EVALUATING VISUAL IMPRESSIONS BASED ON GAZE ANALYSIS AND DEEP LEARNING: A CASE STUDY OF ATTRACTIVENESS EVALUATION OF STREETS IN DENSELY BUILT-UP WOODEN RESIDENTIAL AREA

T. Oki^{a,*}, S. Kizawa^a

^aTokyo Institute of Technology, 2-12-1-M1-27 Ookayama, Meguro-ku, Tokyo 152-8550, Japan –
(oki.t.ab, kizawa.s.aa)@m.titech.ac.jp

KEY WORDS: Gaze Analysis, Convolutional Neural Network, Grad-CAM, Densely Built-up Wooden Residential Area, Attractiveness, Google Street View, Questionnaire, Semantic Segmentation

ABSTRACT:

This paper examines the possibility of impression evaluation based on gaze analysis of subjects and deep learning, using an example of evaluating street attractiveness in densely built-up wooden residential areas. Firstly, the relationship between the subjects' gazing tendency and their evaluation of street image attractiveness is analysed by measuring the subjects' gaze with an eye tracker. Next, we construct a model that can estimate an attractiveness evaluation result using convolutional neural networks (CNNs), combined with the method of gradient-weighted class activation mapping (Grad-CAM) - these in visualizing which street components can contribute to evaluating attractiveness. Finally, we discuss the similarity between the subjects' gaze tendencies and activation heatmaps created by Grad-CAM.

1. INTRODUCTION

1.1 Background

Today, many densely built-up wooden residential areas still remain in Japan. In these areas, buildings are aging, and the percentage of vacant houses is increasing. Therefore, improving safety from the perspective of crime and disaster prevention is an urgent issue. On the other hand, people's visual impression of these areas is not necessarily negative. For example, they may feel nostalgia and warmth when they see a group of wooden buildings, or they may feel a sense of comfort from the dense concentration of buildings. It is also known that in the alleys of these areas, "overflowing" (*afuredashi* in Japanese) occurs as self-expression of territorial appeal and personal identity, forming a unique local community. The impression of miscellaneous and intricate alleys may also bring about interest. Thus, there are both positive and negative aspects to densely built-up wooden residential areas, and the characteristics of each area vary. In the future, there is deep concern that the individuality of such townscape will be lost through the promotion of uniform urban development.

In investigating the impressions that people receive from objects, subjective evaluation methods such as the Semantic Differential (SD) method (Osgood, et. al, 1957) are commonly used. For example, Nishio and Ito (2020) compared the impression structure of sequential streetscapes gained from field experiments with that from head-mounted display virtual reality experiments.

In recent years, devices for measuring people's psychophysiological responses (e.g., wearable sensors and smartwatches) have become relatively inexpensive, and their use in architecture and urban planning research has been increasing. There are various examples of measuring eye gaze, heartbeat, perspiration, electroencephalogram, cerebral blood flow, etc., and using them for analysis (Ohno, 2018; Millar, et. al, 2021), but the development of how to utilize them is still in

its infancy. In addition, not only is it becoming relatively easy to obtain big data on architecture and urban images, such as Google Street View, but also, with the development of deep learning methods (Dubey, et. al, 2016; Takizawa and Kinugawa, 2020; Nagata, et. al, 2020), it is becoming common to construct discriminant and regression models using images as input data.

1.2 Research Objective and Process

Although there are many existing excellent researches on subjective impression evaluation methods (e.g., Osgood, et. al, 1957), the analysis results need to be interpreted carefully, and there are cases where interpretation is difficult. It is also noteworthy to acknowledge the difficulty of grasping the subconscious aspects which are not expressed in the subject's answers, which may in turn be influenced by the psychophysiological responses described above. In other words, developing an objective method for evaluating impressions by utilizing rapidly advancing measurement technology would be a significant milestone. In addition, if the relationship between the image and the subject's impression evaluation could be directly learned by a machine using deep learning methods, which are currently making remarkable progress, it is expected that the impression evaluation with general applicability will become possible.

Therefore, the purpose of this paper is to construct an impression evaluation model based on the subject's gaze analysis and deep learning, using streets in densely built-up wooden residential areas as an example. The overview of research process is shown in **Figure 1**. In section 2, we conducted an experiment to obtain the pairs of photos taken in those areas and attractiveness evaluation values by subjects. The pairs were used not only for comparing the results with the tendency of gaze movement in section 3 but also for the grand truth data of the deep learning model in section 4. In section 3, to analyse the relationship between the subjects' gazing tendency and their impression evaluation, we measured the

Figure 1. Overview of research process in this paper.

gaze of subjects evaluating their impressions of street images on a monitor using an eye tracker. In section 4, we constructed an impression evaluation estimation model using convolutional neural network (CNN: Krizhevsky, et. al., 2012), and visualized which components of the street may contribute to the impression evaluation using the method of gradient-weighted class activation mapping (Grad-CAM: Selvaraju, et. al., 2016). Furthermore, the similarity between the obtained class activation map (CAM) and the gazing tendency of subjects is discussed.

1.3 Related Works

Many studies were conducted regarding the evaluation of architectural and urban spaces based on eye tracking. For example, Ohno et al. (2002) conducted an image-presentation experiment, using CG animations and an eye mark recorder to examine the effect of the appearance of scenes from the occluding edge during movement on the formation of impressions of outdoor scenery. In a later study, Ohno (2018) discussed the relationship between the order of appearance of landscape components, gazing tendency, and impression factors. As an example of evaluating streetscapes, Simpson et al. (2019) analysed which pedestrians visually engage with urban street edges and how social and spatial factors impact such engagement. They used a mobile eye-tracking tool and systematically recorded the gaze distribution of 24 study participants as they carried out everyday tasks on differing streets. Besides, Zhang et al. (2018) attempted to integrate virtual reality, 3-D eye-tracking, and protocol analysis to study people's perception of street space characteristics and apply these for re-designing street spaces.

In recent years, highly accurate gaze prediction models have been developed using deep learning, based on the gaze measurement results of subjects (e.g., Deep Gaze II by Kümmerer et. al. (2016), PathGAN by Assens et al. (2018)). Using these approaches, Wang et al. (2019) utilized a Deep Convolutional Neural Network (DCNN) to model legibility in indoor spaces and compared the behaviour pattern and mechanism with human samples. In addition, the above authors used model results to visually explain legibility differences resulting from architectural program, building age, etc.

There has been an increase in the number of studies that use CNNs to automatically extract features from images of architectural and urban spaces and apply them to the construction of various estimation models. For example, Dubey et al. (2016) introduced a large crowdsourced image dataset from 56 cities with six perceptual attributes (safe, lively, boring, wealthy, depressing and beautiful) and trained a CNN model to predict human judgements. Yamada and Ohno (2019) built a highly accurate CNN model to estimate street names and subjects' willingness to visit from street images mainly of tourist spots and downtown areas. Takizawa and Kinugawa (2020) also showed that adding depth information, estimated from a 3D city model to RGB obtained from Google Street View omnidirectional images, improves the accuracy of the model for predicting subjects' impression ratings. In addition, there are some examples of using semantic segmentation (an application of deep learning) to automatically segment landscape images and evaluate urban visual walkability (Zhou et al., 2019). Another example is Oki and Ogawa (2021), who built a CNN model for estimating building structure and construction age from big data of building appearance images included in real estate rental property data, and tried to understand the model structure using Grad-CAM.

In relation to the above previous studies, we believe that the novelty of this research lies in the following points:

- (1) The targets of the analysis are not urban downtown areas or tourist destinations, but general residential areas, and especially densely built-up wooden ones.
- (2) For each of the various impression evaluation indices used in the SD method, the differences in eye distribution and activation heat maps are visualized and compared.

2. ATTRACTIVENESS EVALUATION EXPERIMENT

2.1 Method

The Experiment was conducted in November 2019. First, we extracted road links with a width of less than 4 m from the maintenance area designated by the Tokyo Metropolitan Government (2016), and then obtained the panorama photo IDs of the points closest to the midpoint of each link using Google Street View API. 100 photos (points) were randomly selected

from these photos (**Figure 2**). All of these photos were taken in densely built wooden residential areas.

For each of the 32 subjects (19 males and 13 females, mostly architecture majors in their 20s and 30s), photographs of streets with a width of less than 4 m, which were randomly selected from abovementioned 100 photos, were presented sequentially (10 photographs in total). Then, 18 questions (13 questions (L1) to (L13) about landscape and 5 questions (A1) to (A5) about attractiveness shown in **Table 1**) were prepared for each photo, and the subjects were asked to rate their honest impressions of the photos on a 7-point scale (e.g., +3: extremely bright, +2: quite bright, +1: slightly bright, ±0: neither, -1: slightly dark, -2: quite dark, -3: extremely dark). The order in which the questions were presented was changed for each subject and for each photo.

Figure 2. Target areas for impression evaluation analysis.

Table 1. Questions prepared for each photo.

	Positive value	Negative value		Positive value	Negative value
(L1)	Open	Closed	(L10)	Neat	Messy
(L2)	Friendly	Alienating	(L11)	lived-in feel	no lived-in feel
(L3)	Lively	Quiet	(L12)	Cosy	Not cosy
(L4)	Comfortable	Uncomfortable	(L13)	Clean	Dirty
(L5)	Full of greenery	Lacking greenery	(A1)	Like	Dislike
(L6)	Calm	Restless	(A2)	Interesting	Boring
(L7)	Bright	Dark	(A3)	Desirable for living	Undesirable for living
(L8)	Old-fashioned	Modern	(A4)	Desirable for going through	Undesirable for going through
(L9)	Safe	Insecure	(A5)	Attractive	Unattractive

We measured the gaze of 30 people (30 people x 10 photos = 300 pairs). Therefore, hereafter, we analyse 211 pairs, whose gazing logs were successfully recorded. Ninety-one of the 100 photos prepared in advance were used for these responses. If the purpose of this study was only to evaluate impressions of streets, it would have been possible to conduct a questionnaire with a larger number of photos and subjects. However, since the main purpose of this paper was to measure eye gaze, the number of photos and subjects had to be limited due to the limitation of the eye gaze measurement device used.

These photos are originally 360-degree panoramic images. However, we here asked the subjects to view only in a fixed direction parallel to a street in each panoramic image for simplicity (as shown in **Figure 3**).

Figure 3. Response form for the impression evaluation questionnaire. (1) Photo display position; (2) Respondent ID; (3) Explanation of the response method (Based on the image you have of each alley, please select the part of each item that you think is appropriate. There are 18 questions per photo); (4) Photo No.; (5) Question No.; (6) Answer field; (7) Button (forward). When you press this button, this answer is recorded and the next question is displayed. When you finish answering all the questions, it will automatically switch to the next photo, and when you finish answering all the photos, the exit screen appears.; (8) Button (quit).

2.2 Results

Figure 4 (**Figure 5**) shows the six photos with the highest (lowest) sum of the average ratings for each of the 18 questions (L1 to L13 and A1 to A5). Although there are large individual differences among the subjects in the evaluation values, the images of ID = 44 and ID = 22 (shown in **Figure 4**) each have a high average evaluation of 4 subjects for all 18 questions. In other words, it is possible that many people have a good impression of the streets. On the contrary, the average rating of the three subjects for the ID = 89 (shown in **Figure 5**) image was low for all 18 questions, suggesting that it may be a street that many people have a bad impression of. Based on the results of the 211 pairs of responses, the correlation coefficients between the 18 questions were calculated and a correlation matrix was created (**Table 2**). Overall, the correlation is positive, and some of the correlation coefficients are relatively high. For example, the correlation coefficient between "comfortable (L4)" and "desirable for living (A3)" is 0.69, which can be said to be an index with relatively strong correlation in evaluating the impression of streets.

Figure 4. Results of impression evaluation by subjects (best 6 photos). ©Google 2019

gazing points is discrete (**Figure 7**). Therefore, to obtain the density distribution of gazing time, we adopted a method of smoothing by applying a kernel density function to the gazing points (**Figure 8**), instead of using the coordinates of the gazing points identified in Section 3.2.

3.4 Correspondence between Semantic Segmentation Results and Gaze Time Density

Each pixel of the 100 street images (384×513 pixels) of the densely built-up wooden residential area acquired by Google Street View API was classified into one of the 19 categories using deep learning-based semantic segmentation. The 19 categories are defined in the CityScapes dataset (Cordts et. al., 2016) used to train the DeepLabV3+ model (Chen et. al., 2018) for semantic segmentation as follows: road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, trunk, bus, train, motorcycle, bicycle. **Figure 9** shows the composition ratio of category in 91 photos estimated by semantic segmentation. It can be seen that there is no clear relationship between the total evaluation value and the category composition ratio.

Figure 9. Composition ratio of category in photos estimated by semantic segmentation.

Figure 10. Mapping semantic segmentation results to gaze time density. The cumulative gaze time of the corresponding pixel is read.

Next, we map this to a 100×100 -pixel gazing time density distribution using Equation (1) (**Figure 10**):

$$x_{\text{gaze}} = \frac{w_{\text{gaze}}}{w_{\text{seg}}} x_{\text{seg}} ; y_{\text{gaze}} = \frac{h_{\text{gaze}}}{h_{\text{seg}}} y_{\text{seg}} , \quad (1)$$

where

$x_{\text{gaze}}, y_{\text{gaze}}$ = Image coordinates on gazing time density heatmap.

$x_{\text{seg}}, y_{\text{seg}}$ = Image coordinates on semantic segmentation map.

$h_{\text{gaze}}, w_{\text{gaze}}$ = Height and width of gazing time density heatmap.

(Here, $h_{\text{gaze}} = 100$ [pixel], $w_{\text{gaze}} = 100$ [pixel].)

$h_{\text{seg}}, w_{\text{seg}}$ = Height and width of semantic segmentation map.

(Here, $h_{\text{seg}} = 384$ [pixel], $w_{\text{seg}} = 513$ [pixel].)

The larger the area occupied in the image, the longer the gazing time is calculated. Therefore, the ratio of the gazing time by category is calculated by standardizing the gazing time of each category by the area occupied by each category. Specifically, the gazing time t_i for class i in each image is calculated as Equation (2):

$$t_i = \frac{r_i}{60} \sum_x \sum_y \delta_i(x_{\text{gaze}}, y_{\text{gaze}}) , \quad (2)$$

where

$$\delta_i(x_{\text{gaze}}, y_{\text{gaze}}) = \begin{cases} 1 & \text{when } c(x_{\text{gaze}}, y_{\text{gaze}}) = i \\ 0 & \text{otherwise} \end{cases} . \quad (3)$$

$$r_i = \frac{\sum_x \sum_y \delta_i(x_{\text{seg}}, y_{\text{seg}})}{w_{\text{seg}} \times h_{\text{seg}}} . \quad (4)$$

$c(x, y)$ is a class index of pixel (x, y) .

3.5 Result 1: Gazing Time and Its Breakdown

The average gazing time required to answer each question was 7.2 seconds, and about 80% of the total gazing time was less than 10 seconds. However, looking at the average values, there is no gazing tendency specific to the question or evaluation value (**Figure 11**).

Figure 11. Average of gazing time composition ratio.

3.6 Result 2: Gazing Tendency in Subjects' Responses for the Same Photo Response

Based on the gazing time composition ratios of each of the 19 categories obtained for each response, the positional relationship between the responses can be plotted on a two-dimensional plane by using the multidimensional scaling (MDS). Based on these XY coordinates, the centre-of-gravity coordinates of questions (A1) to (A5) for the same photograph were calculated for each subject.

In **Figure 12**, the distance between each question and the centre-of-gravity coordinates is shown as a stacked bar graph. The longer the bars are, the greater the difference in the gazing time composition between the questions. Even in cases where the same subject responds to the same photo, there are some subjects whose difference in gazing tendency is small depending on the question, while there are others whose difference is large. This can also be confirmed by looking at the actual gazing time density distribution.

Figure 12. Relationship between the distance from the centre of gravity of each response and the gazing time density distribution. The horizontal axis represents the number of valid photos, and the vertical axis represents the total distance from the centre of gravity.

4. IMPRESSION EVALUATION ESTIMATION BASED ON DEEP LEARNING

4.1 Overview of the Impression Evaluation Estimation Model

In this section, we propose a multi-class classification model to estimate the impression evaluation values (L1) to (L13) and (A1) to (A5) (seven classes: integer from -3 to +3) for each image using the same street images as in the subject questionnaire and eye measurement. In other words, the attractiveness evaluation values (as shown in Section 2.2) and the gazing time density heatmaps (as shown in Section 3.4) were used as the ground truth. The CNN is a multilayer neural network that consists of two types of layers, called the convolutional layer and the pooling layer, stacked alternately. In this paper, we use six layers of convolutional layers (Figure 13), referring to the literature (Team Carpo, 2019; Scholle, 2018), due to limited computational resources, despite the low number of classes. We will also create activation heatmaps with Grad-CAM, one of the CNN visualization methods, and try to understand the decision structure of each model based on these heatmaps.

Figure 13. Structure of impression evaluation estimation model based on CNN.

CAM (Class Activation Map) is a method for generating a heat map of class activation from an input image, which also enables to visualize what a CNN has learned. While deep learning models are essentially black boxes, CAM can help us understand which parts of the image were decisive for the final classification of the CNN. In particular, Grad-CAM is a method that uses the gradient of the classes when weighting the output

feature map of the convolutional layer and the feature map based on the input image.

4.2 Training Results of Impression Estimation Models

For the input data, totally 320 pairs (= 32 subjects x 10 photos) of photo and evaluation for each question were used. After dividing the data into 60% training data, 20% validation data, and 20% test data, the impression evaluation estimation model was trained (loss function: categorical cross entropy, optimizer: RMSProp, number of epochs: 300, learning rate: 0.0003 to 0.001). The estimation accuracy of applying the training model to the test data is shown in Table 3.

Table 3. Accuracy of estimation by impression evaluation model based on deep learning. When calculating the fit rate in this table, the class with zero predicted total is included, and when calculating the reproduction rate, the class with zero correct total is excluded and the average value is calculated.

	L1	L2	L3	L4	L5	L6	L7	L8	L9
Precision (mean)	28.9%	28.7%	42.2%	30.0%	43.0%	46.9%	27.4%	42.5%	16.8%
Recall (mean)	25.1%	20.2%	22.3%	19.3%	24.8%	16.7%	21.0%	15.5%	17.5%
F-measure	26.9%	23.7%	29.2%	23.5%	31.5%	24.6%	23.8%	22.7%	17.1%
Total accuracy	37.5%	36.9%	39.4%	38.5%	32.3%	46.9%	36.9%	35.4%	43.1%
% of major response	28.1%	38.5%	36.4%	35.4%	27.7%	46.9%	30.8%	35.4%	36.9%
Major response No.	+1	+1	-1	+1	+1	+1	+1	+1	+1

	L10	L11	L12	L13	A1	A2	A3	A4	A5
Precision (mean)	31.1%	18.2%	45.0%	56.3%	25.6%	22.6%	27.2%	31.9%	24.5%
Recall (mean)	23.9%	16.8%	25.5%	23.7%	24.5%	14.1%	25.5%	22.5%	31.3%
F-measure	27.0%	17.5%	32.6%	33.4%	25.1%	17.4%	26.3%	26.4%	27.5%
Total accuracy	41.5%	40.0%	43.8%	44.6%	38.5%	29.2%	35.4%	30.8%	33.8%
% of major response	32.3%	38.5%	32.8%	32.3%	27.7%	30.8%	30.8%	30.8%	30.8%
Major response No.	+1	+1	+1	-1	+1	-1	-1	+1	-1

The F value, which is the harmonic mean of the fit rate and the reproduction rate, is generally around 30%, and the overall accuracy is around 40%, so the accuracy of the model is not yet sufficient.

On the other hand, when we look at the activation heat map visualized by Grad-CAM (Figure 14), we can see some interesting results which may indicate that the features of the images are captured by CNN. For example, in the column A3 of Figure 14, the components of the buildings along the street

Figure 14. Example of visualization of activation heat map by Grad-CAM. Each image corresponds to the photo/item combination illustrated in Figure 12.

Figure 15. Example of activation heat map on landscape evaluation.

tend to be emphasized, while in the column A4 of **Figure 14**, the difference in gradient between the road and other parts of the street is relatively clear.

In addition, the activation heat map for landscape evaluation also shows some cases of successful feature extraction (**Figure 15**). For example:

- (1) The activation heat map generated from the model that evaluates the impression of "friendly / alienating (L2)" shows features on the walls of roadside buildings and road surface areas.
- (2) Viewing the activation heat map generated from the model that evaluates the impression of "comfortable / uncomfortable (L4)", we can see that many images have sky extracted. This result suggests that it may be related to sky coverage.
- (3) The activation heat map generated from the model that evaluates the impression of "full of greenery / lacking greenery (L5)" shows that the trees and plantings are extracted.
- (4) The activation heatmap generated from the model that evaluates the impression of "clean / dirty (L13)" shows that building components and overflow are extracted.

5. DISCUSSION

In section 2, we quantitatively showed the relationships among the 18 impression evaluation items that characterize streets in densely built-up wooden residential areas. Based on the aggregated results of the evaluation values by the subjects, we confirmed that an overall good (or bad) impression of the cityscape was consistent with our intuition. However, there were many images with significant individual differences in the evaluation values. For improving the statistical reliability of the results obtained, it would be necessary to conduct experiments with a larger number of images, subjects, and evaluation items by utilizing crowdsourcing services. Regarding the subjects' attributes, more diversity (in terms of age, gender, and area of residence) is also necessary. Further study is also needed to present the photos to make them more like the actual spatial experience. For example, VR head-mounted display is a possible method, but it is not easy to collect a large amount of data.

In section 3, we established a method: (1) to measure subjects' gaze during impression evaluation using a screen-based gazing device; and (2) to obtain the gazing time density distribution and the gazing time by street component. As a result, we could not find a clear relationship between impression evaluation and gazing tendency. Since the experiment was conducted using a normal-sized display, the subjects' eye movements were not very large, which may have affected the results. For example, it is worth examining how the results would differ if street photos are projected on a larger screen or if a similar experiment is conducted using a glass-type

eye tracking device in actual streets. In addition to gazing time density, it is also possible to use various other indicators such as gaze trajectory, eye movement speed, and pupil diameter.

In section 4, we developed a discriminant model using CNN to estimate the impression evaluation values of subjects using street images as input. The model's accuracy was not very high. This can be attributed to a few factors: the amount of training data is small, the data of correct answers are biased toward answers with specific evaluation values (+1 and -1), and there are large individual differences in evaluation even for the same image and question. However, we were able to verify the effectiveness of the model to some extent by visualizing the impression evaluation structure of the streets by CNN using Grad-CAM.

Here, we compared the gazing time density distribution (**Figure 12**) with the activation heat map (**Figure 14**), but we can see that there is not much similarity between them. In more detail, we compared all images used as test data on a pixel-by-pixel basis, and found that the maximum cosine similarity was about 0.70, and the average was 0.24. This result quantitatively supports the possibility that the gazing points and evaluation structure of impression evaluation are different between humans (subjects) and AI (artificial intelligence).

From the results of the series of analyses, we could not find a clear relationship between the subjects' gazing tendency and their impression evaluation, and we could not construct a highly accurate impression evaluation estimation model. One of the reasons for this is the influence of individual differences in impression evaluation. The influence of individual differences is not only seen in the evaluation values obtained directly from the questionnaire, but also in the differences in gazing tendency among subjects, photos, and questions. On the other hand, although the deep learning-based impression evaluation estimation model constructed in this paper use training data that includes individual differences, due to its structure, it is not a model that can take individual differences into account.

6. SUMMARY AND CONCLUSIONS

In this paper, using the evaluation of street attractiveness in a dense wooden housing area as an example, we first analysed the relationship between the classification-specific composition ratio and density distribution of gazing time obtained from subjects' eye measurements, the evaluation items and their evaluation values. Next, we constructed a deep learning model to estimate subjects' attractiveness ratings from street images, evaluated its estimation accuracy, visualized the model's decision structure using Grad-CAM, and discussed its characteristics.

In the future, we will continue to pursue the possibility of impression evaluation based on gaze measurement and deep learning by collecting more subject data and investigating how to account for individual differences.

ACKNOWLEDGEMENTS

This research was supported by the 2020 Tokyo Tech Research Award. We would like to express our gratitude to those who cooperated in the attractiveness evaluation experiment and gaze measurement. In addition, many valuable comments from anonymous reviewers led to the improvement of this paper.

REFERENCES

- Assens, M., Giro-i-Nieto, X., McGuinness, K., O'Connor, N. E., 2018: PathGAN: Visual scanpath prediction with generative adversarial networks, arXiv:1809.00567.
- Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam H., 2018: Encoder-decoder with atrous separable convolution for semantic image segmentation, arXiv:1802.02611.
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016: The Cityscapes dataset for semantic urban scene understanding, arXiv:1604.01685.
- Dubey, A., Naik, N., Parikh, D., Raskar, R. Hidalgo, C. A., 2016: Deep learning the city: Quantifying urban perception at a global scale, arXiv:1608.01769.
- François, C., 2018: Deep Learning with Python and Keras, translated by Queap Inc. and supervised by Sugome, Y., Mynavi Publishing.
- Fukuda, R., Sakuma, M., Nakamura, E., Fukuda, T., 1996: An experimental study on the definition of gazing point, *Ergonomics*, 32(4), 197-204 (in Japanese).
- Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012: ImageNet classification with deep convolutional neural networks, *Advances in neural information processing systems*, 25(2), 1097-1105.
- Kümmerer, M., Wallis, T. S. A., Bethge, M., 2016: DeepGaze II: Reading fixations from deep features trained on object recognition, arXiv:1610.01563.
- Millar, G. C., Mitas, O., Boode, W., Hoeke, L., Kruijff, J., Petrasova, A., Mitasova, H., 2021: Space-time analytics of human physiology for urban planning, *Computers, Environment and Urban Systems*, 85, 1-11.
- Nagata, S., Nakaya, T., Hanibuchi, T., Amagasa, S., Kikuchi, H., Inoue, S., 2020: Objective scoring of streetscape walkability related to leisure walking: Statistical modeling approach with semantic segmentation of Google Street View images, *Health & Place*, 66, 1-11.
- Nishio, S., Ito, F., 2020: Statistical validation of utility of head-mounted display projection-based experimental impression evaluation for sequential streetscapes, *Environment and Planning B: Urban Analytics and City Science*, 47(7), 1167-1183.
- Ohno, R., 2018: Studies on environmental perception during locomotion—a review of empirical studies by the Ohno laboratory, *Japan Architectural Review*, 1(2), 194-206.
- Ohno, R., Udagawa, A., Soeda, M., 2002: Effects of emerging scenes from the occluding edges on visual attention and evaluation of the landscape, *Journal of Architecture and Planning (Transactions of AIJ)*, 67(556), 197-203.
- Oki, T., Ogawa, Y., 2021: Model for estimation of building structure and built-year using building facade images and attributes obtained from real estate database, *Journal of Architecture and Planning (Transactions of AIJ)*, 86(779), 357-367.
- Osgood, C. E., Suci, G. J., Tannenbaum, P. H., 1957: Measurement of meaning, Urbana, University of Illinois Press.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2016: Grad-CAM: Visual explanations from deep networks via gradient-based localization, arXiv:1610.02391.
- Simpson, J., Freeth, M., Simpson, K. J., Thwaites, K., 2019: Visual engagement with urban street edges: insights using mobile eye-tracking, *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 12(3), 259-278.
- Takizawa, A., 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.
- Team Carpo, 2019: Deep learning theory and implementation with just the math you need, Shuwa System.
- Tobii, 2016: Tobii Studio user's manual version 3.4.5. <https://www.tobii.com/siteassets/tobii-pro/user-manuals/tobii-pro-studio-user-manual.pdf> (25 March 2020)
- Tobii Technology, 2020: Tobii Pro nano product introduction page. <https://www.tobii.com/ja/product-listing/tobii-pro-nano/>. (24 March 2020)
- Tokyo Metropolitan Government, 2016: Disaster mitigation urban development promotion plan (revised), <http://www.toshiseibi.metro.tokyo.jp/bosai/bosai4.htm>. (4 February 2021)
- Wang, Z., Liang, Q., Duarte, F., Zhang F., Charron, L., Johnsen, L., Cai, B. Ratti, C., 2019: Quantifying legibility of indoor spaces using deep convolutional neural networks: Case studies in train stations, *Building and Environment*, 160, 1-15.
- Yamada, M., Fukuda, T., 1986: Definition of gazing points in images and its application to image analysis, *Transactions of the Institute of Electrical Communication Engineers of Japan D*, 69(9), 1335-1342 (in Japanese).
- Yamada, S. and Ohno, K. 2019: Development and verification of the impression deduction model for city landscape with deep learning, *Journal of Architecture and Planning (Transactions of AIJ)*, 84(759), 1323-1331.
- Zhang, L., Jeng, T., Zhang, R., 2018: Integration of virtual reality, 3-D eye-tracking, and protocol analysis for re-designing street space, Proceedings of the 23rd International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), 1, 431-440.
- Zhou, H., He, S., Cai, Y., Wang, M., Su, S., 2019: Social inequalities in neighborhood visual walkability: Using street view imagery and deep learning technologies to facilitate healthy city planning, *Sustainable Cities and Society*, 50.