda et al. [29] trained a CNN based on triplet loss calculated from three instances, for the purpose of recognizing revisited places under significant appearance changes. Sünderhauf et al. [30] investigated the performance of CNNs as an image descriptor and its robustness to appearance and viewpoint changes. At the category-level recognition, Zhou et al. [31] used a CNN for scene recognition in the Places 205 dataset. Finally, Uršič et al. [32] proposed a part-based model of household space categories based on CNNs.

While there have been several methods using range images to recognize object categories [33], [34], estimate object shapes [35], or detect vehicles for autonomous driving [36], not many works have focused on range images to solve the place recognition task using CNNs. Sizikova et al. [37] used generated synthetic 3D data to train a CNN for indoor place recognition, however, they presented this task as an image matching problem at the instance level. As for works closely related to us, Goeddel et al. [20] proposed a place categorization technique using a CNN to classify household places such as a room, a corridor, or a doorway. Song et al. [38] proposed the SUN RGB-D indoor scene database and performed scene categorization by concatenating color-based and depth-based CNN features. Still, these works are limited to indoor applications.

In this paper, we aim to predict generic categories in outdoor environments by integrating panoramic depth images and CNNs. Depth images are more robust to changes in illumination, which is a big advantage for autonomous vehicles navigating outdoors.

III. Panoramic Depth Image Categorization using Convolutional Neural Networks

This section describes CNNs used in this work to automatically learn feature representations from panoramic depth images. Panoramic depth images are represented by grayscale values which are proportional to the distance measure. Our CNN is implemented with the deep learning framework PyTorch [39], and the learning processes are performed on the NVIDIA Geforce GTX Titan X.

TABLE I
Network Architecture

Layer type	Details
Convolution	3×3 kernel with stride 1, 32 filters, ReLU
Max Pooling	2×2 kernel with stride 2
Convolution	3×3 kernel with stride 1, 32 filters, ReLU
Max Pooling	2×2 kernel with stride 2
Fully-connected	128 units, 50% Dropout, ReLU
Fully-connected	6 units
Softmax	

A. Preprocessing

The panoramic depth images given to our networks are converted from point clouds measured by 3D LiDAR scans. Each point on the scan is mapped into a 2D plane by cylindrical projection around the vertical axis. The obtained 2D map is scaled by the max limit value of the LiDAR and fed into the network.

B. Network Architecture

Table I shows the architecture of the CNN used in this work. Our network consists of two convolutional layers and two fully connected layers and it is empirically designed. As an activation function of the convolutional layers, we employ rectified linear units (ReLUs), which only pass element-wise positive values. Then max-pooling with a 2×2 window is applied to the outputted feature map without overlapping, achieving translation invariance. The first fully connected layer is also activated by the ReLU function and followed by the second layer. The final output is fed to a softmax function to infer class probabilities.

C. Data Augmentation

In general, millions of images are required for training deep networks. However, we can improve the generalization capabilities by artificially augmenting the training data. In our case, we extend the training data by applying two types of random transformations to the original image set. First, the input image is horizontally flipped. Second, we perform a random circular shift on the image in the horizontal direction; this is equivalent to rotating panoramic images in the yawing direction. The number of shifted pixels is randomly selected from zero to the image width.

D. Training

The weights of the network are trained by backpropagation algorithm using stochastic gradient descent. In each propagation, a mini-batch of images is fed into the network, which estimates a categorical distribution of each image. In order to solve the N-class classification problem, the network is trained by minimizing a cross entropy cost between the outputted probabilities and their ground-truths, which is defined as follows:

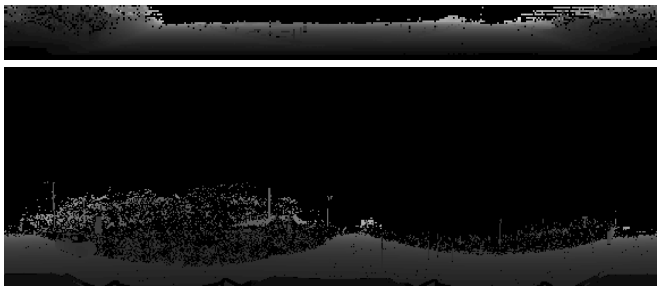

Fig. 2. Panoramic depth images of coast scenes from the Sparse MPO (top) and Dense MPO (bottom) [13]. They have different resolution vertically and horizontally

TABLE II
DISTRIBUTION OF DEPTH PANORAMIC IMAGES BY CATEGORY IN THE SPARSE-MPO DATASET AND THE DENSE-MPO DATASET

Category	**Number of scans**	
	Sparse MPO	**Dense MPO**
Coast	4,298	103
Forest	6,479	116
Indoor parking	4,780	105
Outdoor parking	5,445	108
Residential area	7,464	106
Urban area	5,734	112
Total	34,200	650

$$\mathcal{L}(\boldsymbol{d}, \boldsymbol{p}) = -\frac{1}{M} \sum_{m=1}^{M} \boldsymbol{d}_{\boldsymbol{m}}^{T} \log \boldsymbol{p}_{\boldsymbol{m}} \tag{1}$$

where M is the size of the mini-batch, $\boldsymbol{p_m}$ is the N-dimensional softmax output of the m-th image, and $\boldsymbol{d_m} \in \{0, 1\}$ is the corresponding ground-truth distribution. Finally, the gradients of the cost are calculated w.r.t. the weights on each layer by backpropagation and the weight updating is performed. We empirically set the training mini-batch size to 64.

IV. EXPERIMENTS

A. Datasets

We conduct experiments on two types of outdoor scene datasets Sparse MPO and Dense MPO presented in our previous work [13]. The datasets are composed of 3D point clouds measured by two different 3D laser scanners. In this work we convert the point clouds to 2D panoramic depth images and feed to the CNNs. The data were acquired in the city of Fukuoka (Japan) while driving and locating a car in 6 different types of areas: *coast*, *forest*, *residential area*, *urban area*, *indoor parking*, and *outdoor parking*. The datasets are publicly available [40], [41]. Example panoramic images of coast scene are shown in Fig. 2. The number of scans in each category is shown in Table II.

The Sparse MPO dataset contains 34,200 point clouds with resolution 2166×32, obtained using a Velodyne HDL-32E LiDAR on top of a vehicle. Each category in this dataset contains 10 different sets of scans, each corresponding to a different trajectory in the same place category.

The Dense MPO dataset contains 650 point clouds with high resolution 5140×1757, obtained using a FARO Focus3D laser scanner installed on top of a vehicle. Each category contains 7 different sets of scans. Each set contains scans obtained at a different place inside the same place category.

B. Experiment Settings

To speed-up the training process we reduced the size of the original panoramic images by downsampling using bilinear interpolation. Panoramic depth images were downsampled to 384×32 in the sparse dataset, and to 576×192 in the dense dataset. Our validation uses a k-fold cross-validation approach in which we use one set in each category for testing and the rest for training, and repeat this process k times with different disjoint training and tests sets. In a training phase, we extract a validation set from training sets to monitor overfitting. The numbers of folds k are 10 for the Sparse and 7 for the Dense MPO datasets, according to the scanning groups of the datasets.

We use the Adam [42] algorithm as a weight updating method to adjust the learning rate automatically. We set the hyperparameters β_1 and β_2 of the Adam to 0.9 and 0.999 respectively as recommended in [42]. As for the step size α, we fix at 1×10^{-4} for the both datasets. Furthermore, we perform L_2-regularization on the network by weight decay [43] to mitigate the over-fitting risk. Weight decay is applied by adding L_2-norm of the weights to the cost defined in Eq. (1). The coefficient of the regularization term is fixed at 5×10^{-4}. Dropout regularization of 50% is added to the output of the first fully-connected layer; this regularization allows us to explore generalized weights while avoiding over-fitting. When the validation loss does not improve over 10 epochs, i.e. it overfits or plateaus on the training sets, the training is stopped early and then we use the parameters for testing.

C. Results

1) Categorization performance: The categorization results for the panoramic depth images in the Sparse and Dense MPO datasets applying CNNs provide correct classification rates (CCR) of 94.31 ± 2.58 and 94.93 ± 6.04 respectively. The corresponding confusion matrices are shown in Table III and Table IV.

We compare the categorization results with the two approaches presented in [13] using the same datasets. The first approach obtains a simple spin image representation from the whole panoramic scan as a feature vector. The second approach applies local binary patterns (LBP) to obtain a global descriptor for the depth images. In both cases support vector machines are used as final classifiers. Table V shows the correct classification results using the approaches in [13] in comparison with our CNN-based method. In both datasets, our approach using CNNs outperforms the previous methods.

TABLE III

CONFUSION MATRIX [%] FOR THE SPARSE DATASET USING CNNS

	Coast	Forest	In. P.	Out. P.	Res.	Urban
Coast	**86.84**	8.27	0.00	1.67	2.83	0.39
Forest	5.37	**93.31**	0.03	0.09	1.16	0.03
Indoor parking	0.10	0.13	**95.98**	1.89	1.78	0.13
Outdoor parking	1.38	0.06	1.82	**94.01**	1.82	1.02
Residential area	1.23	0.58	0.06	1.49	**96.34**	0.30
Urban area	0.35	0.11	0.00	0.48	1.27	**97.79**

TABLE IV

CONFUSION MATRIX [%] FOR THE DENSE DATASET USING CNNS

	Coast	Forest	In. P.	Out. P.	Res.	Urban
Coast	**94.12**	2.94	0.00	0.00	2.94	0.00
Forest	5.04	**90.76**	0.84	3.36	0.00	0.00
Indoor parking	0.00	0.00	**100.0**	0.00	0.00	0.00
Outdoor parking	0.00	0.00	10.81	**87.39**	0.00	1.80
Residential area	0.00	0.00	0.00	0.96	**99.04**	0.00
Urban area	0.00	0.88	0.00	0.00	0.00	**99.12**

2) Majority vote on sequential predictions: For the continuous trajectories of the Sparse MPO dataset, we additionally apply a majority vote approach to classify the current location using the previous n predictions as suggested in [13]. This smoothing method has shown to improve categorization results in continuous trajectories. We applied this technique using the labels outputted by the three methods and compared the results for different number of votes n. As Fig. 3 shows, CNNs always outperform the previous methods also when using majority vote. In addition, Table VI compares results for the three approaches when the optimal number of votes is selected for each method. Our CNN provides the better results with the minimum number of necessary votes.

D. Qualitative Analysis

In order to verify the way the CNN solves our categorization problem, we carried out two additional experiments. First, we analyzed the CNN response when the test images are partially hidden by an image patch [44] as the one shown in Fig. 4. The size of the partial image patch is 8×8 and the mean value of all training images is assigned uniformly to the pixels. The softmax values of the correct categories (certainty score) calculated by shifting the patch on a whole image is defined as occlusion sensitivity [44]. The obtained maps of the occlusion sensitivity for all categories are depicted as "Occlusion Map" in Fig. 5. Note that the size of the score map is 377×25, which is smaller than the input image since it was not padded around the border. Thus, the score map in Fig. 5 is stretched and the pixel position in the Score map does not coincide with the input image.

Next, we visualize pixel-space features contributing to correct classification using Guided Backpropagation [45]. In accordance with the method, we first propagated an image and obtained scores without applying softmax. Then the score of the correct category was backpropagated to the pixel-space with only positive gradients. The obtained featuremaps are depicted as "Guided Backpropagation" in Fig. 5.

TABLE V

COMPARISON TO PREVIOUS APPROACHES

Method	**CCR [%]**	
	Sparse MPO	**Dense MPO**
Spin Image + SVM [13]	79.23 ± 4.51	89.43 ± 2.65
LBP + SVM [13]	92.00 ± 4.62	91.30 ± 2.74
Ours	$\mathbf{94.31 \pm 2.58}$	$\mathbf{94.93 \pm 6.04}$

TABLE VI

CCR WITH OPTIMAL NUMBER OF VOTES n'

Method	**Number of votes n'**	**CCR [%]**
Spin Image + SVM [13]	40	88.34
LBP + SVM [13]	30	93.74
Ours	**7**	**99.42**

We focus on and discuss the three distinctive categories, *coast*, *forest*, and *outdoor parking*. From Fig. 5(a), the top yellow area of the image is very sensitive to occlusion. In this area, the range data are hardly obtained due to the sky or the sea, and thus this empty area is important to classify images as *coast*. If this area is covered by crossing cars or trees, this can produce misclassification for the *coast* label because this category is also influenced by the left and right wooded areas, where the nodes in the CNN are highly activated. For the *forest* category, however, the center and side wooded areas are activated and very sensitive to occlusion as shown in Fig. 5(b). This means that the areas wooded widely are important to classify images as *forest*. In addition, Fig. 5(c) shows that the upper area is important to classify images as *outdoor parking*. If this upper area is covered by buildings, the images are likely to be classified as *indoor parking*. For example, the bottom featuremaps in Fig. 5(c)(d) indicates that the lower area, which captures parked cars, is slightly activated. Over all the categories, border areas between the measurable things and the sky, i.e. skylines, are activated strongly.

V. CONCLUSION

In this paper we presented a new approach for the outdoor scene categorization using the convolutional neural networks (CNNs) which take panoramic depth images as input. We applied our method to two different outdoor panoramic datasets and obtained high categorization rates (over 94%) for six outdoor scene categories: *coast*, *forest*, *indoor parking*, *outdoor parking*, *residential area*, and *urban area*. In addition, we compared our approach to previous methods and obtained higher categorization performance than previous methods. Finally, we validated our CNN by presenting its score and gradients.

Future work will include the optimization of the CNN structure and the selection and combination of optimum features depending on the environment to improve the categorization performance. Moreover, the number of outdoor scene categories could be increased, although it is difficult to clearly define them even when using human perception.

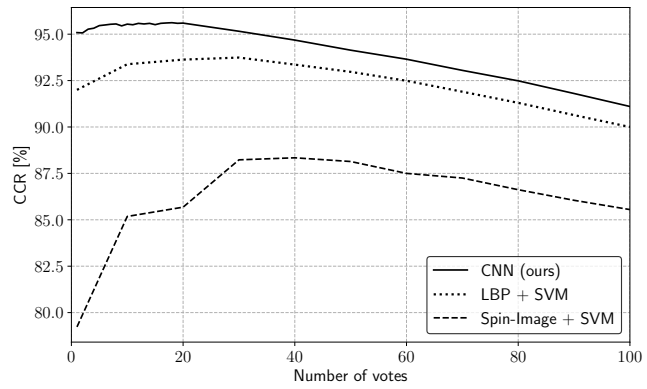

Fig. 3. Number of votes n VS. CCR

Fig. 4. Example of an occluded image by a mean-color patch.

ACKNOWLEDGEMENT

The present study was supported in part by a Grant-in-Aid for Scientific Research (A) (26249029).

REFERENCES

[1] H. Zender, O. M. Mozos, P. Jensfelt, G.-J. M. Kruijff, and W. Burgard, "Conceptual spatial representations for indoor mobile robots," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 493–502, June 2008.

[2] A. Pronobis and P. Jensfelt, "Large-scale semantic mapping and reasoning with heterogeneous modalities," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, Saint Paul, MN, USA, May 2012, pp. 3515–3522.

[3] C. Stachniss, O. M. Mozos, and W. Burgard, "Efficient exploration of unknown indoor environments using a team of mobile robots," *Annals of Mathematics and Artificial Intelligence*, vol. 52, no. 2-4, pp. 205–227, April 2008.

[4] ——, "Speeding up multi-robot exploration by considering semantic place information," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2006, pp. 1692–1697.

[5] H. Zender, P. Jensfelt, and G.-J. Kruijff, "Humanand situation-aware people following," in *Proc. of the IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2007, pp. 1131–1136.

[6] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*. IEEE, 2003, pp. 273–280.

[7] T. Kollar and N. Roy, "Utilizing object-object and object-scene context when planning to find things," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2009, pp. 2168–2173.

[8] O. M. Mozos, C. Stachniss, and W. Burgard, "Supervised learning of places from range data using adaboost," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2005, pp. 1730–1735.

[9] A. Pronobis, P. Jensfelt, K. Sjöö, H. Zender, G.-J. M. Kruijff, O. M. Mozos, and W. Burgard, "Semantic modelling of space in cognitive systems," *ser. Cognitive Systems Monographs, H. I. Christensen, A. Sloman, G.-J. M. Kruijff, and J. Wyatt, Eds.*, pp. 165–221, 2010.

[10] H. Zender, O. M. Mozos, P. Jensfelt, G.-J. M. Kruijff, and W. Burgard, "Conceptual spatial representations for indoor mobile robots," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 493–502, 2008.

[11] H. I. Christensen, G. Kruijff, and E. J. Wyatt, *Cognitive Systems, ser. COSMOS*. Springer Verlag, 2010.

[12] J. Ibañez-Guzman, C. Laugier, J.-D. Yoder, and S. Thrun, "Autonomous driving: Context and state-of-the-art," in *Handbook of Intelligent Vehicles*. Springer, 2012, pp. 1271–1310.

[13] H. Jung, Y. Oto, O. M. Mozos, Y. Iwashita, and R. Kurazume, "Multi-modal panoramic 3D outdoor datasets for place categorization," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Daejeon, Korea, October 2016, pp. 4545–4550.

[14] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1808–1817.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[16] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2006, pp. 2169–2178.

[17] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[18] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, *et al.*, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, June 2010, pp. 3485–3492.

[19] E. Fazl-Ersi and J. K. Tsotsos, "Histogram of oriented uniform patterns for robust place recognition and categorization," *The Int. Journal of Robotics Research*, vol. 31, no. 4, pp. 468–483, 2012.

[20] R. Goeddel and E. Olson, "Learning semantic place labels from occupancy grids using cnns," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2016, pp. 3999–4004.

[21] O. M. Mozos, H. Mizutani, R. Kurazume, and T. Hasegawa, "Categorization of indoor places using the kinect sensor," *Sensors*, vol. 12, no. 5, pp. 6695–6711, 2012.

[22] E. Fernandez-Moral, W. Mayol-Cuevas, V. Arvalo, and J. Gonzalez-Jimenez, "Fast place recognition with plane-based maps," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, May 2013, pp. 2719–2724.

[23] O. M. Mozos, H. Mizutani, H. Jung, R. Kurazume, and T. Hasegawa, "Categorization of indoor places by combining local binary pattern histograms of range and reflectance data from laser range finders," *Advanced Robotics*, vol. 27, no. 18, pp. 1455–1464, 2013.

[24] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *Int. Journal of Robotics Research*, vol. 29, no. 2-3, pp. 298–320, 2010.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. of the European Conf. of Computer Vision (ECCV)*. Springer, 2016, pp. 21–37.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[28] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5297–5307.

[29] R. Gomez-Ojeda, M. Lopez-Antequera, N. Petkov, and J. Gonzalez-Jimenez, "Training a convolutional neural network for appearance-invariant place recognition," *arXiv preprint arXiv:1505.07428*, 2015.

[30] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 4297–4304.

[31] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 487–495.

[32] P. Uršič, R. Mandeljc, A. Leonardis, and M. Kristan, "Part-based room categorization for household service robots," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, May 2016, pp. 2287–2294.

[33] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 681–687.

[34] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *Proc. of the IEEE/RSJ*

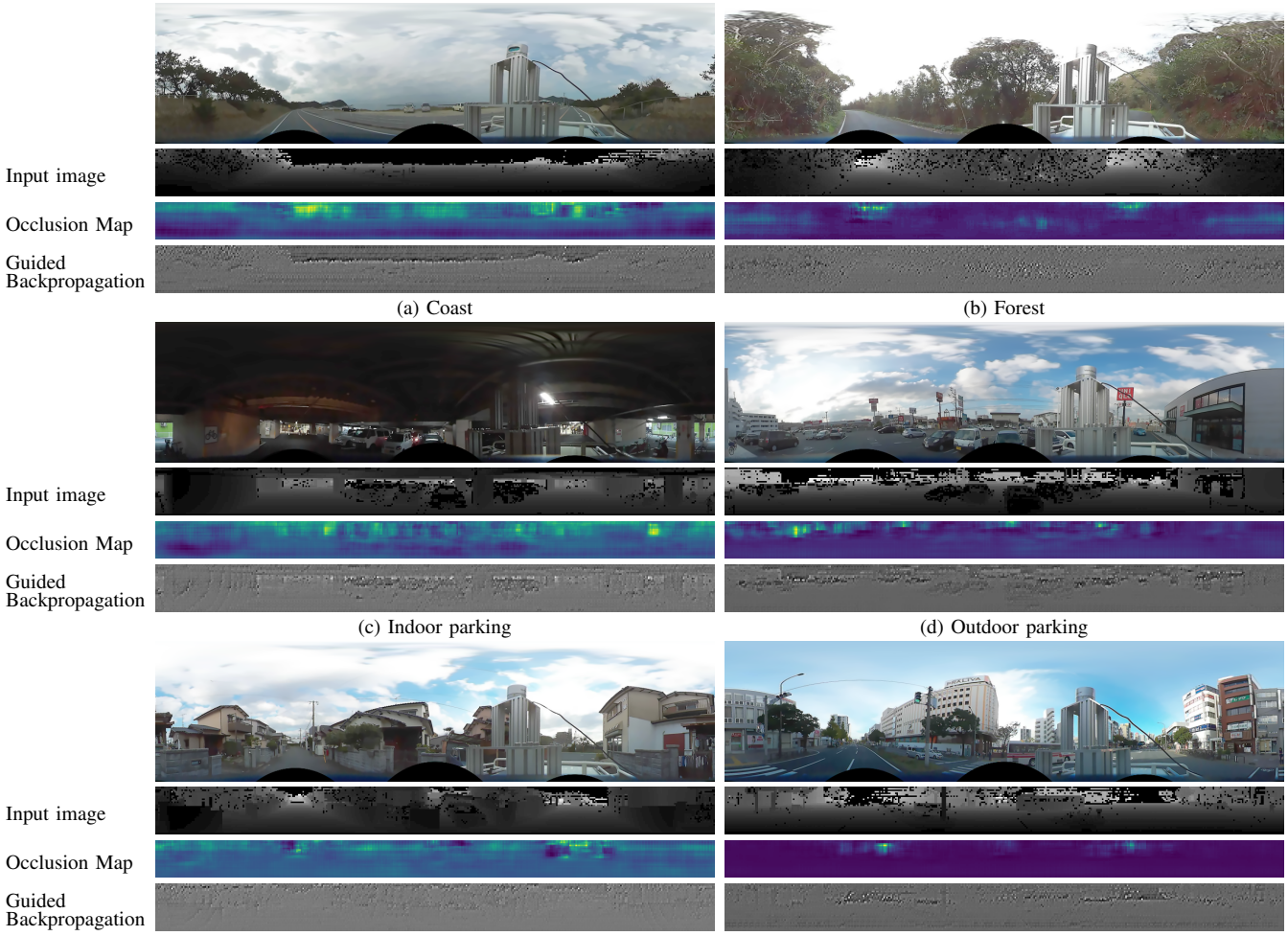

Fig. 5. Each group shows, from top to down, an example of color scenery images, a corresponding depth image to input, "Occlusion Map" which is a map of softmax scores w.r.t. the correct category generated by moving a occlusion patch [44], and "Guided Backpropagation" visualizing contributing features in pixel-space, which is reconstructed by backpropagating the activations of the last layer with positive gradients [45]. In Occlusion Maps, the high and low scores appear in purple and yellow respectively. A yellow area indicates that it is more important for correct classification. Best viewed in color.

Int. Conf. on Intelligent Robots and Systems (IROS). IEEE, 2015, pp. 922–928.

[35] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1912–1920.

[36] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3d lidar using fully convolutional network," *arXiv preprint arXiv:1608.07916*, 2016.

[37] E. Sizikova, V. K. Singh, B. Georgescu, M. Halber, K. Ma, and T. Chen, "Enhancing place recognition using joint intensity-depth analysis and synthetic data," in *Computer Vision–ECCV 2016 Workshops*. Springer, 2016, pp. 901–908.

[38] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, Massachusetts, June 2015, pp. 567–576.

[39] "Pytorch," http://pytorch.org/.

[43] J. Moody, S. Hanson, A. Krogh, and J. A. Hertz, "A simple weight

[40] "Sparse Multi-modal Panoramic 3D Outdoor Dataset for Place Categorization," 2016, http://robotics.ait.kyushu-u.ac.jp/~kurazume/research-e.php?content=db#d08.

[41] "Dense Multi-modal Panoramic 3D Outdoor Dataset for Place Categorization," 2016, http://robotics.ait.kyushu-u.ac.jp/~kurazume/research-e.php?content=db#d07.

[42] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
decay can improve generalization," *Advances in Neural Information Processing Systems (NIPS)*, vol. 4, pp. 950–957, 1995.

[44] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. of the European Conf. of Computer Vision (ECCV)*. Springer, 2014, pp. 818–833.

[45] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2015.