PlaceNet: Neural Spatial Representation Learning with Multimodal Attention

Chung-Yeon Lee[1,2] , **Youngjae Yoo**[1] , **Byoung-Tak Zhang**[1,3]
[1]Seoul National University,
[2]Surromind,
[3]AIIS
{cylee, yjyoo, btzhang}@bi.snu.ac.kr

Abstract

Spatial representation capable of learning a myriad of environmental features is a significant challenge for natural spatial understanding of mobile AI agents. Deep generative models have the potential of discovering rich representations of observed 3D scenes. However, previous approaches have been mainly evaluated on simple environments, or focused only on high-resolution rendering of small-scale scenes, hampering generalization of the representations to various spatial variability. To address this, we present PlaceNet, a neural representation that learns through random observations in a self-supervised manner, and represents observed scenes with triplet attention using visual, topographic, and semantic cues. We train the proposed method on a large-scale multimodal scene dataset consisting of 120 million indoor scenes, and demonstrate that PlaceNet successfully generalizes to various environments with lower training loss, higher image quality and structural similarity of predicted scenes, compared to a competitive baseline model. Additionally, analyses of the representations show that PlaceNet activates more specialized and larger numbers of kernels in the spatial representation, capturing multimodal spatial properties in complex environments.

1 Introduction

While extensive research on spatial perception has been carried out in AI, a significant challenge still remains for a generic representation capable of learning a myriad of environmental features. A flexible, efficient, and informative spatial representation is required for natural spatial understanding of mobile AI agents, which enables the agents to reliably navigate and interact with the objects they encounter [Gupta *et al.*, 2017; Herweg and Kahana, 2018].

Current mobile technology in robots and autonomous vehicles recognizes the surrounding space by a comparison directly between sensory inputs and a metric map represented by simple topographical structures or visual cues (e.g. position of landmarks, obstacles). However, metric map representations have limited scalability for large-scale and dynamic environments, inefficient mapping processes, and lack of expressiveness [Bailey and Durrant-Whyte, 2006; Cadena *et al.*, 2016].

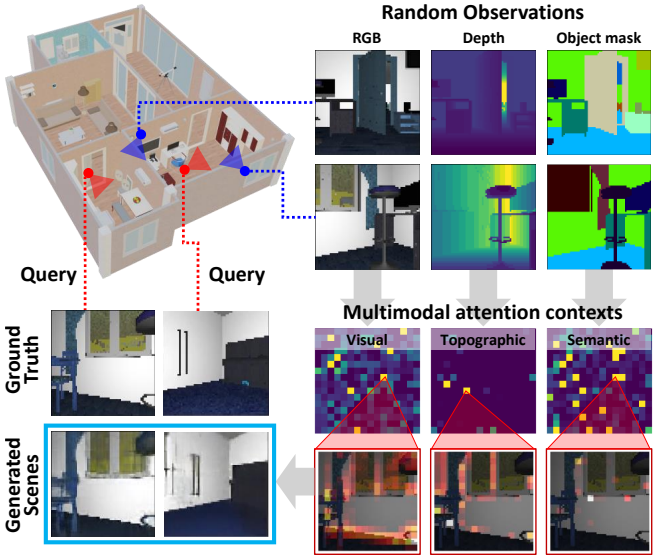

Figure 1: Generating scenes from queried viewpoints based on a spatial representation learned with multimodal contexts of random observations in a 3D indoor environment.

On the other hand, recent progress in deep generative networks has given rise to a new approach that discovers rich representations of 3D scenes, namely neural rendering [Tewari *et al.*, 2020]. Neural rendering approach encodes a scene representation by learning to generate novel views of the scene. However, this approach has been focused only on high-resolution rendering of specific small-scale scenes [Sitzmann *et al.*, 2019; Mildenhall *et al.*, 2020], or evaluated on simple artificial environments with only a few objects [Eslami *et al.*, 2018; Tobin *et al.*, 2019], both hampering generalization of the representation to spatial variability.

To address this, we train deep generative networks with a large-scale 3D house dataset to generalize a neural representation across various indoor environments. Our spatial representation network, named PlaceNet*, deals with spa-

*Named after *Place cells* [O'Keefe, 1976; Moser *et al.*, 2008].

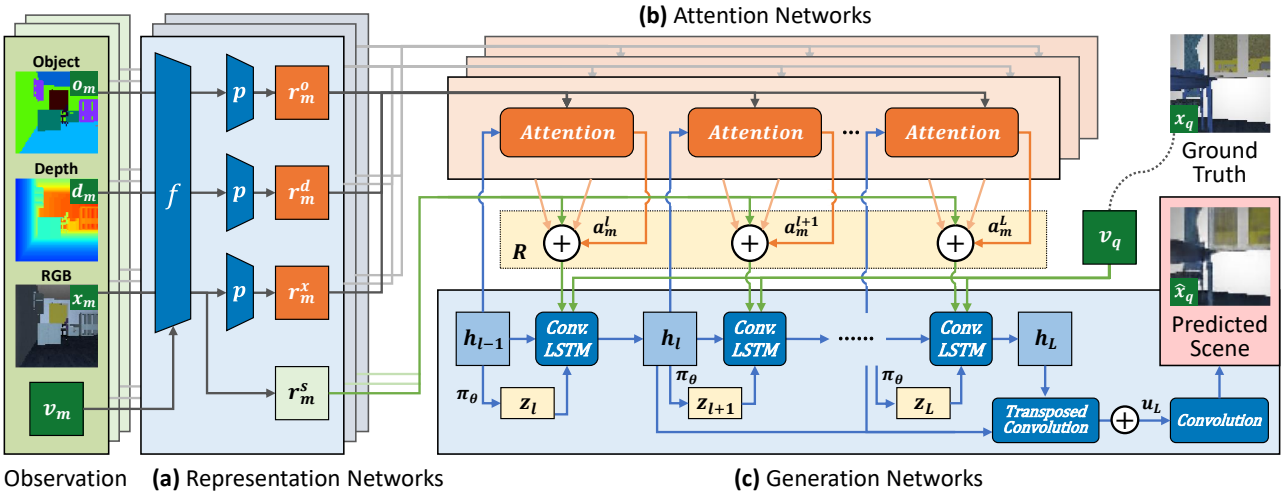

Figure 2: **Network architecture of PlaceNet.** (a) Representation networks encode multimodal observations. (b) Attention networks extract spatial contexts a_m^l and update the spatial representation R. (c) Generation networks predict a target scene $\hat{x}_q$ corresponding to a queried viewpoint v_q by recurrently sampling latent variables z from hidden states h based on R. Note that "+" denotes element-wise summation, and yellow boxes are intermediate representations in each training step.

tial variability of indoor scenes, with an attention mechanism that leverages multimodal scene components, including color, depth images, and semantic masks, as shown in Figure 1.

Generative Query Networks (GQN), one of the breakthroughs in neural rendering, impressively demonstrated that a single neural architecture could learn to perceive, interpret and represent various synthetic 3D scenes without any explicit supervision [Eslami *et al.*, 2018]. However, the single global scene representation network of GQN has limitations in learning diverse spatial features, and shows degraded performance when rendering more complex scenes such as the Minecraft world [Rosenbaum *et al.*, 2018] and 3D house environments [Wu *et al.*, 2018].

In contrast, PlaceNet's attention mechanism propagates additional relevant spatial contexts extracted from different scenes to the hidden states of the model's decoder, allowing the representation network to learn the various spatial features in complex scenes. Consequently, PlaceNet encodes random observations, produces visual, topographic, and semantic contexts with the triplet attention networks, and generates unobserved visual scenes based on the spatial representation updated by the multimodal attention contexts.

Our approach also can be positioned in the context of Novel-View Synthesis (NVS) which constructs 3D structures of the scene from multiple camera views to render unseen parts of the scene [Chen and Williams, 1993; Zhou *et al.*, 2016]. The difference between our work and other NVS methods is the range of scenes generalized and the targets to be optimized. Recent NVS approaches, such as DeepVoxels [Sitzmann *et al.*, 2019] and NeRF [Mildenhall *et al.*, 2020] mainly focus on representing intra-scene variability, limiting their scope to rendering high-resolution views on a few specific scenes. These NVS methods optimize neural networks for volumetric representation, using scene- or object-specific continuous views collected from various angles. In contrast, our approach focuses more on learning generalizable spatial properties by optimizing the neural representations based on multimodal inter-scene variability across numerous scenes. While training, our approach samples a small number of discontinuous contexts from arbitrary viewpoints, allowing for a better representation of the scene that is less dependent on local features that can be inferred by continuous views.

2 PlaceNet

2.1 Problem Formulation

As a neural scene rendering problem, we consider a framework that trains an implicit neural representation that can accurately generate a visual scene at an arbitrary viewpoint based on one or more observed scenes.

In this framework, a scene consists of a collection of multimodal images including an RGB image x, depth image d, object-aware segmentation mask o, and the corresponding viewpoint v. The observations C contains M randomly sampled scenes: $C = \{(x_m, d_m, o_m, v_m)\}_{m=1}^{M}$, where M is also randomized at every training step for better generalization.

PlaceNet receives a set of multimodal observations C and a randomly sampled query viewpoint v_q as inputs, and should be able to render a plausible scene $\hat{x}_q$ corresponding to v_q. In detail, the model parameterizes a conditional distribution $P_\theta(x_q|C, v_q)$, from which the images are generated.

2.2 Encoding

Firstly, given multimodal observations of a scene, a convolutional encoder f encodes them into low-dimensional latent vectors, as illustrated in Figure 2(a). These latent vectors are aggregated to context representations r_m^x, r_m^d, r_m^o, respectively, through an average pooling layer p, as following:

$$r_m^c = p(f(c_m, v_m)),\ c \in \{x, d, o\} \tag{1}$$

The triplet context representations are then fed into the attention networks to produce an attention context a_m^l, as

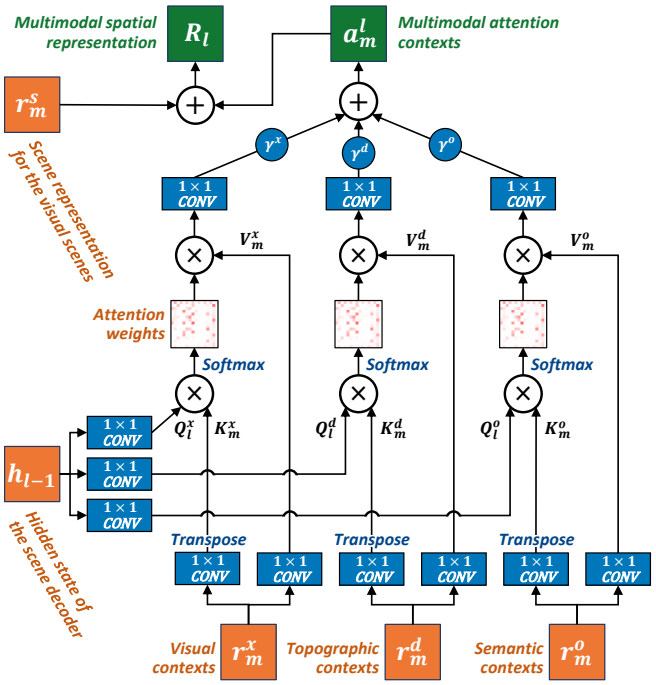

Figure 3: **Multimodal attention mechanism.** The key K_m and value V_m are computed by linear projection of each modal's representation, and the previous hidden state h_{l-1} from the scene decoder is deployed to the query Q_l by linear projection. Then, K_m and Q_l are matrix-multiplied to form attention weights, and multiplied again by V_m to get an attention score. All of the attention scores are linearly projected to equalize the output dimension and summed element-wise, and aggregated with the scene representation r^s to produce the multimodal spatial representation R. Note that orange boxes are input contexts, and green boxes are the aggregated attention contexts. "×" denotes batch-wise matrix multiplication.

shown in Figure 2(a-b). Also, latent vectors of the image modality to be generated (i.e. RGB in here) are independently aggregated to the scene representation r^s_m via an element-wise summation. Here, we use the *Tower* representation network from GQN [Eslami *et al.*, 2018] for the convolutional encoder f in order to take advantage of its spatially arranged scene representation.

2.3 Multimodal Attention

The attention networks map the triplet context representations $\{r^x_m, r^d_m, r^o_m\}$, given from the representation networks, and a hidden state h_{l-1} in the previous step of the generation network to an attention context a^l_m with:

$$a^l_m = \sum_{c\in\{x,d,o\}} \gamma^c \,\mathrm{Softmax}\left(Q^c_l \cdot (K^c_m)^\top\right) \cdot V^c_m, \quad (2)$$

where γ^c is a learnable parameter for each representation, and Q, K, V stands for queries, keys, and values for attention mechanism, respectively, as illustrated in Figure 3.

The attention contexts a^l_m represent the weighted contributions of aspects of visual, topographic, and semantic features to each spatial position within the scene representations r^s_m. a^l_m and r^s_m are then aggregated in a permutation-invariant way to obtain the spatial representation R_l as follows:

$$R_l = \sum_{m=1}^{M} r^s_m + \sum_{m=1}^{M} a^l_m \quad (3)$$

Then, we use R_l as the input to the generation networks to produce the next hidden state h_l, as depicted in Figure 2(c). Note that the spatial representation is updated in each layer l.

2.4 Generation

The generation networks parameterize an auto-regressive prior distribution over latent variables: $\pi_\theta(z|v^q, R)$, using a decoder based on the convolutional Long-Short Term Memory network (CLSTM) [Xingjian *et al.*, 2015; Gregor *et al.*, 2015], as shown in Figure 2(c).

At every iteration l ($l \leq L$), the generation networks recurrently sample the latent variable z from a prior π that is parameterized by the hidden states h, by conditioning on R_l and v_q. The hidden states h_l are updated as follows:

$$(c_l, h_l, u_l) = CLSTM(v_q, R_l, c_{l-1}, h_{l-1}, u_l, z_l), \quad (4)$$

where c refers to the cell states in the CLSTM structure, and u is a canvas matrix that represents a scene to be generated.

The updated h_l is decoded into a residual update to u_l. Finally, $\hat{x}_q$ is sampled from u_L with regularization.

2.5 Training

Given randomly sampled observation C and a query (x_q, v_q) from the dataset, the training objective is to find the model parameter θ that maximizes the expected log probability of the scene: $\mathbb{E}[\log P_\theta(x_q|C, v_q)]$.

The training is done by minimizing the expected value of the negative log-likelihood of the query image x_q given the target distribution, regularized by the cumulative KL divergence between the obtained posterior ρ^l and prior π^l distributions from l^{th} generation step as following:

$$\mathcal{L} = -\mathbb{E}_{C,z\sim\rho_\phi}\left[-\ln\mathcal{N}(x_q|\hat{x}_q) + \sum_{l=1}^{L}\mathrm{KL}(\rho^l_\phi\|\pi^l_\theta)\right], \quad (5)$$

where ρ_ϕ is an approximation to the true posterior density with variational parameter ϕ [Kingma and Welling, 2013].

Please see the supplementary material's section C for further details on our implementation.

3 Dataset

To train PlaceNet, we built a large-scale dataset consisting of about 120 million complex indoor scene images[†]. Since houses are distinguished by different layouts, interior designs, and distinct objects arranged in the space, they are considered a suitable environment as a source of data for learning complex spatial representations. We collected the scene and pose information an agent encounters while traversing the realistic 3D virtual houses, previously introduced by [Song *et al.*, 2017], and the House3D simulator [Wu *et al.*, 2018].

It is important to find optimal trajectories allowing an agent to observe every aspect of the scene efficiently while avoiding

[†]Available at: https://github.com/jamixlee/placenet

redundant visits. To accomplish this, we developed a trajectory generation algorithm, which creates natural paths for the 3D house exploration. Further details on this algorithm are given in the supplementary material A.

Our dataset consists of 115,781 training samples and 10,081 evaluation samples extracted from the 25,071 houses as the training dataset, but from different viewpoints. We further generated 6,501 samples for evaluation from 929 houses which were not used to make the training dataset. The two evaluation datasets were labeled as "seen" and "unseen" in the evaluation phase. Each data sample includes 300 frames of color, depth images, object masks, and the corresponding viewpoints. All images were rendered at size 64×64 pixels while moving through the generated trajectories. Viewpoint vectors were composed of the pose $\{x, y, z\}$ and orientation $\{yaw, pitch\}$ on the trajectory for each of the 26,000 houses.

To evaluate our model's performance according to the degree of scene complexity, we also made three subsets of our dataset. Here, we assume that the greater the number of unique pixel values constituting each image, the higher the complexity. The data complexity is thus derived by averaging the number of unique pixels within each RGB image, depth image, and object-aware segmented mask from all frames. Consequently, we divided our dataset into the upper 30% (High), the lower 30% (Low), and the remaining 40% (Medium) of the average complexity distribution for each modality of the dataset.

4 Experiments

The experiments are mainly conducted to evaluate the learning performance and scene generation performance of PlaceNet, as well as to compare our method to a competitive baseline, GQN. The proposed models are implemented with a scene encoder's dimension of 256 channels, and the convolutional LSTM's hidden state of 128 channels. The generation network has 12 layers, and weights are not shared between generation steps for better performance. The number of observations given to the models is determined randomly for the training (maximum of 20), and is fixed to 5 for the evaluation phase. For training, we use the Adam optimizer [Kingma and Ba, 2014] with initial learning rate $5e-4$, which linearly decays by a factor of 10 over 1.6M optimizer steps according to the scheduler. Additional hyper-parameters used for the training are shown in the supplementary material B.

4.1 Learning Performance

PlaceNet showed improved learning performances, showing a lower training loss and KLD than the baseline results, as seen in Figure 4(a-b). The structural similarity [Wang *et al.*, 2004] between the query and generated scene converges to above 0.6 in PlaceNet's validation result, while the similarity of GQN stops at approximately 0.53, as seen in Figure 4(c). Also, GQN's performance decreased when using multi-modal scenes with increased input channels. This is presumably because the three types of heterogeneous inputs disturb the scene representation of the model from recognizing the correct spatial structure in the scene. The loss difference between PlaceNet and the baseline is more distinct in the training of high-complexity datasets, as seen in Figure 4(d-f).

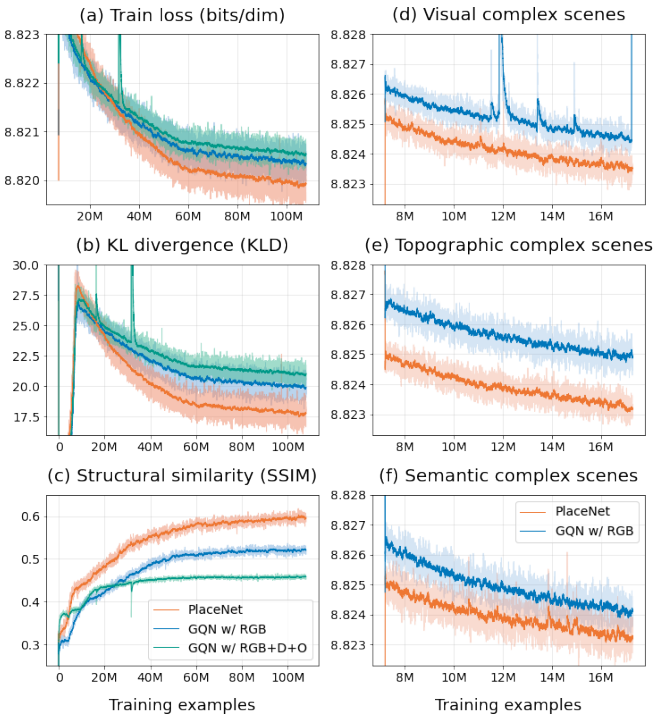

Figure 4: Training losses and structural similarity scores of the rendered scene against the number of training steps. PlaceNet outperforms the baseline with (a) higher likelihood (lower train loss), (b) lower difference between posterior and conditional prior upon observing ground-truth images, and (c) higher similarity of the rendered images at query viewpoints. (d-f) Learning scene data of higher complexity shows larger loss differences.

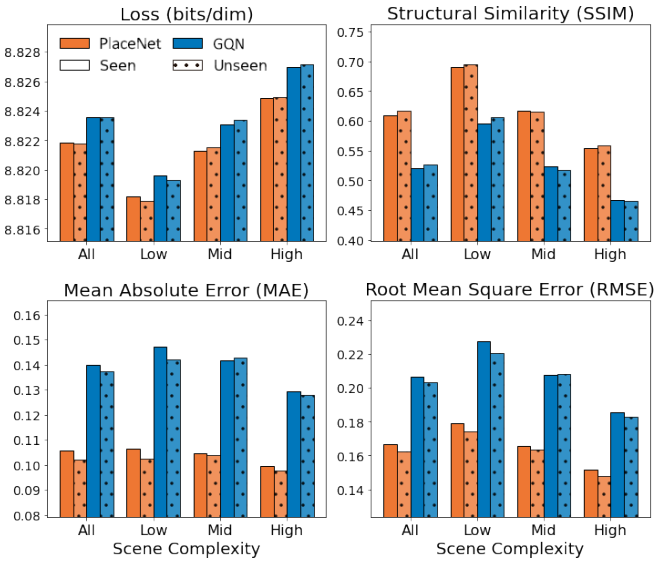

Figure 5: Evaluation results on 10K test samples according to scene complexity and novelty. The loss is calculated from negative ELBO [Kingma and Welling, 2013] for each sample, where the minimum value indicates the theoretical minimum error. MAE, RMSE, and SSIM are obtained by comparing the generated image and query.

4.2 Scene Generation Performance

We conducted extensive evaluations with several types of evaluation datasets, which are divided by scene complexity and whether the scenes are seen/unseen during training. For all evaluation settings, PlaceNet consistently shows lower errors and higher similarities, compared to the baseline model, as shown in Figure 5. Interestingly, there was no significant difference in the results between seen and unseen scenes. From this, we can infer that both models were trained sufficiently to achieve acceptable generalization performance.

The two examples shown in Figure 6 provide a glimpse into the spatial comprehension capabilities of trained representations. Here, PlaceNet and the baseline model generated scenes, given five multimodal observations and a corresponding query viewpoint on the trajectory in the house. We found PlaceNet correctly generated complicated topographic structures and specific objects in the scene, misrendered in the results of the baseline model. In Figure 6(a), PlaceNet improved the rendering quality (rmse=0.0680; ssim=0.8692) of the GQN model (rmse=0.3340; ssim=0.7217) by rendering the ground truth scene, including the window and a curtain hanging beside the window. Figure 6(b) shows that PlaceNet succeeded in rendering the wall and the window with exact colors and shapes (rmse=0.0520; ssim=0.8853), which were misrepresented in the results rendered by the baseline model without attention (rmse=0.2260; ssim=0.7112).

More examples are shown in Figure 7, including failure cases that show incorrect colors and vague patterns in Figure 7(b-c), and wrong perspective, different shape or missing parts of objects in Figure 7(c-d).

Additional results are shown in the supplementary material D and E, including scenes generated according to spatial complexity and incorrectly generated scenes.

4.3 Transferability to Real-World Scenes

We demonstrated that the spatial representation of PlaceNet trained with the virtual house dataset is also capable of predicting scenes in real-world environments. In this experiment, a mobile robot—Toyota HSR [Yamamoto *et al.*, 2018]—recorded RGB and depth images while navigating a real house, and pose information of the robot-self on the map was acquired using SLAM [Bailey and Durrant-Whyte, 2006]. The semantic segmentation masks were extracted from the RGB images using the DeepLabV3 model [Chen *et al.*, 2017] trained with the ADE20K dataset [Zhou *et al.*, 2017; Zhou *et al.*, 2019]. Finally, a pre-trained PlaceNet generated visual scenes for randomly chosen robot poses among the collected data.

As shown in Figure 8, PlaceNet successfully generated the real-world scenes containing correct visual and topographic structures of the environment as well as the objects in the environment. However, in some results, misrendered or blurry objects such as windows or sinks were found, which was expected due to the low resolution of the observed image. In contrast, the GQN model did not succeed in rendering exact spatial structures in the real-world scenes, which shows its limitation of learning complex indoor environment. These results indicate that PlaceNet can sufficiently learn the general properties of indoor scenes from the virtual house dataset.

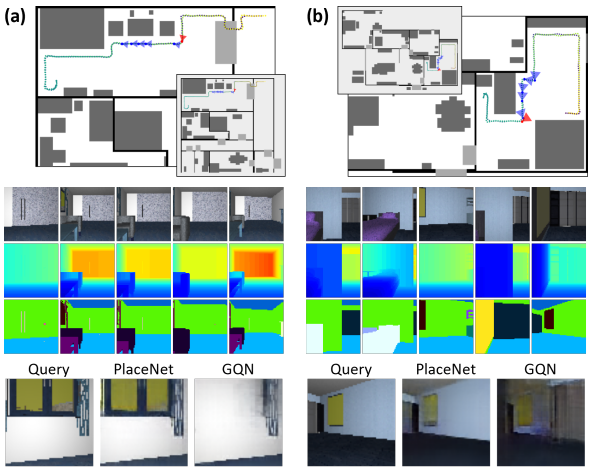

Figure 6: **Examples of our experimental results.** (Top) Observations (blue ▲) and a query (red ▲) on the trajectory in the house. (Middle) Observed multimodal scenes. (Bottom) Ground truth image corresponding to the query viewpoint, and the generated scenes.

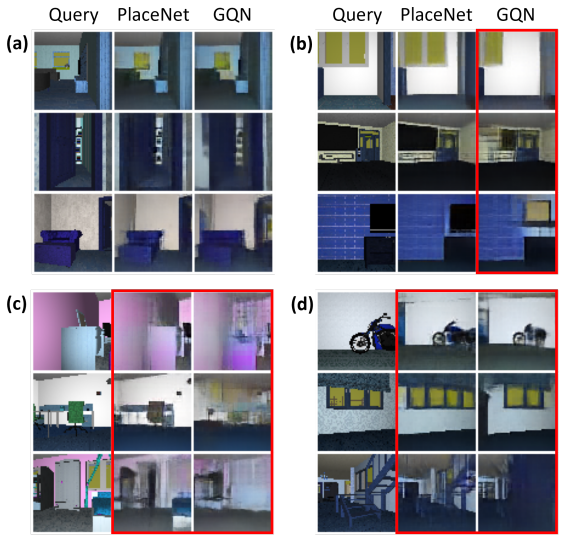

Figure 7: Examples of generated scenes that: (a) are well-rendered, (b) have misrepresented parts (w/o attention), (c) express appearance incorrectly (w/ attention) or are fully misrendered (w/o attention), and (d) have different object shapes or topographic properties.

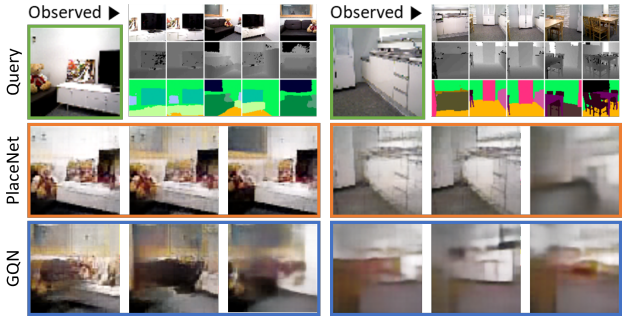

Figure 8: Examples of real-world scene generation.

4.4 Results of Scene Depth Generation

In this section, we provide experimental results of scene depth generation that show the potential of PlaceNet for cross-modal generation. While the model successfully predicted the complex topographic structures in the generated scenes, simple planar parts were generated with noise, as shown in Figure 9. We speculate this problem happened because the model reconstructs scenes using the approximate variational posterior while training (Eq. 5).

4.5 Representation Analyses

We analyzed each modality's context representations r^x, r^d, r^o in PlaceNet to see the effect of multimodal attention. This was done by visualizing saliency maps of kernel activations in the spatial representations.

The activation values shown in Figure 10(a) indicate the significant kernel positions of the spatial representation R, in terms of visual, topographic, and semantic attention.

Figure 10(b) shows that the kernel activations were spatially arranged on the regions expected to contain relevant features for the target scene. Kernels activated by visual attention overall represented distinguishable areas, generally on the edges. In contrast, topographic attentional activations focused more on planar areas, such as surfaces of objects, walls, or the floor. Semantic attentional activations have a unique characteristic that showed object-awareness while being sparsely activated overall. Also, some kernels highlighted in visual representation were occasionally indicated by representations through other modalities as well. This is presumably because visual properties are most strongly reflected in spatial representations as training losses are computed by comparing RGB scenes. Kernels of the GQN representations without attention also showed activations on reasonable areas but did not reflect spatial features precisely.

Also, we found differences in the distribution of the activation values of kernels, as shown in Figure 10(c). First, the ratio of activated kernels in PlaceNet ($82.62 \pm 4.5\%$) was higher than in GQN ($59.84 \pm 6.6\%$). Second, while GQN's activations were uniformly distributed into kernels, PlaceNet's activations distributed to relatively specific kernels and showed different patterns for each modality's representation.

To summarize, these results support that PlaceNet activates more specialized and larger numbers of kernels in the spatial representation, and the activations accurately reflect multimodal spatial properties in complex environments.

5 Concluding Remarks

In this paper, we introduced PlaceNet which learns powerful neural representations by leveraging visual, topographic, and semantic spatial features in complex 3D environments. PlaceNet can generalize across various indoor scenes by allowing hidden states of the representation decoder to depend on relevant spatial contexts extracted from the observed scenes. Experimental results show that PlaceNet is capable of learning various indoor scenes with significant improvements in neural scene rendering performance. Furthermore, the generalized representation learned by PlaceNet can be transferred to predict real-world scenes as well.

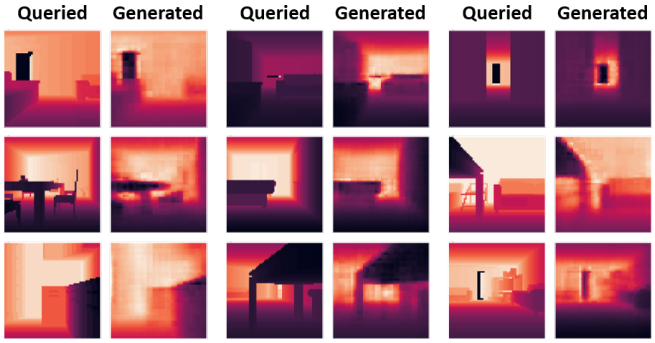

Figure 9: Examples of depth image generation.

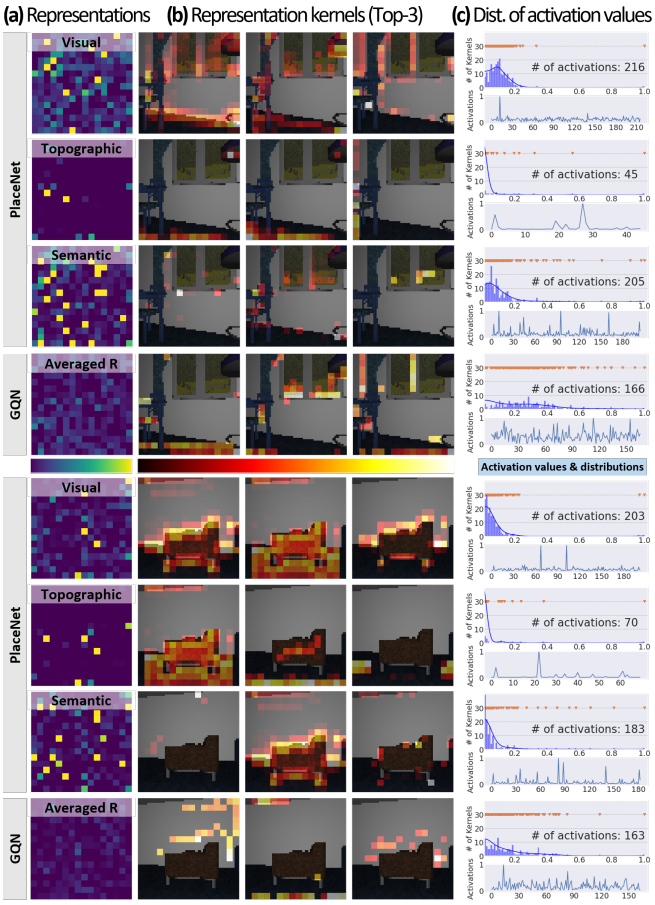

Figure 10: **Kernel activations of the spatial representations.** (a) Representations for each modality indicating averaged kernel activations. (b) Sailency maps of kernel activations. (c) Distributions of the normalized kernel activation values.

While this gives us hope that PlaceNet can potentially be used for down-stream tasks such as localization and navigation, we leave that for future work. The limitation of our method requiring object-aware masks can potentially be tackled by state-of-the-art methods for semantic segmentation [Fu *et al.*, 2019; Poudel *et al.*, 2019]. We will further investigate whether the latent size of PlaceNet is sufficient for encoding the expanded scene complexity at higher resolutions.

Acknowledgments

The authors would like to thank Min Whoo Lee, Christina Baek, Dong-Sig Han, Kibeom Kim, and Chris Hickey for their insightful comments and discussion. This research was partly supported by IITP (2015-0-00310-SW.StarLab/40%, 2019-0-01371-BabyMind/10%, 2021-0-02068-AIHub/10%, 2021-0-01343-GSAI/10%, 2022-0-00166-PICA/10%, 2022-0-00951-LBA/10%) and KIAT (P0006720-ILIAS/10%) grants funded by the Korean government (MSIP, MOTIE).

References

[Bailey and Durrant-Whyte, 2006] Tim Bailey and Hugh Durrant-Whyte. Simultaneous localization and mapping (SLAM): Part II. *IEEE Robotics & Automation Magazine*, 13(3):108–117, 2006.

[Cadena *et al.*, 2016] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

[Chen and Williams, 1993] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 279–288, 1993.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[Eslami *et al.*, 2018] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

[Fu *et al.*, 2019] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3146–3154, 2019.

[Gregor *et al.*, 2015] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. DRAW: a recurrent neural network for image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 37, pages 1462–1471, 2015.

[Gupta *et al.*, 2017] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2616–2625, 2017.

[Herweg and Kahana, 2018] Nora A Herweg and Michael J Kahana. Spatial representations in the human brain. *Frontiers in Human Neuroscience*, 12:297, 2018.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Mildenhall *et al.*, 2020] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020.

[Moser *et al.*, 2008] Edvard I Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31:69–89, 2008.

[O'Keefe, 1976] John O'Keefe. Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, 51(1):78–109, 1976.

[Poudel *et al.*, 2019] Rudra PK Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-SCNN: fast semantic segmentation network. *arXiv preprint arXiv:1902.04502*, 2019.

[Rosenbaum *et al.*, 2018] Dan Rosenbaum, Frederic Besse, Fabio Viola, Danilo J Rezende, and SM Eslami. Learning models for visual 3D localization with implicit mapping. *arXiv preprint arXiv:1807.03149*, 2018.

[Sitzmann *et al.*, 2019] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. DeepVoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2437–2446, 2019.

[Song *et al.*, 2017] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1746–1754, 2017.

[Tewari *et al.*, 2020] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. *Computer Graphics Forum*, 39(2):701–727, 2020.

[Tobin *et al.*, 2019] Joshua Tobin, Wojciech Zaremba, and Pieter Abbeel. Geometry-aware neural rendering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 11559–11569, 2019.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Wu *et al.*, 2018] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with

a realistic and rich 3D environment. *arXiv preprint arXiv:1801.02209*, 2018.

[Xingjian *et al.*, 2015] Shi Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NIPS)*, pages 802–810, 2015.

[Yamamoto *et al.*, 2018] Takashi Yamamoto, Koji Terada, Akiyoshi Ochiai, Fuminori Saito, Yoshiaki Asahara, and Kazuto Murase. Development of the research platform of a domestic mobile manipulator utilized for international competition and field test. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 7675–7682, 2018.

[Zhou *et al.*, 2016] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301. Springer, 2016.

[Zhou *et al.*, 2017] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 633–641, 2017.

[Zhou *et al.*, 2019] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.