370

First-Person Viewpoint Type Spatial Analysis Method Based on Deep Learning Integrating Texture, Semantic, and Geometric Spatial Features

HINA KINUGAWA & ATSUSHI TAKIZAWA,

OSAKA CITY UNIVERSITY

ABSTRACT

Images are used in studies evaluating the impressions of architecture and urban spaces. However, given the large effect of subjectivity in spatial evaluations, it is difficult to clarify all pertinent influencing elements. In recent years, because convolutional neural networks (CNNs) have features automatically recognized from images, the number of studies applying CNNs for modelling such subjective evaluations have been increasing. On the other hand, geometric features of the spaces such as openness are also important. The geometric features of the spaces have been studied in the Space Syntax community. Isovist is a fundamental model for expressing the local spatial features. Although Batty proposed various spatial feature quantities of 2D isovist, isovist tends to result in complicated shapes, having the same limitation as an image analysis approach.

In this study, we report on the improvement of our new spatial analysis method using multiple deep learning methods that treats 3D isovist-like geometric information within the framework of image processing. Specifically, we take new omnidirectional images in the Osaka city, and improve the method of estimating a depth map from a real image that approximates a 3D isovist. Next, an impression evaluation experiment on spatiality is conducted with the omnidirectional images. Then, the impression evaluation value is predicted by CNNs using images with different properties as input data, such as an omnidirectional RGB image and a depth map. In the evaluation of spatiality, we will evaluate what combination of images is best, and verify the effectiveness of the depth map.

KEYWORDS

Omnidirectional image, Depth map, Deep learning, Cityscape, Impression of spatiality

1 INTRODUCTION

Studies evaluating landscape impressions of street spaces have been widely conducted. When using spatial image analysis techniques for an impression evaluation, it is generally necessary to extract meaningful computable features from images in advance. Such features include the sky view factor, green luminous rate, and color distribution. However, because subjectivity is greatly related to the impression evaluation of the space, it seems to be difficult to explicitly define all elements affecting such subjectivity. In recent years, convolutional neural networks (CNNs) have been rapidly advancing and have been increasingly applied to the spatial analysis of cities using images because they can automatically learn features from such images. A CNN is a network architecture for deep learning that are used primarily for image data and can learn directly from the data without manual feature extraction. For example, Liu et al. (2017) proposed a method for allowing specialists to directly predict the values of impression evaluation experiments conducted in an urban space of China using Street View images when applying a CNN. In addition, landscape research using CNNs has recently increased (e.g. Sereshinhe et al. 2017, Law et al. 2018).

These studies used images with a standard angle of view rather than omnidirectional images. However, the space spreads in all directions from the viewpoint, and geometric features such as a sense of openness and visibility are also important for a spatial evaluation. Such geometric spatial features have traditionally been studied through a Space Syntax (Hillar and Hanson 1984) analysis. Among them, isovist (Benedikt 1979) is a fundamental and important model for expressing local spatial features. The isovist represents the visible region from the viewpoint, a polygon in two dimensions, and a polyhedron in three dimensions. Batty (2001) proposed various spatial features such as the average distance and area of a two-dimensional isovist. However, isovist tends to have a complicated shape, and an approach that explicitly uses the number of shape features has a limit equal to that of a conventional image analysis.

A depth map is the depth information recorded as the image data in a game, and omnidirectional mapping has an information quantity approximately equivalent to the 3D isovist (Artem et al. 2018). Based on this fact, Furuta and Takizawa (2017) captured a large number of omnidirectional images and their depth maps in real time within a virtual urban space built using a game engine. By applying an image photographed beforehand as the input data, using a CNN, a model was constructed for predicting the evaluation value obtained from a preference research experiment using virtual reality (VR), and the usefulness of the depth map was confirmed. In addition, using a deep learning method called pix2pix (Isora et al. 2017), which is a technique for generating corresponding images from a single input image by learning the relationship between a set of paired images, Kinugawa and Takizawa (2020) estimated the depth map of a real space where an acquisition was difficult to achieve, and carried out the prediction of the preference, including a spherical CNN, for a Street View image. However, Kinugawa and Takizawa's study left some problems such as insufficient consideration of depth scaling methods, low quality of the

generated depth images, spatial preferences that do not always correspond to spatiality, and failure to take into account information other than depth and texture.

Therefore, in the present study, the omnidirectional image of the suburb of Osaka city is photographed at the sidewalk position, and the depth map acquired in the CG image is scaled based on the results. In addition, deep learning techniques called WCT2 (Yoo et al. 2019), which is a technique that brings the style of an image closer to that of a specified image, is used to approximate CG images as real images. In this process, an image processing task called semantic segmentation (SS) is used to divide an image into basic landscape components such as the sky, buildings and vegetations pixel by pixel. From these images, we attempt to generate more accurate depth maps using pix2pixHD (Wang et al. 2018) which is an improved version of pix2pix. In addition, the authors conducted a questionnaire for the examinees, increasing to 40 the number of factors related to 3 spaces at the survey points, using the average value as an input image of multiple channels by combining the RGB, depth map, and SS images, among other factors. They then learned and estimated it as a regression problem using ResNet (He et al. 2016), which is a general CNN, and verified the usefulness of the proposed method.

2 SHOOTING OMNIDIRECTIONAL IMAGES IN REAL SPACE

For the impression evaluation experiment described later, omnidirectional photographs of the streets of Sumiyoshi Ward, Osaka city were taken. Although this is an area where residential areas far from the city centre are predominant, there are various residential areas from high-class residential areas to narrow residential areas. In addition, there are various spatial compositions such as avenues, viaducts, large parks, downtown areas, schools, alleys, and streetcars, and the area was judged to be appropriate when the present impression evaluation experiment.

A Richo THETA Z1(Richo 2019) was used as the omnidirectional camera. Considering the intense contrast that occurs in all directions in an outdoor area, a hand-held high dynamic range mode was used. At the same time, the location of the imaging site was recorded through real-time kinematic (RTK) surveying, and the image was recorded as exchangeable image file format (EXIF) data using a smartphone. A Drogger GNSS DG-PRO1 RWS (BizStaton 2019) was used as the surveying instrument. Higashi-Nada Ward of Kobe City (JP-RJBE 10) was set as a reference station for the RTK survey. The theoretical measurement error reached a maximum of 1.5 cm. The height of the omnidirectional camera was adjusted to 2.05 m above the ground, which is the camera height of Google Street View in Japan.

The photographs were taken on November 5, 9, 18, and 19, 2020, all of which were sunny days. A total of 415 images were captured at appropriate intervals on foot. From the EXIF data of the photographed image, the longitude and latitude information of the captured site was extracted, and the site was mapped using GIS (see Figure 1). To maximize the diversity of the images, we constructed a mathematical programming problem that maximizes the Euclidean distance

between imaging points and selected 200 of the 415 images taken. As a result, the minimum distance between imaging points was 53 m. Finally, the 200 images were reduced to a pixel resolution of 1,024 × 512 for subsequent image processing.

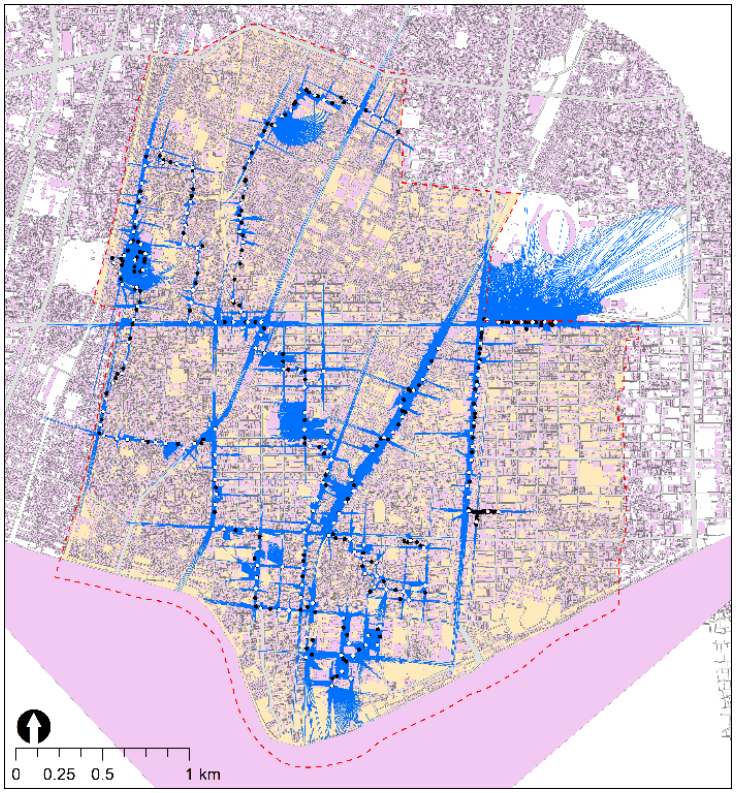


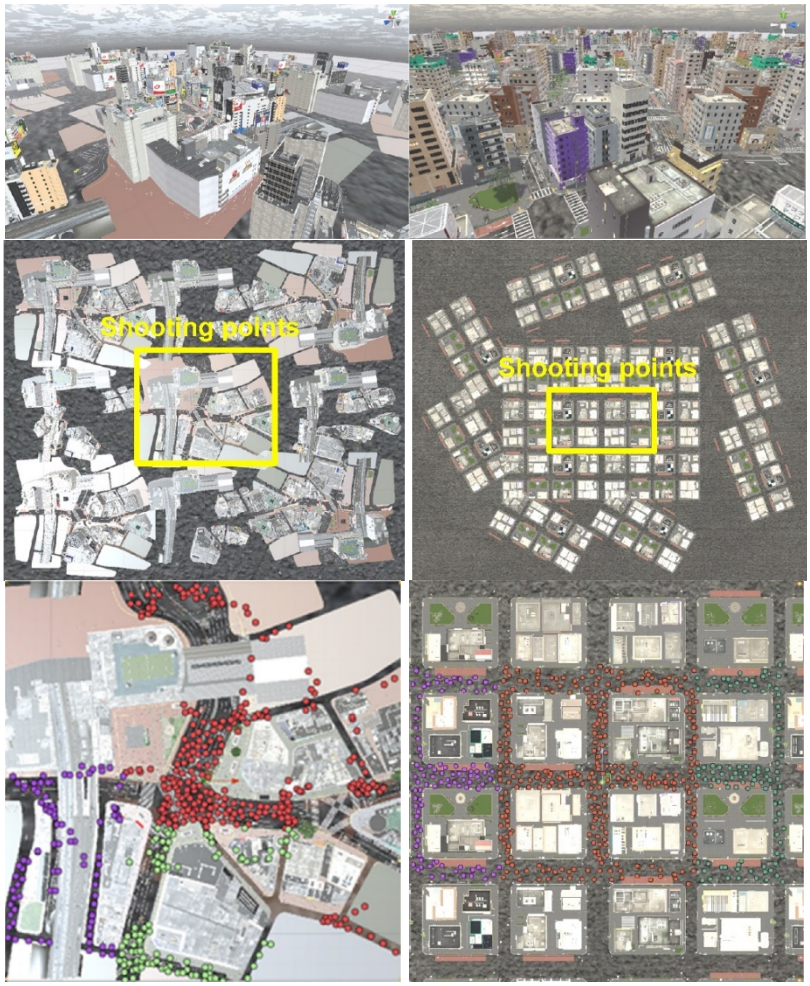
Figure 1: Distribution of shooting points and line of sight of omnidirectional images. The white dots indicate the shooting points, the black dots are unused photographing points, and the blue lines are the lines of sight.

3 LEARNING DEPTH MAPS

In this section, we describe the learning method of omnidirectional depth maps.

3.1 Preparing a virtual city model

A three-dimensional virtual urban space model was prepared for use in learning to generate depth maps from omnidirectional RGB images using pix2pixHD. As in the previous study, two urban models, a Shibuya model that reproduced the busy streets of Japan with a high level of reality, and a local city model, were purchased from a foreign 3D model site (NoneCG 2012 and 2015) and imported into the Unity game engine (see Figure 2). Because the spatial extent of the original model was not sufficiently large to photograph the depth map, each model was copied multiple times to extend the region of the model into the surroundings. Omnidirectional RGB images were also photographed after people and cars were added. Photographs of actual streets vary greatly according to weather, season, time, and other factors. In particular, the sky has a significant influence on the depth map generation, and thus we used the Tenkoku Dynamic Sky asset (Tanuki Degital 2016) to set two sky conditions: clear and cloudy.



(a) Shibuya (b) Local city

Figure 2: Two city models and their locations. The upper images are full view, the middle images show the entire layout, and lower images are enlarged views of the photographed location.

3.2 How to set the depth of the depth map

To generate depth maps of the virtual space, the depth buffer information of a general processor unit (GPU) was used. It is necessary to set the cutoff value of the maximum distance to express this information in an image form. Briefly, the distance from 0 to the cutoff distance is uniformly divided, and the distance is discretized for expression as a pixel value. Hereinafter, we call this method a “linear” scale. Because the amount of information per channel of the general image is only 256 gradations, the resolution of the distance becomes coarse when the distance of the cutoff is lengthened. There is a trade-off between the distance resolution and the extent to which the objects are distinguished. When the 200 photographed sites used in this study were plotted using GIS, a line of sight of 1,000 m in length was generated at 1° intervals in all horizontal directions at each of 185 observation points, except for 15 points that were within the contour polygon (Osaka city 2020) of the building based on the positional error. The line segment crossing of the lines of sight and building polygons was then determined, and the 2D isovist shown in Figure 1 was made. The median length was 21.9 m, and the mean was 49.4 m. As

shown in Figure 3, the length of line of sight was 88% within 100 m, 97% within 250 m, and 99% within 500 m, and it was confirmed that they also followed a lognormal distribution of $LN(3.08,1.29)$. Because the depth map made in this study corresponds to a 3D isovist, the distribution of line-of-sight lengths differs; however, such a distribution in the horizontal direction seems to be important for evaluating the impression of a space. Based on this result, considering the balance between distance resolution and depth, in this study, depth maps were recorded as monochrome images in real time using a GPU with the cutoff distance of the linear model set at 100 m.

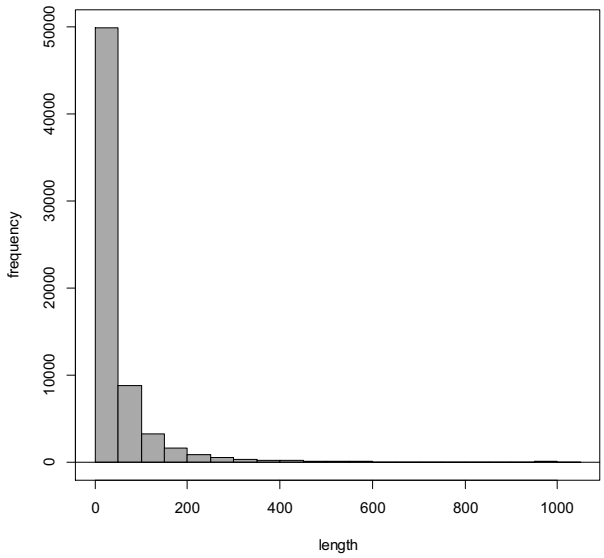


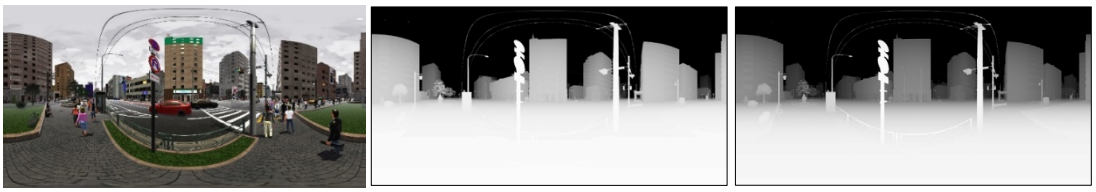
Figure 3: Histogram of the line-of-sight lengths at the photographed points

However, as Weber-Fechner law (Fechner 1860) indicates, many human senses are proportional to the logarithm of the stimulus quantity. In this section, the depth of the space corresponds to the stimulus. For example, the depth of the view in landscape evaluation (Yui et al. 1993) and modelling the depth perception of drivers of automobiles (Yotsutsuji and Kita 2009) use logarithmic transformation of the depth. Therefore, this study also follows the same approach.

Because the length of the line of sight follows a lognormal distribution, a curve is drawn such that the cumulative distribution function of the probability density function is multiplied by 256 vertically to correspond to the pixel value. Next, the range of the vertical axis of this curve is divided into 256 equal sections, and through the cumulative distribution function, it corresponds to the section of the distance of the horizontal axis, and the depth map is made at this corresponding distance. In other words, the interval of the distance is changed such that the appearance probability of the number of the line of sight becomes equal. At this time, the cut-off distance was approximately 625 m. We call this method the “nonlinear” scale.

Figure 4 shows the differences between the two distance scales. In the depth map used in this study, white indicates a close distance, and black indicates a distance greater than or equal to the

cutoff distance. The resolution on the short-distance side is higher in the nonlinear scale, and the building in the back is recognized.



(a) RGB (b) Linear scale (c) Nonlinear scale.

Figure 4: Comparing two distance scales in omnidirectional depth maps

3.3 Photographing omnidirectional images in virtual spaces

Next, omnidirectional images of the virtual spaces were photographed. As shown in Figure 1, an image of a layout drawing of each city model was input into the GIS, and 500 photographed points were randomly set along the roads at the centre of the space. Next, the photographed points were divided into three categories: learning, verification, and testing of pix2pixHD learning. The number of images for each category is 300, 100, and 100, and the images were color-coded red, green, and purple, respectively (see Figure 1). When the distance to the adjacent photographed point is as close as several meters, and the image is randomly divided without considering the position of the point, images that are similar to those in each dataset were mixed in; however, because this might not be a meaningful test, the dataset was divided into the space described above.

The height of the camera was set at 2.05 m, and an omnidirectional image was taken for each photographed point using a “Spherical Image Cam,” which was previously sold as an asset of Unity but is no longer available. These images were taken and recorded on linear and nonlinear depth scales using a pixel resolution of 1,024 × 512 for an equal-area cylindrical projection. The RGB images were photographed with pedestrians and cars in consideration of the actual conditions; however, the depth map was photographed by excluding the models of the pedestrians and cars because we aimed to grasp only the spatial conditions.

3.4 Applying style transformations

The purpose of using pix2pixHD is to input an omnidirectional RGB image of a real cityscape and generate a depth map. Learning is conducted using computer graphic (CG) images of the virtual space. It is therefore expected that the quality of the generated depth map will be higher if the atmosphere of the CG images is closer to the actual images. The style transformation of the CG images was then carried out using WCT2. With WCT2, the actual referenced image is prepared for conversion into a CG image. Semantic segmentation (SS) was applied to both images, and the same style was applied to the same type of objects. In this study, we used DeepLab v3+ (Chen et al. 2018) learned from the Cityscapes dataset (Cordts et al. 2016) as the

SS model. Cityscapes is a large database focused on the semantic understanding of urban cityscapes. The database provides pixel annotations for 30 classes grouped into eight categories (planes, people, vehicles, buildings, objects, nature, sky, and blank).

The 200 photographs described in Section 2 were used as the actual images. WCT2 was applied by selecting one actual image that was the most similar to the components of each CG image. With WCT2, there are several conversion methods to choose from; here, we chose option unpool=cat5 and transfer_at_decoder. Because the color tone of the image with a cloudy sky became unnatural, a style transformation was carried out only for the CG images with clear skies. Figure 5 shows an example of a style transformation. The original RGB image has strong blue and green colors and appears to be CG; however, the style transformation brings the colors closer to those of the real image.

(a) CG image (b) Actual image of the style

(c) Style transferred CG image (d) Operator overlayed image

Figure 5: Style transformation example using WCT2

3.5 Other image processing

In the omnidirectional image photographed in Section 2, the photographer and photographic equipment are in the lower part of the image, but not in the CG image. To match the conditions of the actual images, we selected 10 actual images with different conditions from the 200 images. We then extracted only the parts of the photographer and photographic equipment through a masking process, and randomly selected 1 of the 10 images to be superimposed on each CG image (see Figure 5(d)). After this process, each CG image and the corresponding depth map were rotated by 22.5° to the vertical axis of the camera to obtain the same image as that captured by the camera. The number of rotated omnidirectional images increased to 16 for each location. Thus, we prepared 4,800 training data, 1,600 validation data, and 1,600 test data as paired RGB and depth map images for CG omnidirectional images taken for each city model.

3.6 Learning and accuracy evaluation of pix2pixHD

A dataset of the above image was constructed using eight combinations of city model = {Shibuya, local}, sky = {fine, cloudy}, and depth scale = {linear, nonlinear}, and pix2pixHD was learned. The original pix2pixHD implemented by PyTorch (NVIDIA 2018) was used. The hyperparameters that varied from the default were epoch number = 120 (100 (constant) + 20 (decay)) and batch size = 10. The learning was conducted using a GeForce RTX 3090 GPU. The weight of pix2pixHD was preserved every 10 epochs. To evaluate the accuracy of the depth map generated by pix2pixHD, the error at the pixel level between the generated depth map and the correct depth map was evaluated using root mean square error (RMSE). Let X, Y, and Z be an ordered set of input images, correct images, and noise images, respectively. In addition, n is the number of images, and m is the number of pixels in one image. The RMSE for the generated image $G(x \in X, z \in Z)$ and the corresponding output image $y \in Y$ is expressed as follows:

$$RMSE(x, y, z) = \sqrt{\frac{1}{m}\|y - G(x, y)\|^2}.$$

The average RMSE for all images is expressed by

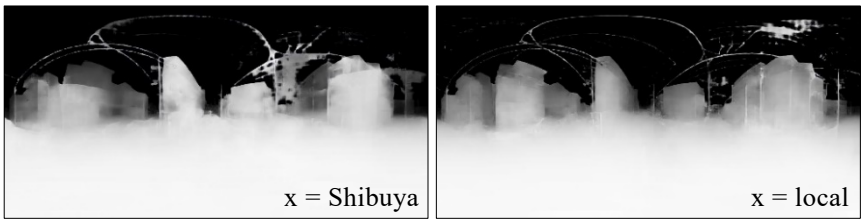
$$\overline{RMSE(X, Y, Z)} = \frac{1}{n} \sum_{x \in X, y \in Y, z \in Z} RMSE(x, y, z).$$

Table 1 lists the results of the average RMSE of the test data obtained using the epoch model, which showed the highest average RMSE in the verification data for each pix2pixHD model. The average RMSE of the model trained using local_fine_linear was the lowest. Figure 6 shows an example of the generation of a depth map using the model. It was concluded that the distance estimated by pix2pixHD has a high generalization for images of the same domain because of the high visual similarity.

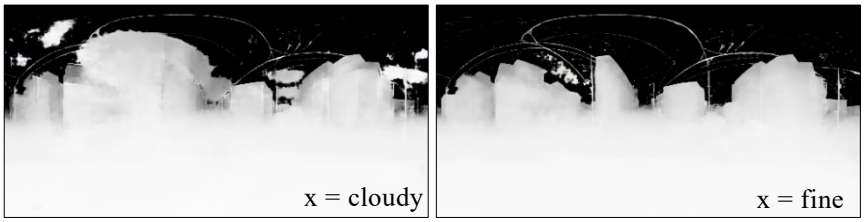
Table 1: Average RMSE of the best model for each dataset

Dataset	Best epoch	Mean RMSE
Shibuya_cloudy_linear	90	4.43
Shibuya_cloudy_nonlinear	110	5.19
Shibuya_fine_linear	120	4.08
Shibuya_fine_nonlinear	120	4.77
local_cloudy_linear	120	4.43
local_cloudy_nonlinear	80	3.95
local_fine_linear	90	**3.92**
local_fine_nonlinear	100	4.62

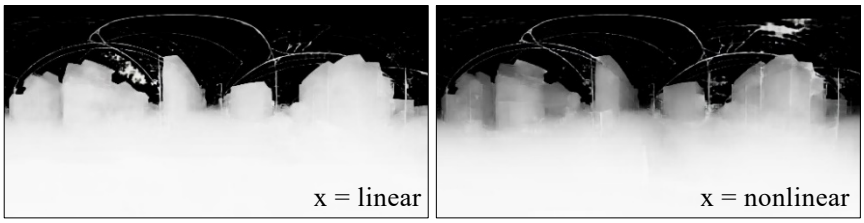
(a) RGB



(b) Difference in city model (x_fine_nonlinear)



(c) Difference in sky condition (local_x_linear)



(d) Difference in depth scale (local_fine_x)

Figure 6: Comparison of depth maps generated by pix2pixHD model with different datasets

4 IMPRESSION EVALUATION EXPERIMENT

The impression evaluation experiment was carried out on-line using Google Photo which can project the omnidirectional image in panorama and Google Form for the answer. Two hundred images photographed in Sumiyoshi Ward of Osaka city were used for the evaluation. In a previous study (Takizawa and Kinugawa 2020), the preference of the place was evaluated in four stages, and it was assumed that the quality of the photograph itself, such as the amount of sunshine, affected the preference, and applying the depth image seemed to be insignificant. Then, referring to the study by Tsumita (1993), three factors related to spaciousness were defined i.e., stereoscopic feeling (planar-stereoscopic), sense of unity (messy-unified), and a sense of space (closed-open), and were evaluated using four grades. The subjects were 20 students majoring in architecture and 20 working people of various occupations. Each image was evaluated by a total of 10 people (5 students and 5 adults), and the subjects were randomly assigned such that the assignment of the subjects was different for all images. Each subject evaluated 50 images.

After the evaluation, for each factor of each image, the mean of the evaluation value of 10 subjects was taken, and the basic statistics on 200 images of the value are listed in Table 2. The median and mean values were within the range of 2.5–2.9, and there were no factors biased to either.

Table 2: Basic statistics of mean evaluation values on 200 images for each factor

Factor	Min	Median	Mean	Max	Std
Stereoscopical feeling	1.7	2.9	2.8	3.7	0.42
Sense of unity	1.5	2.5	2.5	3.9	0.45
Sense of openness	1.3	2.8	2.7	3.9	0.58

Figure 7 shows example images in which the evaluation values of each factor are low, medium, and high. Although an inverse relationship seems to appear between the sense of solidity and the sense of openness, the solidity tends to be affected by the fine ruggedness of the buildings, such as the eaves. However, when there is no sense of unity, a tendency to capture superficial features, such as building advertisements, can be read.

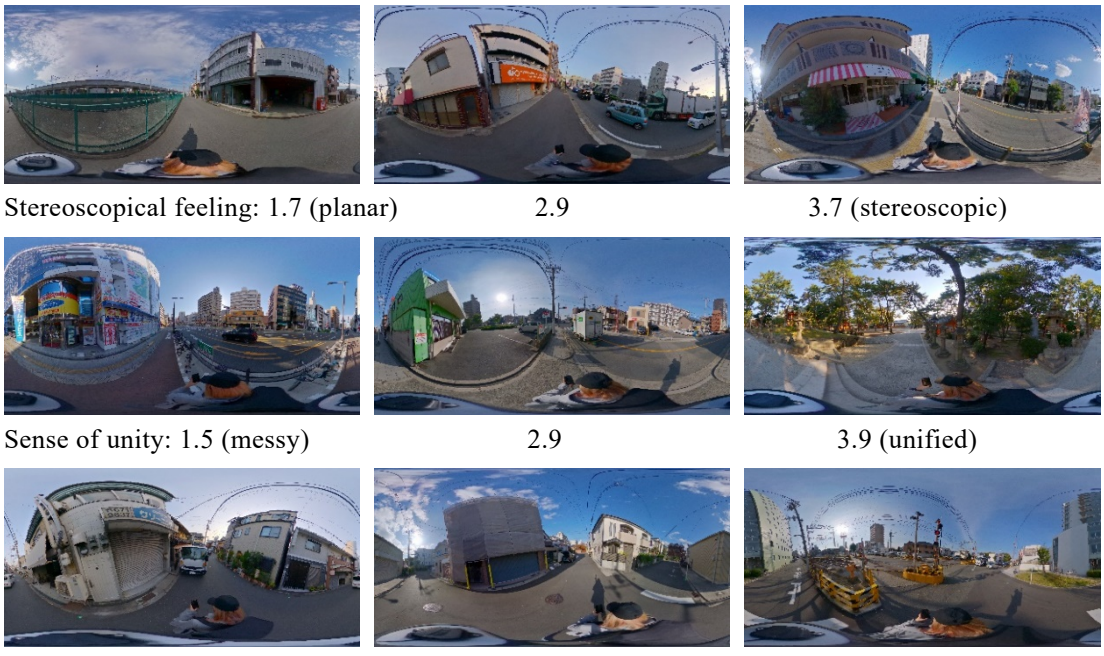


Fig. 7 Examples of images with low (left), medium (centre), and high (right) ratings for each factor

5 LEARNING OF IMPRESSION EVALUATION VALUE ESTIMATION MODEL

Finally, using the results described thus far, we constructed an impression evaluation model for spatiality using a CNN.

5.1 Preparing various image channels

To create the CNN used to learn the model for estimating the impression evaluation value described in the next chapter, an estimation depth map, an SS image, and a grayscale image were prepared on the basis of omnidirectional RGB images of the 200 photographed points described in Section 2. First, the estimated depth map was generated from real images using the learned pix2pixHD model described in Section 3.6. As a result of the improvement, the accuracy of the estimated depth map was relatively good even for the sky part; however, the noise tended to increase when there were many clouds, and we therefore generated as SS image of the real image in the same way as when WCT2 was applied, and then filtered the pixel value of the sky part of the estimated depth map as the maximum distance. In addition, the SS image generated at this time was used as the new input image. The original SS image is output as an RGB image; however, the information is organized to reduce the number of channels to one. Among the 20 spatial components in the CityScapes dataset, only the 11 fixed spatial components listed in Table 3 were used, and the rest were set as “others.” Except for the others, the pixel values were set within the range of [0, 255] in descending order, in order of their appearance from the bottom to the top of the image and were set to integer values with approximately equal intervals. To create a grayscale image, we first removed the gamma correction from the real RGB image, and then converted the image into the CIE XYZ color space (BT. 709) and gamma corrected it again. A grayscale image was then created in the Y-channel.

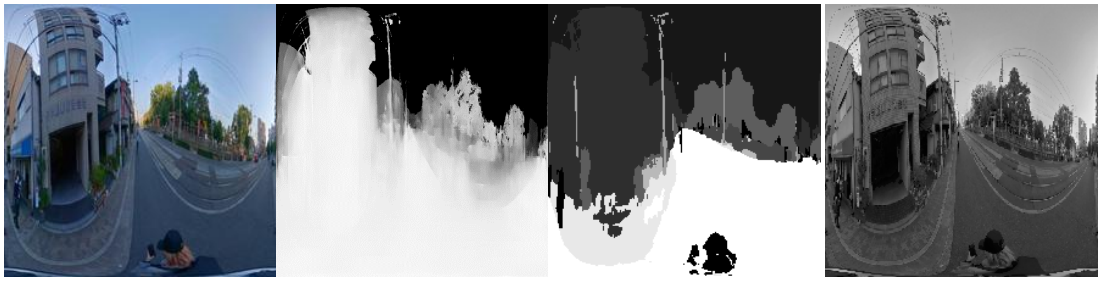
Table 3: Components and pixel values used in 1-ch SS images

Component	Pixel value	Component	Pixel value
Road	255	Traffic light	116
Sidewalk	232	Vegetation	93
Fence	209	Wall	70
Terrain	185	Building	46
Pole	162	Sky	23
Traffic sign	139	Others	0

Although CNNs for omnidirectional images have recently been proposed, in our previous study (Kinugawa and Takizawa, 2020), the estimation accuracy of the impression ratings when applying the CNNs was lower than that with CNNs used for general rectangular images, and thus we also utilized general CNNs in the present study. In our previous study, we resized the omnidirectional image represented by a cylindrical equirectangular projection with an aspect ratio of 2:1 into a square, which is a common input format for a CNN. However, there is a problem with a cylindrical equirectangular projection in which the areas of objects above and below the image become too large. In the case of cityscape images, these areas are mainly occupied by the sky and roads and lack information. In this study, we transformed the images using Tobler’s world in a square (Tobler and Chen 1986), which is a type of cylindrical equal-area projection. An example of a set of transformed input image channels is shown in Figure 8. For reference, the upper image is a simple square-resized RGB image. It can be seen that the objects projected near the eye level occupy more area in the image of the Tobler’s projection.

Finally, each image was resized to a pixel resolution of 224 × 224, rotated by 22.5° along the vertical axis, and inverted left and right to increase the number of images. The total of 32 (images/point) × 200 (points) = 6,400 images were generated.

(a) RGB (simply resized from a cylindrical equirectangular projection)



(b) RGB (c) Depth (d) SS (e) Gray scale

Figure 8: Input image channel set of projections using Tobler's world in square except for (a)

5.2 CNN configuration and learning

Using the prepared images, CNN models were constructed to predict the impression evaluation value of each factor obtained during the experiment. Although there were differences in individual preference, for the impression evaluation value, outliers were deleted to estimate the average impression. For each factor of each image, the lower boundary of the outliers of the evaluation values of the 10 subjects was set to $Q_{1/4} - IQR \times 1.5$, and the upper boundary was set to $Q_{3/4} + IQR \times 1.5$. In addition, the values outside this range were omitted as outliers, and the mean impression evaluation values were calculated. Here, $Q_{1/4}$, $Q_{3/4}$, and IQR are the first quartile, third quartile, and interquartile ranges of the evaluation values of the 10 subjects, respectively.

Standard ResNet-18 and ResNet-50 were used for the CNNs, and the weights pre-trained by ImageNet (Deng et al. 2009) were applied as the initial values in the Torchvision (PyToch 2022) implementation. However, because the number of input image channels varies from one to five, we modified the input layer to match the number of channels. In addition, because this is a regression problem for estimating the impression evaluation values, the loss function was changed to the mean sum of squared error (MSE). The main hyperparameters of ResNet are as follows. A total of 50 epochs were used, the batch size is 16, the momentum SGD was applied as the optimization method, the momentum is 0.9, the initial learning rate is 0.001, and the learning

rate reduction method was diminished to 1/2 for every 1/3 of the total number of epochs. Six image channels, i.e., r, g, b (RGB), d (depth), s (SS), and y (grayscale), were used in the input images, as shown in the combinations listed in Table 4. To begin with, the learning was carried out in a simple ResNet-18, and only the combination in which the accuracy was expected from the test result was learned in ResNet-50. Among the 200 images, 160 were classified as learning data, and the remaining 40 were classified into 20 + 20 points. The roles of the verification data and the test data were exchanged, and the accuracy evaluation was carried out through an approximate 10-fold cross-verification. Each point was divided into its own dataset to be as similar as possible to the variance of the impression evaluation values of the images of the original 200 points.

Table 4: Combination of image channels input into a CNN

CNN	1ch	2ch	3ch	4ch	5ch
ResNet-18	r g b d s	ds	rgb dys	rgbd rgbs	rgbds
ResNet-50		ds	rgb dys	rgbs	rgbds

5.3 Results

The MSE and coefficient of determination (R2) were used to evaluate the accuracy of the final model. Depending on the validation data, the model with the epoch having the smallest value of MSE+(1-R2) for each factor was selected, and its accuracy was evaluated using the test data. The results are listed in Table 5. The channels with “50” at the end of the channel combination are the results of ResNet-50. The d model was the best in terms of the sense of three-dimensionality, the rgbds50 model was the best in terms of the sense of unity, and the dys model was the best in terms of the sense of openness. A plot of the averaged evaluation for all factors is shown in Figure 9. In addition, dys and rgbds50 were the best models. Among all of the best models, a model with explicit depth maps was used. In addition, the dys model was a ResNet model with 34 layers, but achieved an accuracy equal to the 5-channel ResNet model with 50 layers.

From the above results, it can be inferred that CNNs trained on data that explicitly include basic spatial information that humans can recognize, such as depth and segmentation, may be able to model an impression evaluation with a simpler CNN than CNNs trained on general RGB images that do not include such information. This means that a CNN trained on RGB images may be able to model an impression evaluation with a simpler CNN than a CNN trained on non-RGB images. Because CNNs are a black box, visualization methods called CAMs (Selvaraju et al. 2017) are

often used to show the basis for their judgments. Although CAMs can point out specific locations, it is difficult to point out overall factors, such as the atmosphere of a space. In this paper, we propose a new method for evaluating the importance of the input channels in a CNN.

Table 5: Mean accuracy of each impression evaluation value of each model with test data

Channel	Stereoscopical feeling		Sense of unity		Sense of openness	
	MSE	R2	MSE	R2	MSE	R2
r	0.233	0.158	0.228	0.211	0.274	0.412
g	0.231	0.124	0.244	0.193	0.218	0.541
b	0.235	0.138	0.243	0.214	0.208	0.533
d	0.206	**0.198**	0.226	0.238	0.278	0.384
s	0.229	0.130	0.248	0.185	0.215	0.546
ds	0.220	0.150	0.220	0.261	0.196	0.577
rgb	**0.202**	0.183	0.213	0.265	0.213	0.557
rgbd	0.222	0.129	0.220	0.233	0.228	0.515
rgbs	0.223	0.144	0.218	0.246	0.202	0.562
rgbds	0.247	0.106	0.225	0.229	0.198	0.575
dys	0.208	0.174	0.217	0.259	**0.178**	**0.621**
ds50	0.212	0.188	0.227	0.248	0.202	0.557
rgb50	0.208	0.173	0.218	0.260	0.190	0.589
rgbs50	0.229	0.133	0.217	0.258	0.192	0.578
rgbds50	0.217	0.169	**0.202**	**0.311**	0.192	0.596
dys50	0.218	0.152	0.223	0.243	0.189	0.591

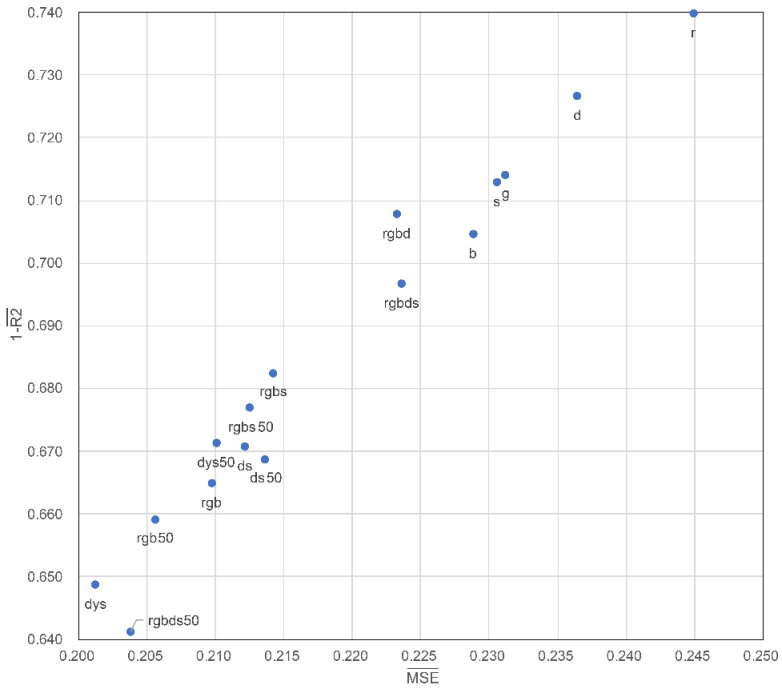


Figure 9: Scatter plot of mean evaluation values of all factors for each model (lower is better)

6 CONCLUSIONS

In this study, we improved a new first-person spatial analysis method based on deep learning, which can handle geometric information such as spatial depth and image features including color and texture within the same framework. For this method, we took omnidirectional images of an urban area in Osaka city and conducted an impression evaluation experiment using factors related to spatiality as the evaluation scale. We then trained a CNN to estimate the impression evaluation values related to spatiality using a combination of processed images such as depth maps and SS images as input data.

As a result, we were able to obtain better quality depth maps from real images than before. In addition, the proposed method was able to train a model with a better generalization performance by using a simple CNN than applying a simple RGB image input. In terms of Space Syntax, the relatively high accuracy of CNNs trained on combinations of depth maps, semantic images, and textures is a major achievement of this study. When CNNs are used for spatial analysis, RGB or SS images at normal viewing angles are often used as input data. Although it could be said in this study that SS may indeed contain important data for spatiality evaluation, it is incomplete by itself, and it may be possible to develop a spatial evaluation model that is closer to human sensation by combining it with omnidirectional depth maps and textures. In the future, it will be necessary to develop such a method to visualize and evaluate the influence of each input channel. However, the absolute accuracy of the proposed method was low, partly because the number of data to be trained was small for deep learning. In the future, it will be necessary to obtain more omnidirectional images and impression evaluation data. In addition, a comparison should be made with spatial analysis methods that make the features explicit.

ACKNOWLEDGEMENT

This work was supported by Grant-in-Aid for Scientific Research C (20K04872).

REFERENCES

Artem C. et al. (2018) ‘Generilized Visibility-Based Design Evaluation Using GPU’, *The 23rd International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA)*, pp. 483-492.

Benedikt M. (1979) ‘To Take Hold of Space: Isovists and Isovist Fields’. *Environment and Planning B*, 6, pp. 47-65. doi: 10.1068/b060047.

Batty M. (2001) ‘Exploring Isovist Fields: Space and Shape in Architectural and Urban Morphology’, *Environment and Planning B: Planning and Design*, 28, pp. 123-150. doi: 10.1068/b2725.

BizStation (2019) DG-PRO1RWS, https://www.bizstation.jp/ja/drogger/dg-pro1rws_index.html

Chen L.C. et al. (2018) ‘Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation’, *The 15th European Conference on Computer Vision (ECCV)*, pp. 833-851.

Cordts M. et al. (2016) 'The Cityscapes Dataset for Semantic Urban Scene Understanding', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213-3223. doi: 10.1109/CVPR.2016.350.

Deng J. et al. (2009) 'ImageNet: A Large-Scale Hierarchical Image Database', *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

Fechner G.T. (1860) Elemente der Psychophysik, Leipzig: Breitkopf & Härtel.

He K. et al. (2016) 'Deep Residual Learning for Image Recognition', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778. doi: 10.1109/CVPR.2016.90.

Hillier B. and Hanson J. (1989) The Social Logic of Space. Cambridge University Press.

Isola P. et al. (2017) 'Image-to-Image Translation with Conditional Adversarial Networks', *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967-5976, doi: 10.1109/CVPR.2017.632.

Law S. et al. (2018) 'Street-Frontage-Net: Urban Image Classification Using Deep Convolutional Neural Networks', *International Journal of Geographical Information Science,* 34(4), pp.681-707. doi:10.1080/13658816.2018.1555832.

Liu L. et al. (2017) 'A Machine Learning-Based Method for the Large-Scale Evaluation of the Qualities of the Urban Environment', *Computers, Environment and Urban Systems*, 65, pp. 113-125. doi: 10.1016/j.compenvurbsys.2017.06.003.

NoneCG (2012) Tokyo Shibuya, https://www.nonecg.com/3D-products/tokyo-shibuya/

NoneCG (2015) Japan - 8 Blocks - 34 Buildings, https://www.nonecg.com/3D-products/japan-8-blocks-34-buildings/.

NVIDIA (2018) pix2pixHD, https://github.com/NVIDIA/pix2pixHD.

Osaka City (2020) Osaka City Topographic Map (Structured Data_ESRI Shapefile), https://www.geospatial.jp/ckan/dataset/r01-esri-shapefile.

PyTorch (2022) torchvision, https://pytorch.org/vision/stable/index.html

Ricoh (2019) THETA Z1, https://thetaz1.com/en/

Selvaraju R. R. et al. (2017) 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization', *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618-626. doi: 10.1109/ICCV.2017.74.

Seresinhe C. et al. (2017) 'Using Deep Learning to Quantify the Beauty of Outdoor Places', *Royal Society Open Science*, 4, 170170. doi:10.1098/rsos.170170.

Takizawa A. and Furuta A. (2017) '3D Spatial Analysis Method with First-Person Viewpoint by Deep Convolutional Neural Network with Omnidirectional RGB and Depth Images', *The 35th Education and research in Computer Aided Architectural Design in Europe (eCAADe)*, pp. 693-702.

Takizawa A. and Kinugawa H. (2020) 'Deep Learning Model to Reconstruct 3D Cityscapes by Generating Depth Maps from Omnidirectional Images and Its Application to Visual Preference Prediction', *Design Science*, 6, E28. doi:10.1017/dsj.2020.27.

Tanuki Digital (2016) Tenkoku Dynamic Sky System, http://www.tanukidigital.com/tenkoku/.

Tobler W. and Chen Z. (1986) 'A Quadtree for Global Information Storage' *Geographical Analysis*, 18(4).

Tumita H. (1993) 'Study of Correlation-Analysis between Space-Consciousness and Physical-Elements on Urban-Open-Spaces', *Journal of Architecture, Planning and Environmental Engineering (Transactions of AIJ)*, 451, pp. 145-154. doi: 10.3130/aijax.451.0_145.

Wang T. et al. (2018) 'High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs', *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8798-8807, doi: 10.1109/CVPR.2018.00917.

Yoo J. et al. (2019) 'Photorealistic Style Transfer via Wavelet Transforms', *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9035-9044. doi: 10.1109/ICCV.2019.00913.

Yotsutsuji H. and Kita H. (2009) 'A Perspective and Review for Modeling the Structure of Driver's Visual Perceptions for Distance and Speed', *IATSS Review*, 34(3), pp.72-79.

Yui M. et al. (1993) 'A Study on the Influence of Artificial Structures on the Natural Landscape with Change of Viewing Distance', *Journal of the Japanese Institute of Landscape Architects*, 56(5), pp.217-222, doi: 10.5632/jila1934.56.5_217.