# 3D Spatial Analysis Method with First-Person Viewpoint by Deep Convolutional Neural Network with Omnidirectional RGB and Depth Images

*Atsushi Takizawa[1], Airi Furuta[2]*
*[1,2]Osaka City University*
*[1]takizawa@arch.eng.osaka-cu.ac.jp [2]furuta@graphics.arch.eng.osaka-cu.ac.jp*

*The fields of architecture and urban planning widely apply spatial analysis based on images. However, many features can influence the spatial conditions, not all of which can be explicitly defined. In this research, we propose a new deep learning framework for extracting spatial features without explicitly specifying them and use these features for spatial analysis and prediction. As a first step, we establish a deep convolution neural network (DCNN) learning problem with omnidirectional images that include depth images as well as ordinary RGB images. We then use these images as explanatory variables in a game engine to predict a subjects' preference regarding a virtual urban space. DCNNs learn the relationship between the evaluation result and the omnidirectional camera images and we confirm the prediction accuracy of the verification data.*

**Keywords:** *Space evaluation, deep convolutional neural network, omnidirectional image, depth image, Unity, virtual reality*

## INTRODUCTION

The fields of architecture and urban planning widely apply spatial analysis based on images. Examples include the use of the sky-view factor [1] to indicate the amount of sky visible and the green view rate (Takeshi et al. 2013, Yakui et al. 2016) to indicate the amount of green space visible to observers. Generally, when using images to analyze and evaluate spaces, we tend to think that it is necessary to extract meaningful calculable indices from the images. However, many features might influence the spatial conditions, not all of which can be explicitly defined. Even if some can be defined, extracting those features from images is a laborious task. To date, there is no analysis method with which we can quickly evaluate spatial designs and efficiently analyze large image databases, such as Google Street View, 3D interior images by Matterport [2], or the MIT Places Database [3].

Currently, deep learning techniques are attracting much attention for their ability to automatically extract features from raw data. Of these techniques, the deep convolution neural network (DCNN) has expanded the application of deep learning by significantly improving the accuracy of image classification, as compared with conventional methods that use explicit features. In this research, we use DCNNs in a spatial preference evaluation based on images.

General spatial images are considered to be insufficient as spatial analysis data. This is because the spaces themselves are three-dimensional, whereas images are two-dimensional. A photographed scene will change depending on the direction of the camera, and not all areas in the scene can be captured by the camera in one image. However, omnidirectional cameras have become available that can record all the information surrounding the shooting point (so-called first-person viewpoint) and, in this research, we utilize omnidirectional images for image analysis and learning.

Our next consideration is that the features acquired from images tend to be mainly related to color and texture. However, in the world of spatial analysis, methods for analyzing geometric features, such as spatial spread, as typified by the isovist concept (Benedict 1979), have long been used. Understandably, the computer graphics modeling procedure considers spatial characteristics to be influenced by both the geometric features of objects within that space as well as the color and texture features. Therefore, although these features should be handled in a unified manner, they are handled separately due to their different analysis methods. Also, it is difficult to acquire geometric spatial information and then match it with color and image information. Today, laser scanner and infrared sensor technologies are being developed for acquiring spatial depth information. Furthermore, the development of structure-from-motion techniques (Westoby et al. 2012) is advancing, in which 3D surfaces are reconstructed from plural views, so the above difficulty is being addressed. Although these techniques are still in development, the creation of three-dimensional data that once required intensive labor and advanced expertise will become easier as these technologies mature.

In this research, we propose a new framework that uses deep learning to extract spatial features without explicitly specifying them and then uses these features for spatial analysis and prediction. First, we defined a DCNN learning problem with omnidirectional images that include depth images as well as ordinary RGB images. We used these images as explanatory variables in a game engine to predict subjects' preferences regarding a virtual urban space. To do so, we conducted preference evaluation experiments using virtual reality (VR) with multiple subjects, used DCNNs to learn the relation between the evaluation result and the omnidirectional camera images, and confirmed the prediction accuracy of the verification data. If we can successfully solve this problem, we believe that our proposed framework will also be effective in the analysis of more complicated and realistic spaces.

Past studies that have applied deep learning to landscape evaluation include those by Yabuki et al. (2015) and Abe et al. (2016). In contrast to our work here, these studies used pre-trained models with generic images stored in the MIT Place Database rather than omnidirectional images and depth information. Returning to the isovist concept, a 3D isovist was developed in recent years (Derix et al. 2008, Morello and Ratti 2009, Varoudis and Psarra 2014). However, complicated geometric calculations are required to obtain this 3D isovist, so there are associated difficulties with its implementation and slow calculation speed. In this research, we are able to directly acquire depth information images with high speed and accuracy that correspond to a 3D isovist from the z-buffer of a GPU. Furthermore, although Batty (2001) proposed indices for 2D isovists, the shape of a 3D isovist can become complicated, so those features alone may not be sufficient to describe the characteristics and it is difficult to explicitly define other features. In this research, we use a novel approach in which we apply a deep learning automatic extraction method of feature quantities.

The remainder of this paper is organized as follows. In the next section, we propose our evaluation system. Next, we describe our preference evaluation experiments and the DCNNs we use for learning. Finally, we present our results and conclusions.

## SPATIAL EVALUATION SYSTEM BASED ON A GAME ENGINE

In this research, we use the popular Unity [4] game engine as the foundation of our 3D space evaluation system. In addition to its good operability in 3D space, this game engine can efficiently measure various kinds of spatial information through its application programming interface (API) and has improved real time rendering. Therefore, it is reasonable to use this game engine as the basis for our spatial evaluation system, the details of which we describe below.

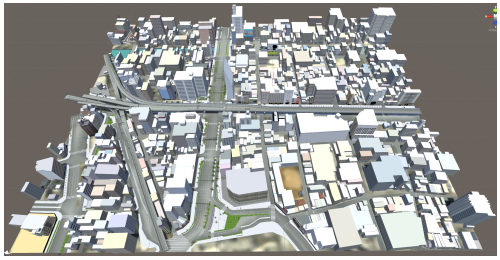


Figure 1
3D model of the study area (Japanese Naniwa City) [5]

### *Spatial data*

Zenrin Co. Ltd., a Japanese map production company, released the ZENRIN City Asset Series [5], which contains the assets of Unity's 3D spatial data. Based on actual measurement data of urban spaces in Japan, this asset comprises lightly arranged textures for games, and can be used as a real spatial model. We used the Namba area of Osaka City (Japanese Naniwa City) as the study space to be evaluated. The area centered in Namba spans a range of about 700 m east-west and about 600 m north-south, as shown in the model in Figure 1.

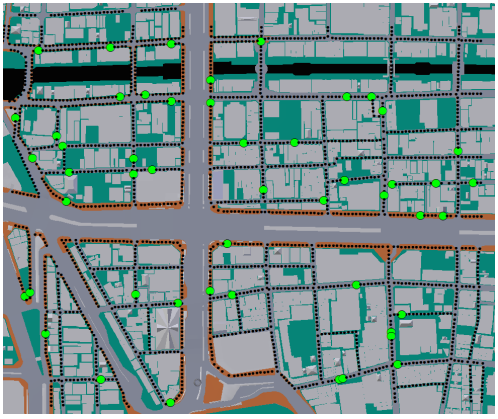
In the preference evaluation experiment, since we asked subjects for their spatial preferences when looking at the city from pedestrian walking areas, we needed to measure the spatial characteristics of each place.

For pedestrian pathways on roadways, we considered the center of the sidewalk (if a sidewalk was present), both sides of the road if there was no sidewalk and the road was wide, and the center of the road if there is no sidewalk and the road was narrow. Then, we set observation points along those pathways at 5-m intervals, for a total of 2,029 observation points (black dots in Figure 2). Since it would be too laborious to conduct a preference evaluation experiment using all these points, we randomly selected 50 points, as illustrated by the green points in Figure 2, and used them as observation points in our preference evaluation experiment.

Figure 2
Candidate observation points (black dots) and those used for evaluation (green dots).

### *Extraction of spatial features*

When extracting the features of each observation point, we considered both the textural and geometric features of the space around each observation point. Using Unity's omnidirectional camera asset, Spherical Image Cam [6], we created a script to capture and store omnidirectional images at each observation point in real time. In addition, we asked the author of this asset to modify it so that we could also take omnidirectional images of depth information from the camera in real time from the z-buffer of a GPU.



Normally, omnidirectional images are outputted as an equirectangular projection map with a height/width aspect ratio of 1:2. Due to the data input limitations in a CNN, when using images for learning and validation, we outputted them as reduced images in a square of 256 × 256 pixels.

Figure 3
Example of omnidirectional images at the same observation point: (upper) RGB image and (lower) D image

We output the depth information as a one-channel grayscale image. When creating a depth image, the original depth information in a z-buffer must be cut by a certain distance value to fit it within the range of the pixel values. In this paper, we omit depth images and call them D images. Then, we created an RGBD image from a pair of omnidirectional RGB and D images taken at the same observation point and angle. Figure 3 shows examples of RGB and D images before being reduced in size to a square.

In the D images, pixels whose distances are closer to the viewpoint are whiter and those farther away are blacker. As we can see from the figure, in an equirectangular projection map, the lateral width of an object is enlarged and distorted from the top or bottom of the image.

### *Construction of the preference evaluation experiment system*

Since we used omnidirectional images as learning data, it was necessary to evaluate subject preferences regarding the whole space visible from each observation point. For this purpose, we used a head-mounted display (HMD) of the Oculus Rift as an interface for conducting our preference evaluation experiment (see Figure 4). If we run a Unity application with an HMD and an ordinary camera instead of an omnidirectional camera, users can experience the immersive feeling of 3D space without any complicated setting. We customized Unity so we could move interactively between the 50 observation points of our preference evaluation experiment and input the evaluation value of each point. Figure 5 shows an example of a screen shot at a certain evaluation time, in which we overlaid the number of the observation point and the evaluation value in virtual space to facilitate our evaluation.

Figure 4
Preference evaluation experiment with VR

### *DCNNs*

Of the several implementation techniques in the deep learning library, we used Chainer [7], which is popular in Japan. In the sample files of Chainer, a Python script train_imagenet.py is available for learning with DCNNs. This is an image classification script for the main 1000 kinds of objects included in the large image database ImageNet. We customized this script to suit our purpose. For our DCNNs, we used AlexNet (Krizhevsky et al. 2012) and GoogLeNet with batch normalization (Szegedy et al. 2015). AlexNet consists of five convolution layers and three layers that are all connected. GoogLeNet comprises much deeper layers. The classification error for the 1000

classes in Imagenet with GoogLeNet is less than that with AlexNet. Since we used depth images from one-channel and RGBD images from four channels as our training data, in addition to the normal 3-channel RGB images, we corrected the input part of the script image data from the original three channels. In addition, as we see later, since we solved the problem of the two-class classification, we reduced the number of output nodes of the final layer from the original one thousand to just two. The padding and stride parameters of the DCNNs are the same as those of the original.

## PREFERENCE EVALUATION EXPERIMENT

In our preference evaluation experiments, our subjects included eight college and graduate students majoring in constructions engineering. The subjects were instructed to make intuitive judgments with respect to four levels: 1 (poor), 2 (slightly poor), 3 (slightly good), and 4 (good). Table 1 shows the evaluation value distribution for each subject in the preference evaluation experiment. Two of the subjects had evaluation values of roughly 2. We averaged all the evaluation values at each point and obtained a maximum average of 3.9 and a minimum of 1.4. Figures 6 and 7 show images of the points with the highest and lowest evaluation values, respectively. Three places received the lowest evaluation values, and we show one of them. When comparing these places, we found them to differ with respect to their geometric features, such as the degree of spatial visibility, and their textural features, such as landscape planting and stone sidewalks. Therefore, as explained in the introduction, we consider the geometric and textural features to be related to spatial evaluation results.

Table 1
Summary of evaluation values for each subject

| | Evaluation value | | | | |
|---|---|---|---|---|---|
| Subject | 1 | 2 | 3 | 4 | Ave. |
| A | 27 | 15 | 8 | 0 | 1.1 |
| B | 17 | 17 | 8 | 8 | 1.2 |
| C | 1 | 22 | 24 | 3 | 2.8 |
| D | 6 | 23 | 13 | 8 | 2.2 |
| E | 7 | 23 | 16 | 4 | 2.5 |
| F | 20 | 20 | 7 | 3 | 1.5 |
| G | 15 | 17 | 10 | 8 | 1.8 |
| H | 13 | 13 | 14 | 10 | 3 |
| Ave. | 13.3 | 18.8 | 12.5 | 5.5 | 2.0 |

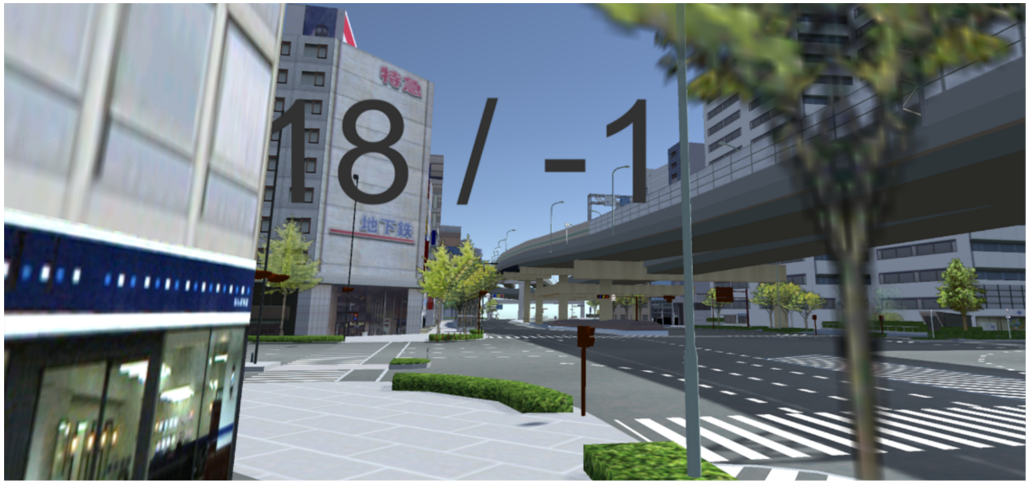



Figure 5
A screen shot presented in our preference evaluation experiment

Figure 6
Images of the point with the best evaluation (Ave. = 3.9)

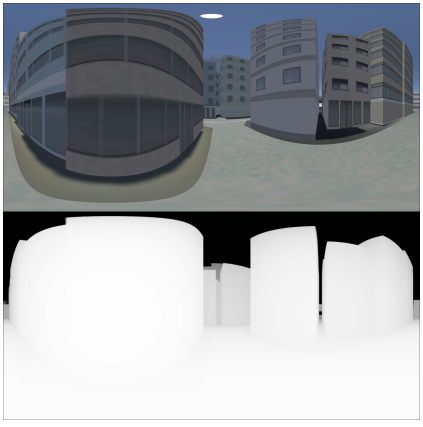


Figure 7
Images of one of the points with the worst evaluation (Ave. = 1.4)

## LEARNING DCNNS

In this section, we described the DCNN learning process with images and evaluation data.

### *Preparation of data*

First, using Unity, we took images to be used for learning from each observation point. When dealing with images in deep learning, data is artificially increased by enlarging and rotating a part of an image, which adds noise. In this study, for each observation point, as shown in Figure 8, we rotated the camera from 0-340 degrees in 20-degree increments with respect to the vertical axis, and generated 18 images for each of the RGB and depth images. As such, for the 50 points, we prepared 900 images for each type of image.

Next, these images are divided for learning and verification data. Since it seems to be difficult to classify to 4 classes because the number of images with an evaluation value of 4 is a little as listed in Table 1, we simplified it to the classification problem with two classes where the evaluation values 1 and 2 belong to Class 0 and the evaluation values 3 and 4 are belong to Class 1. Table 2 lists the number of observation points in the case of two class classification. In consideration of the ratio of each class, the points were randomly divided for each subject so that the data for learning and verification are approximately 3:1 (Table 3). Therefore, the image data set of an observation point appears only in the data set for learning or verification. Through the above procedure, a total of 900 image data were divided into 666 images for learning and 234 images for verification.

Learning was performed for three types of image data: RGB, depth, and RGBD images. We set the mini-batch size for learning to 32. We conducted preliminary learning, and set the number of learning times (number of epochs) to 40 epochs for AlexNet and 20 epochs for GoogLeNet. The learning error of the learning data converged almost to 0 in both cases.

| | Evaluation value: Class | |
|---|---|---|
| Subject | 1 or 2: Class 0 | 3 or 4: Class 1 |
| A | 42 | 8 |
| B | 34 | 16 |
| C | 23 | 27 |
| D | 29 | 21 |
| E | 30 | 20 |
| F | 40 | 10 |
| G | 32 | 18 |
| H | 26 | 24 |

Table 2
Number of observation points for each class

Figure 8
A set of images generated for learning and validation by rotating 20 degrees at the same observation point


### *Results*

Based on the above setting, the DCNN learned using three kinds of image data sets for each subject, and for verification, we applied the data sets to the obtained model to examine their error rates. Tables 4 and 5 list the error rate for each DCNN. Overall, GoogLeNet tended to have a lower error rate than AlexNet.

GoogLeNet is a more recent DCNN than AlexNet and is known to have better classification performance for general images, and our results are consistent with this. Next, we found that the error rate differs greatly depending on the subjects. Comparing the evaluation values listed in Table 1 for subjects F and G, who

had a particularly low error rate, with subjects C and E, who had a large error rate, the subjects with small error rates have low average evaluation values of between 2 and 3.

| | Learning data | | Validation data | |
|---|---|---|---|---|
| Subject | Class 0 | Class 1 | Class 0 | Class 1 |
| A | 31 | 6 | 11 | 2 |
| B | 25 | 12 | 9 | 4 |
| C | 17 | 20 | 6 | 7 |
| D | 22 | 15 | 7 | 6 |
| E | 22 | 15 | 8 | 5 |
| F | 30 | 7 | 10 | 3 |
| G | 24 | 13 | 8 | 5 |
| H | 19 | 18 | 7 | 6 |

Table 3
Number of observation points for each dataset and class

| Subject | RGB | D | RGBD |
|---|---|---|---|
| A | 0.36 | 0.14 | 0.24 |
| B | 0.23 | 0.32 | 0.31 |
| C | 0.48 | 0.53 | 0.42 |
| D | 0.23 | 0.18 | 0.22 |
| E | 0.40 | 0.32 | 0.47 |
| F | 0.14 | 0.13 | 0.17 |
| G | 0.11 | 0.05 | 0.03 |
| H | 0.36 | 0.38 | 0.28 |
| Ave. | 0.29 | 0.26 | 0.27 |

Table 4
Error rate of the validation dataset for each subject by AlexNet

| Subject | RGB | D | RGBD |
|---|---|---|---|
| A | 0.29 | 0.25 | 0.21 |
| B | 0.20 | 0.32 | 0.20 |
| C | 0.45 | 0.43 | 0.46 |
| D | 0.19 | 0.12 | 0.10 |
| E | 0.36 | 0.42 | 0.42 |
| F | 0.04 | 0.12 | 0.01 |
| G | 0.14 | 0.03 | 0.09 |
| H | 0.23 | 0.29 | 0.27 |
| Ave. | 0.24 | 0.25 | 0.22 |

Table 5
Error rate of the validation dataset for each subject by GoogLeNet

As such, when the evaluation value near the preference boundary is large, the error rate is considered to be large. In addition, the error rate of the model for depth images with only one-channel information is the smallest in AlexNet, and even in the case of GoogLeNet, there is not much difference compared with the models of other data sets. This is a significant point-a depth image has less information than an RGB image, but the DCNN has constructed a model with depth images comparable with or superior to the model learned with RGB images. However, the error rate tends to differ for each person. The depth image models of subjects A, D, and G have lower error rates than the RGB-image models. On the other hand, the model learned with RGB images of subject B has a lower error rate than the depth image model. Thus, a DCNN seems too simple to precisely model human spatial cognition, but it is sufficient for considering the different weights of geometric and textural features for the spatial evaluation of each person.

Overall, we found the error rate of the GoogLeNet model using RGBD images to be the lowest. There is a possibility for constructing a spatial evaluation model with higher precision by a DCNN by integrating color and textural information with geometric information.

## DISCUSSION

In this research, we addressed the problem of predicting spatial preference and confirmed that a DCNN can learn from omnidirectional images with some degree of accuracy. Since this method automatically extracts features from complex spaces, we believe that the framework of this research can be widely applied to fields dealing with spatial features. For example, this framework might be used to predict how often actions and events are likely to be triggered by certain spatial features. In the past, we built a model to estimate the location at which urban street crimes are likely to occur (Takizawa 2013). During this research, it took a long time to create detailed street space data, such as the areas of each building's wall openings as viewed from each observation point on the street. The framework used in this research can be expected to efficiently and highly accurately analyze a large amount of 3D spatial information from the first-person viewpoint.

However, further basic research is necessary prior to any full-scale applications. To clarify what kinds of spatial features DCNNs can extract from omnidirectional images, we can use visualization methods such as Grad-CAM [8]. We must also improve model accuracy. In this study, we performed learning and verification using data from only 50 observation points. However, since the sampling data of the study space was more than 2000 points, in principle, more data can be used for learning. However, it is not easy to increase the number of points for preference evaluations due to evaluator fatigue. Therefore, to compress high-dimensional spatial information to a low dimension, we could use an auto encoder, which can perform learning without objective variables, and then use that information as an explanatory variable. Furthermore, since the lateral distortion of shapes in an omnidirectional image increases with greater widths, there is a possibility that the convolution of the square area of DCNNs is not appropriate. Therefore, we plan to develop a new method that can directly convolve a spherical surface, rather than projecting a spherical image onto a plane.

## CONCLUSIONS

In this study, we used a CG model of a street in Osaka to capture omnidirectional images of observation points, including depth images, conducted a VR preference evaluation experiment using 50 street observation points, and learned the data using two types of DCNNs. Our results show that the model error rate was lowest for RGBD images, which suggests the necessity of a new spatial evaluation method that integrates color/texture and geometric features. Since this method automatically extracts features from complex spaces, we believe the framework used in this research can be widely applied to fields dealing with spatial features, although further basic research is necessary prior to full-scale applications.

## ACKNOWLEDGEMENT

This study was partially supported by a Grant-in-Aid for Scientific Research (C) (No.16K06652) and (A) (No. 16H01707).

## REFERENCES

Abe, H, et al. 2016 'Possibility of applying deep learning to the evaluation of street scape', *Proceeding of the annual symposium of Japan Society of Graphic Science*, pp. 85-90

Batty, M 2001, 'Exploring isovist fields: space and shape in architectural and urban morphology', *Environment and Planning B: Planning and Design*, 28(1), pp. 123-150

Benedikt, M 1979, 'To take hold of space: isovists and isovist fields', *Environment and Planning B*, 6, pp. 47-65

Derix, C, et al. 2008 '3d isovists and spatial sensations, two methods and a case study, Evaluating Design Rationale and User Cognition', *Evaluating Design Rationale and User Cognition*, pp. 67-72

Krizhevsky, A, et al. 2012 'ImageNet classification with deep convolutional neural networks', *Advances in Neural Information Processing Systems 25*, pp. 1097-1105

Morello, E and Ratti, C 2009, 'A digital image of the city: 3D isovists in Lynch's urban analysis', *Environment and Planning B: Planning and Design*, 36(5), pp. 837-853

Szegedy, C, et al. 2015 'Going deeper with convolutions', *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9

Takeshi, F, et al. 2013, 'Measurement of the ratio of visible green spaces in the omnidirectional field of vision using cg models and their potential applications', *AIJ Journal of Technology and Design*, 19(43), pp. 1067-1072

Takizawa, A 2013, 'Emerging Pattern Based Street Crime Analysis: Street Level Spatial Analysis of Crime Location Associated with Built Environment in Fushimi Ward, Kyoto City', *Journal of Architecture and Planning (Transactions of AIJ)*, 78(686), pp. 957-967

Varoudis, T and Psarra, S 2014, 'Beyond two dimensions: Architecture through three-dimensional visibility graph analysis', *The Journal of Space Syntax*, 5(1), pp. 91-108

Westoby, MJ, et al. 2012, ''Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications', *Geomorphology*, 179, pp. 300-314

Yabuki, K, et al. 2015 'Basic study on the method of landscape evaluation with deep learning', *Proceeding of the annual symposium of Japan Society of Graphic Science*, pp. 23-28

Yakui, D et al. 2016 'Automatic measurement system of visible greenery ratio using augmented real-

ity', *Proceedings of the 21th International Conference on Computer-Aided Architectural Design Research in Asia*, pp. 703-712
[1] http://www.mlit.go.jp/jutakukentiku/jutaku-kentiku.files/keikan/keikankisei.pdf
[2] https://matterport.com/gsv/
[3] http://places.csail.mit.edu/
[4] https://unity3d.com/
[5] http://www.zenrin.co.jp/product/service/3d/asset/
[6] https://www.assetstore.unity3d.com/jp/#!/content/21979
[7] http://chainer.org/
[8] https://arxiv.org/abs/1610.02391