Subjective Functionality and Comfort Prediction for Apartment Floor Plans and Its Application to Intuitive Online Property Search

Taro Narahara, *Member, IEEE,* and Toshihiko Yamasaki, *Member, IEEE,*

***Abstract*—This paper presents a new user experience for online apartment search using functionality and comfort as query items. Specifically, it has three technical contributions. First, we present a new dataset on the perceived functionality and comfort scores of residential floor plans using nine question statements about the level of comfort, openness, privacy, etc. Second, we propose an algorithm to predict the scores from the floor plan images. Lastly, we implement a new apartment search system and conduct a large-scale usability study using crowdsourcing. The experimental results show that our apartment search system can provide a better user experience. To the best of our knowledge, this is the first work to propose a highly accurate machine learning model for predicting the subjective functionality and comfort of apartments.**

***Index Terms*—Real estate floor plans, crowdsourcing, graph analysis, attractiveness prediction**

I. Introduction

In recent years, the real estate industry has been showing increasing interest in applying machine learning-assisted solutions such as price prediction [1]–[4] and apartment-searching [5] tools. Some online platforms can help users search for properties by specifying metadata, such as the type of apartment and room size. However, many users inspect floor plans based on more intuitive sensory impressions, such as *living comfort*, *openness*, and *privacy*. This makes it difficult to estimate the perceptive values of apartments through any currently available retrieval system, as there are no quantifiable data that represent such subjective characteristics of apartments. Moreover, apartment properties listed with the same size and type in their metadata (e.g., two-bedroom apartments) could feature different room arrangements, which will have a significant impact on their functionality and overall livability. Therefore, further understanding the relationships between floor plan images and structured data, including the connectivity of rooms and metadata, could improve the user experience on real estate search platforms.

Information on floor plans has been widely adopted by users to evaluate the values of properties over the years, as can be seen in many real estate portal sites today. A floor plan image of an apartment includes various room types, room sizes, and connections and spatial layouts of the rooms. Following a customer survey conducted in Japan[1], a floor plan was found to be among the top five priorities for customers during their apartment search. Moreover, it was found that customers are very reluctant to compromise on their preferred floor plans and are often willing to invest more for their pursuits. Essential elements that largely influence functional, environmental, and some perceptive characteristics of apartments, such as locations of walls, columns, windows, and wet areas, are already set in the floor plans, and cannot be changed no matter the finish materials or furniture used. Although we can estimate the subjective quality of apartments, such as *living comfort*, simply by looking at the floor plan images, no related work nor dataset has been reported for such a task.

In this study, therefore, we first constructed a new dataset that contains $1,000$ floor plan images (hereafter, “dataset A”). Each image has a subjective score from nine perspectives relating to perceived quality and functionality of the apartments. Examples of the generated dataset are shown in Fig. 1. Based on this dataset, we developed a multimodal neural network-based framework to predict subjective apartment scores via their floor plan images, graph representations, and various metadata. The experimental results showed that we can predict the scores with a correlation coefficient of 0.701 on average. This is relatively high, considering that they are all subjective scores. Our study shows that the baseline model, which uses features based on images alone, has a much lower average correlation coefficient of 0.491, even using ResNet50, the state-of-the-art network for image recognition tasks.

The contributions of this paper can be summarized as follows.

- We created and analyzed a large-scale dataset of subjective evaluations of both perceived quality and functionality of real estate floor plan images using crowdsourcing.

Manuscript received March 15, 2022; revised July 3, 2022; accepted September 25, 2022. Date of publication XXXX; date of current version October 8, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abderrahim Benslimane. (Corresponding author: Taro Narahara.)

Taro Narahara is with the Hillier College of Architecture and Design (HCAD), New Jersey Institute of Technology (NJIT), Newark, New Jersey 07102, USA (e-mail: narahara@njit.edu).

Toshihiko Yamasaki is with the Department of Information and Communication Engineering and Research Institute for an Inclusive Society through Engineering (RIISE), University of Tokyo, Bunkyo-ku, Tokyo 113-8654, Japan (e-mail: yamasaki@cvm.t.u-tokyo.ac.jp).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes one MP4 video file demonstrating the proposed search tool. This material is 68 MB in size.

[1] https://suumo.jp/article/oyakudachi/oyaku/chintai/fr_data/hikkoshi-sumikae2017/, accessed 09/16/2020

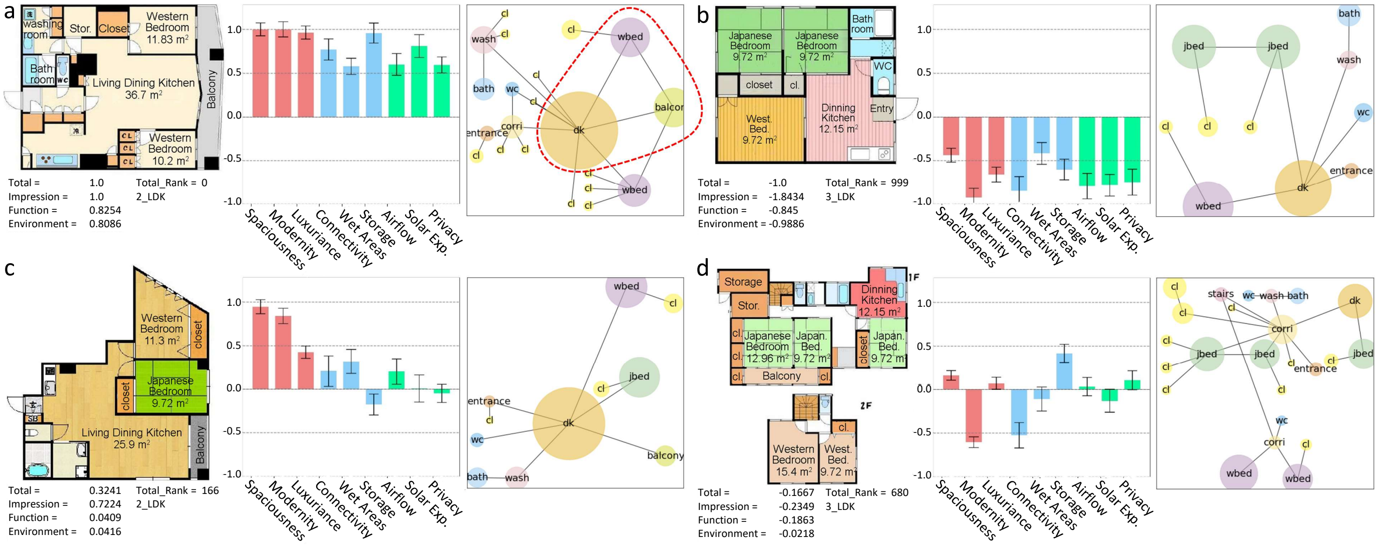


Fig. 1. Examples of our generated functionality evaluation dataset using real estate floor plan images. Each example shows a floor plan image, a bar graph for nine evaluation measures, and a graph from left to right. (a) Highest scoring floor plan; (b) lowest scoring floor plan; (c) floor plan that scored high on modernity; (d) floor plan that scored low on modernity.

(i.e., 3,128 participants rated 1,000 floor plans based on nine subjective criteria.)

- Our proposed prediction model, which extracts features from floor plan images and their graph structures, proved to be effective and highly accurate with average correlation coefficient of 0.701. We also developed a workflow to extract semantically segmented floor plan images, graphs, and graph-related features for our multimodal deep neural network model.
- Upon using a new set of floor plan images (with predicted scores by our model as a dataset), our proposed apartment search tool was found to provide a significantly better user experience than the baseline tool without the proposed feature.

This paper is organized as follows. In Section 2, we report the results of our literature survey. Section 3 explains dataset creation. Section 4 describes the methodology utilized in the study. Section 5 explains the experiments and their results. Section 6 describes the usability study of our proposed apartment search tool. Section 7 discusses the limitations of our approach, with concluding remarks included in Section 8.

II. Related Works

A. Real Estate Tasks using Property Images

Several researchers have worked on tasks related to real estate using property images. In [2], the authors tried to improve the accuracy of real estate price prediction by predicting the luxury levels of the rooms using the appearance and interior images of properties. Law et al. [6] showed that street view and satellite images are also helpful when predicting house prices. In [7], the researchers attempted to predict the construction age of the property by combining the predictions for each of its salient image patch, resulting in greater accuracy than human prediction.

Moreover, deep learning has been applied to real estate property images, and some studies have analyzed real estate images themselves. In [8], real estate images were classified into different types (e.g., bedroom, kitchen, living, and garden) by employing contrast-limited adaptive histogram equalization (CLAHE) and applying long short-term memory (LSTM) in both vertical and horizontal directions.

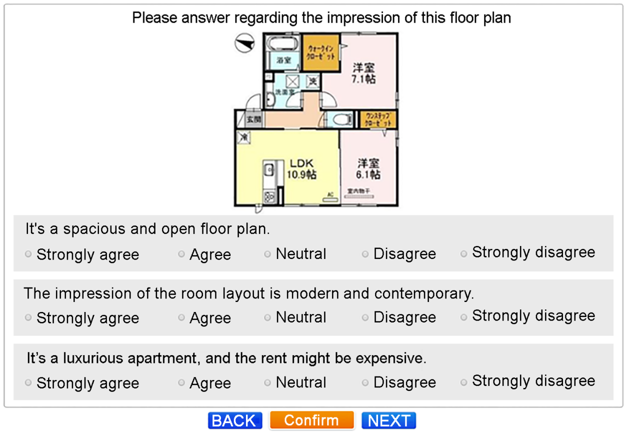
Wang et al. [9] predicted which among two images of the same property is more attractive using a pairwise comparison network.

These works demonstrated the use of images in specific tasks that are related to the appearance of the property. In our task, to comprehensively represent the user experience in term of the perceived quality and functionality, we assumed that floor plan information with related metadata can provide more structure and detailed information on the property.

B. Real Estate Tasks using Floor Plan

Several related works have been conducted on floor plan image analysis. Before the development of deep learning, some researchers manually and graphically analyzed floor plans using adjacency graphs (with rooms, corridors, and other features as labeled nodes) and used them to classify apartments into different types [10]. Takizawa et al. [11] analyzed the relationships among adjacency graph structures of apartment floor plans and their rental fees. They extracted subgraphs from the adjacency graphs and effectively estimated the rent from the presence and absence of common subgraphs. However, the cost of creating adjacency graphs by hand was very high.

Floor plan images have also proved to be useful for rent price prediction [12]–[14]. In [12], [13], it was demonstrated that conventional bag-of-features (BoFs) [15] has the potential to achieve lower-error prediction with smaller variance, although the contribution of BoFs was smaller than that of other apartment attribute information. In [14], image features


Please answer regarding the impression of this floor plan

It's a spacious and open floor plan.
○ Strongly agree ○ Agree ○ Neutral ○ Disagree ○ Strongly disagree

The impression of the room layout is modern and contemporary.
○ Strongly agree ○ Agree ○ Neutral ○ Disagree ○ Strongly disagree

It's a luxurious apartment, and the rent might be expensive.
○ Strongly agree ○ Agree ○ Neutral ○ Disagree ○ Strongly disagree

Fig. 2. A screenshot of the crowdsourcing page.

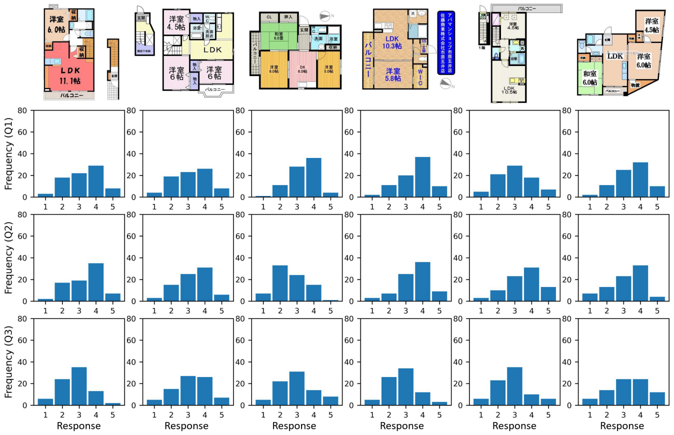


Fig. 3. Histograms of the responses (5-point Likert scale) for several examples of randomly selected floor plan images.

were applied to hedonic price models [16] to predict the apartment rent price after controlling for locational and structural characteristics of an apartment.

Recently, machine learning has been applied to the analysis of floor plan images. Yamasaki et al. [5] used deep neural networks (DNNs) to conduct semantic segmentation of floor plan images. They further developed a method to systematically generate adjacency graphs of floor plans from the semantically segmented images. Takada et al. [17] utilized multi-task learning for floor plan images and retrieved similar floor plans to the query. The floor plan recognition was then applied to property recommendation [18], [19] and retrieval [20]. Additionally, a toolbox for converting floor plan images to a vector format was developed in [21]. Furthermore, generating floor plans using graphs [22], [23], panoramic images [24], or 3D scans [25]–[30] is emerging. Generating furniture layouts using graphs was also discussed in [31].

To the best of our knowledge, this study is the first to propose an accurate prediction model for subjective scores of apartments using machine learning.

C. Prediction of Subjective Scores of Multimedia Content

There are extensive surveys in the literature introducing quality assessment studies and methods using images [32] and videos [33], [34].

Assessing the perceived low-level quality of images [35]–[38] and videos [39]–[42] has been an important topic in multimedia. These works tried to predict the perceived quality when the quality of the multimedia content is somewhat degraded by low-level factors such as compression and noise.

Predicting higher-level subjective scores has also been an active research area. For instance, analysis of image aesthetics relates more to color usage, composition, etc., and not to low-level noise in the content [43]–[48].

Sentiment and emotion classification and affective analysis constitute another direction of research for analyzing subjective evaluation of multimedia content. In this regard, many related works can be found in the literature for texts [49]–[51], images [52]–[59], speech [60], music [61], videos [53], [62]–[65], and their combinations [66]–[68].

Skills and creativeness have also been research targets, including skill assessment [69], creativity [70], click through rate prediction for online advertisements [71], [72], presentation slide assessment [73], [74], and so on.

To the best of our knowledge, this is the first work on subjective evaluation of apartment floor plans. In the paper, we show that predicting the subjective functionality and comfort is possible, and we also demonstrate possible applications of such predictions.

III. Dataset Creation

A. Subjective Scores

In this study, we used crowdsourcing to create a new dataset based on subjective evaluation of real estate floor plan images through a set of statements that question their levels of comfort, openness, privacy, and other characteristics.

In total, $3,128$ workers participated in this evaluation. We recruited 400 participants separately from 10 groups: two genders over five age ranges (20-29, 30-39, 40-49, 50-59, and 60+) using a crowdsourcing service.

We used floor plan images of Japanese rental apartments from the "Home's dataset" released by LIFULL Co., Ltd. with the cooperation of National Institute of Informatics[2], which has been widely used as a general floor plan image dataset in the international research community [22], [75]–[77]. We randomly selected $1,000$ floor plan images that included apartments with one, two, three, and four or more bedrooms in balanced proportions and prepared the following nine question statements for each image (questions Q1 to Q3 are about impressions, Q4 to Q6 are about functionality, and Q7 to Q9 are about environmental criteria):

- Q1 (Spaciousness): It is a spacious and open floor plan.
- Q2 (Modernity): The impression of the room layout is modern and contemporary.
- Q3 (Luxuriance): It is a luxurious apartment, and the rent might be expensive.

[2] National Institute of Informatics (NII), https://www.nii.ac.jp/dsc/idr/lifull/homes.html (accessed: 05.05.2020)

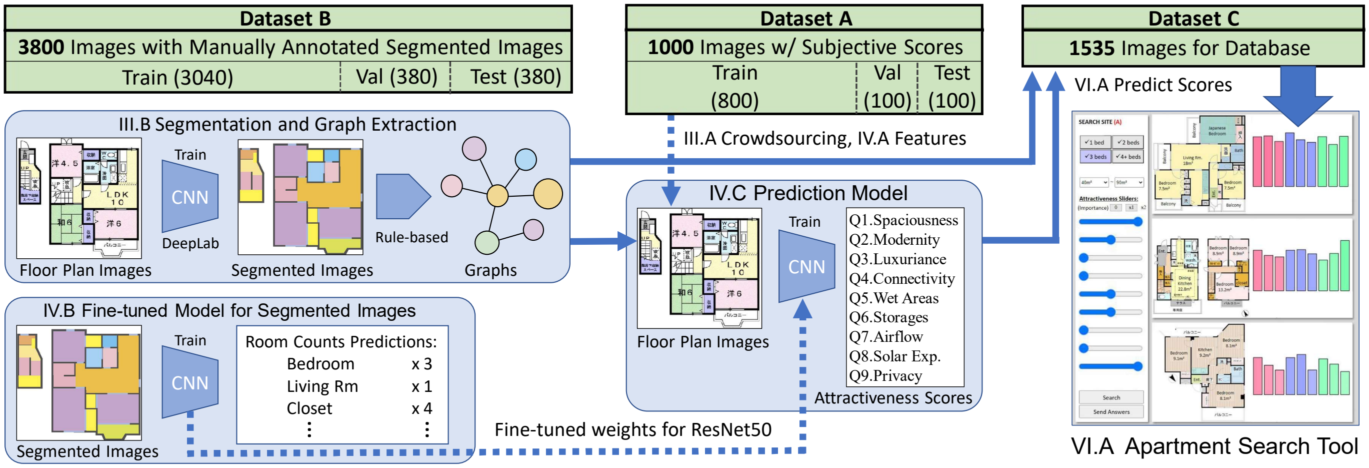


Fig. 4. Overview of our framework.

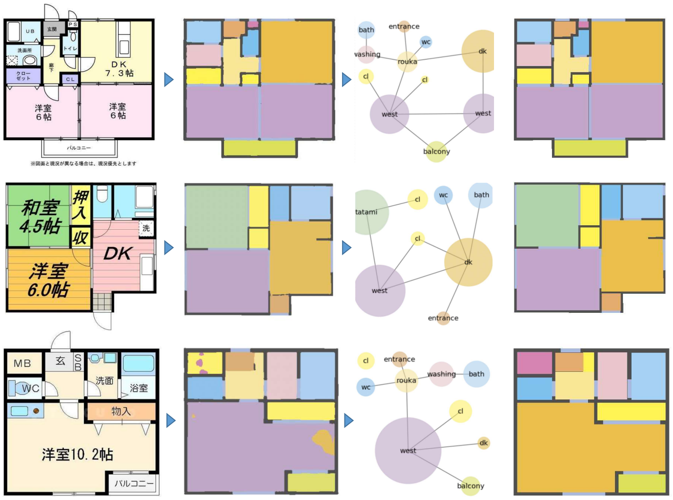


Fig. 5. Examples of input images (first column on the left), inferred semantically segmented images (second column), extracted graphs (third column), and grand truth segmented images (forth column).

- Q4 (Connectivity): Connectivity, adjacency, and layout of rooms and circulation are efficient and look comfortable.
- Q5 (Wet Areas): The traffic paths for the kitchen, bathroom, and restroom are good.
- Q6 (Storage): Locations and sizes of storage are good.
- Q7 (Airflow): Airflow inside the floor plan is good overall.
- Q8 (Solar Exp.): Solar exposure of the floor plan is good.
- Q9 (Privacy): The arrangement and adjacency of rooms fully consider the privacy of each family member.

In Q9, participants were provided a family size to evaluate the privacy of the image. Each floor plan was evaluated based on a five-grade score on a scale of 1 (strongly disagree) to 5 (strongly agree) by participants. Only the floor plan images were shown to participants for the evaluation (Fig. 2). Each participant was asked to evaluate 25 floor plan images. A task completion control was implemented, and the participants were required to answer all the assigned questions in order to complete the task; otherwise, their responses were not included in the study. After removing those from whom we did not receive all responses and also those who chose the same rating for all (i.e., "straight-lining"), we obtained 3,128 valid participants' results out of the 4,000 participants that were recruited. The remaining 871 include those who either dropped out or were not included for the reasons stated above. Thus, each floor plan was evaluated by 78 to 80 individuals (i.e., 2 genders × 5 age ranges × 400 participants × 25 images / 1,000 total images = 100 participants per image; with approximately 20% drop outs, this coincides with 3128/4000 = 0.782). We verified that the histograms of responses by participants were approximately normally distributed from random selections of over 200 floor plans and that no isolated peaks with unusual values were found. Figure 3 shows some examples of histograms from our randomly selected floor plans. Each participant would have different mean and standard deviation in their evaluation scores, and therefore the scores were standardized before taking the average among the participants. We used the value obtained by subtracting the mean value from each raw score and dividing it by the standard deviation. The resulting scores were normalized between -1 and 1 for each question. Figure 1 shows selected examples of floor plans with the corresponding results of the nine scores of subjective measure. The error bar represents the standard deviation. Next, for each floor plan, we averaged the scores from the nine questions. Since we averaged nine scores that originally came from different populations of scores with different distributions, we standardized the mean values from all floor plans and normalized them between -1 and 1; we called this value the "Total Score" of the floor plan. The Total Score represents the overall performance of each floor plan across all nine questions.

B. Segmentation and Graph Extraction

A floor plan can also be viewed as a graph with nodes representing rooms and with edges representing connections between them [5], [20], [22], [23]. Therefore, we extracted graphs from floor plan images and used them as features to predict the nine subjective scores defined above. This is a reasonable assumption because floor plan images are inspected for the connections among rooms and their adjacency.

In order to extract corresponding graphs automatically from floor plan images, we used the following two steps (Fig. 4 top left). First, we prepared $3,800$ new floor plan images (no overlap with dataset A), which we call dataset B, with their manually annotated segmented images using an online annotation tool (please see Fig. 4). They were consistently color-coded and semantically segmented into the following 15 classes of elements: wall, western bedroom (wbed), Japanese bedroom (jbed), dining kitchen (dk), restroom (wc), bathroom (bath), washroom (wash), balcony (balc), entrance (ent), corridor (corri), stairs, closet (cl), door, window, and unknown elements that do not belong to any category (abbreviations in parentheses are used in figures representing graph nodes). Then, we prepared a segmentation prediction network based on the method introduced in [5] using an improved network architecture, DeepLab v3+ [78], [79], instead of fully convolutional networks. Using $3,040$ images for training, 380 for validation, and 380 for testing, we trained the network using the segmented images as ground truth data. We automatically obtained $1,000$ segmented images with subjective scores for our dataset by feeding dataset A into this pre-trained network using DeepLab v3+.

Second, we used the rule-based method to extract $1,000$ graphs from the inferred segmented images using images in dataset A following the same procedure in [20]. Figure 5 shows examples of floor plan images, inferred segmented images, extracted graphs, and ground truth (GT) segmented images, which were manually annotated using an online annotation tool. We used 11 elements for nodes of the graphs, excluding the wall, door, window, and unknown elements from the above 15 elements. We added edges between two rooms only if they are directly accessible through a physical opening or door. Nodes were created by extracting regions representing rooms with a certain area in the inferred segmented images. The resulting dataset was used to determine whether the differences in graph structures influence the impression and functionality of apartments from subjective evaluations.

In Figure 5, it is noticeable that the inferred segmented images using the segmentation prediction network are noisy and slightly degraded compared to the GT segmented images. As a result, the extracted graphs from the inferred images are not as accurate as the grand truth graphs extracted from the GT segmented images. In Section V-B, we discuss how the imperfections in the use of inferred images and extracted graphs affects the prediction accuracy of our proposed model compared to the best-case using GT segmented images and GT graphs as input features as the upper bound performance. In fact, the average performance drop in Pearson correlation coefficient (PCC) is only 0.046 between two cases, which is statistically insignificant (see Section V-B). It is thus acceptable to use the inferred segmented images instead of GT segmented images, which allows us to automatically generate all required features only from floor plan images without preparing manually annotated floor plan images for new data inputs.

C. Overview of Our Datasets

Figure 4 shows an overview of our framework. To predict subjective scores only from the real-estate floor plan images, we prepared three sets of floor plan images without any overlaps to avoid bias for network models' training.

The first set of 1,000 images (dataset A) was used for the dataset with subjective scores and to train our prediction model in Section IV.

The second set with 3,800 images (dataset B) was used to obtain the network that outputs segmented images from the floor plan images as input, which as explained in Section III-B. We also used dataset B to train an ImageNet-pretrained feature extractor network using color-coded semantically segmented images, as shown in bottom left of Fig. 4. We additionally developed this network since our proposed network in Figure 8 explained in Section IV-C shows improved performance using the new weights fine-tuned by this network for ResNet50 [80] instead of using the weights only pre-trained on ImageNet on segmented images. The network was based on ResNet50 and used the segmented image as input. It was pre-trained first using ImageNet and then fine-tuned for the multi-task classification task to predict the number of rooms of the 13 room types, excluding the wall and unknown elements, from the 15 types defined in Section III-B.

Finally, we predicted subjective scores using our model from a separate set of $1,535$ floor plan images (hereafter, "dataset C"; please see Fig. 4), and it was also used for our proposed apartment search tool in Section VI.

IV. Proposed Methods

In this section, we introduce our prediction model. We propose to use two types of inputs for the model: floor plan images and structured data features. For the images, we used both floor plan images and semantically segmented images from Section III. In Section IV-A, we explain our method to extract four features from the structured data based on graphs and metadata of floor plans. The image features are described in Section IV-B. In Section IV-C, we introduce our prediction model in detail.

A. Features

1) Emerging Subgraphs: To extract frequently appearing common subgraphs that are very important and contribute to either higher or lower evaluation scores for each question regarding the floor plans, we first performed frequent subgraph mining on a total of $1,000$ graphs in dataset A using graph-based substructure pattern mining (gSpan) [81]. For the condition to evaluate subgraph isomorphism, we considered node attributes that represent the room types for graph matching. The edge attributes that represent door and window types were not considered. As a result, $162,470$ subgraphs were extracted by setting the minimum support threshold to 5 (i.e., the condition for extracting the common subgraph corresponding to at least 5 out of $1,000$). The following three steps were further performed to extract subgraphs that are more relevant to each subjective score.

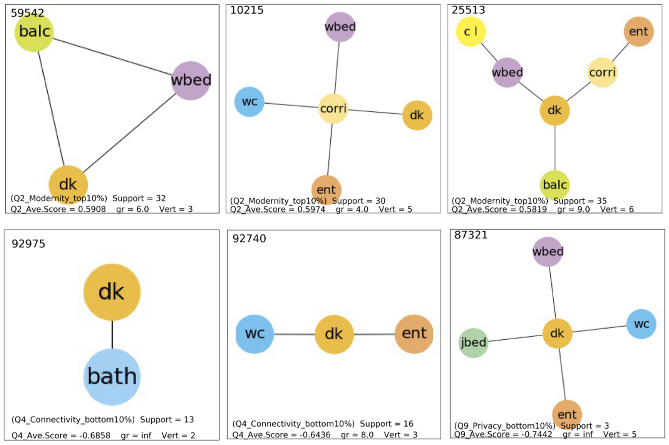

59542
balc
wbed
dk
(Q2_Modernity_top10%) Support = 32
Q2_Ave.Score = 0.5908 gr = 6.0 Vert = 3
10215
wbed
wc
corri
dk
ent
(Q2_Modernity_top10%) Support = 30
Q2_Ave.Score = 0.5974 gr = 4.0 Vert = 5
25513
c l
ent
wbed
corri
dk
balc
(Q2_Modernity_top10%) Support = 35
Q2_Ave.Score = 0.5819 gr = 9.0 Vert = 6
92975
dk
bath
(Q4_Connectivity_bottom10%) Support = 13
Q4_Ave.Score = -0.6858 gr = inf Vert = 2
92740
wc
dk
ent
(Q4_Connectivity_bottom10%) Support = 16
Q4_Ave.Score = -0.6436 gr = 8.0 Vert = 3
87321
wbed
jbed
dk
wc
ent
(Q9_Privacy_bottom10%) Support = 3
Q9_Ave.Score = -0.7442 gr = inf Vert = 5

Fig. 6. Examples of emerging subgraphs frequently found in floor plans having a high Q2 (modernity) score (top row). The top left also appears in Fig. 1a (circled in red). Examples found in floor plans having a low Q4 (connectivity) (bottom left & middle) and a Q9 (privacy) score (bottom right).

Step 1: For each question from Q1 to Q9 and the "Total Score" from Section III-A (10 items in total), common subgraphs that were included in the floor plan graphs with evaluation scores of the top and bottom 10% were extracted and separated into a total of 20 classes. We further narrowed down the selection of subgraphs by setting the following three thresholding strategies. First, the minimum support threshold was set to 10. Second, the average score of the apartments that contain the target subgraph should be 0.25 or larger for those in the top 10% classes and -0.25 or lower for those in the bottom 10% classes. Third, we also used the growth rate (GR), which is widely used for discovering discriminative patterns in emerging pattern mining [82]. The ratio between GR for the top 10% and that for the bottom 10% should be larger than 4 or smaller than $1/4$. These thresholds were set by our empirical study. *Step* 2: To further extract relevant common subgraphs in the above 20 classes, we considered identification numbers of extracted subgraphs as "words" in 20 different "documents." Then, we performed term frequency - inverse document frequency (TF-IDF), which evaluates how relevant a word (subgraph) is to a document (class) in a collection of documents (classes that represent each question). Based on the obtained TF-IDF weights, we sorted important subgraphs that are relevant to the top or bottom 10% of floor plans evaluated based on a specific question in each class. The use of TF-IDF helped eliminate frequently appearing subgraphs found in multiple classes that are not yet relevant to any specific class.

For any subgraph that included other subgraphs (i.e., inclusion dependency) within the same class, subgraphs with a minimum number of nodes were always kept. If the larger-size subgraph in the top 10% class has a larger mean evaluation score than those of the minimum-size subgraphs defined above, it is also kept (and vice-versa for the bottom 10% classes).

Step 3: We sorted all the extracted subgraphs from the previous steps based on the mean evaluation score in each class and obtained the top 20 subgraphs for each of the 20 classes. By eliminating duplicates that appeared in multiple classes from a total of 400 extracted emerging subgraphs, we obtained 230 unique emerging subgraphs. For each floor plan in the dataset, a 230-dimensional feature vector that contained 0 or 1 based on the absence or presence of 230 emerging subgraphs, respectively, were extracted.

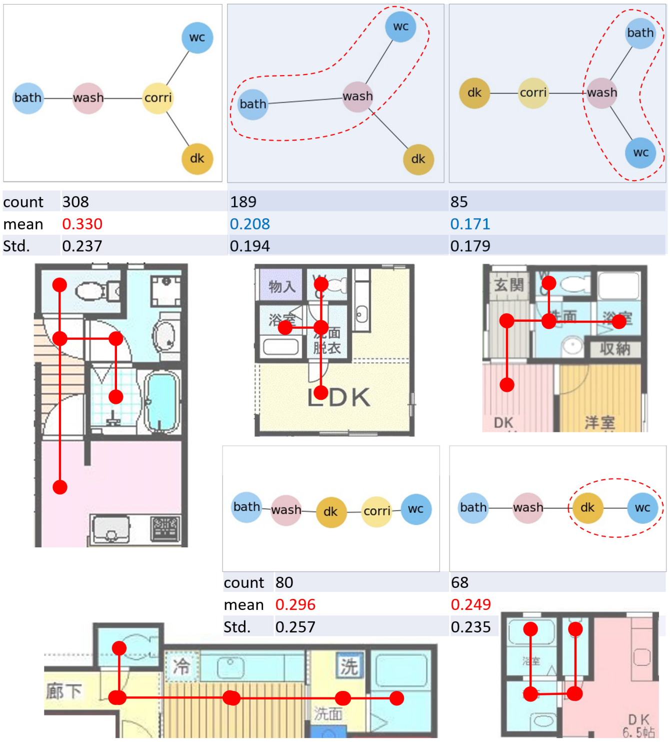


Fig. 7. The main five types of subgraphs for wet areas that make up 76% of all floor plans. Two subgraphs that include the linkage (wc)–(wash)–(bath), circled in red, have lower mean scores for Q5 (wet areas) than the others.

Figure 6 shows examples of the emerging subgraphs that were extracted using the above three steps. The subgraph on the top left is a triangle graph connecting (wbed)-(dk)-(balc) as a subgraph that appears in the top 10% of the Q2 (modernity) class. For example, the floor plan in Fig. 1a includes this subgraph and has a very high evaluation score for Q2. It represents the wide balcony space that is open to both the bedroom and kitchen and is often considered a contemporary layout among Japanese apartments (advertised as a "wide-span balcony"). The subgraph with two nodes, (dk)–(bath), was in the class representing the bottom 10% of the Q4 (adjacency and connectivity) class (bottom left in Fig. 6). This has a bathroom door that immediately opens to a living room, allowing anyone to enter a communal space directly after taking a bath, which is not considered desirable. Adding a washing room between the two nodes is a common practice in Japan, as it provides a room to change clothes before entering the bathroom. This subgraph was also included in the class for the bottom 10% for the Q9 (privacy) class. Our method was able to extract emerging subgraphs representing such characteristics of floor plans in 20 classes.

2) Subgraphs for Wet Areas: Wet areas are particularly important because they are more personal spaces. Therefore, special attention was paid to them. For each floor plan in dataset A, a connected subgraph containing rooms related to the usage of water, including dining kitchen (dk), washroom (wash), bathroom (bath), and restroom (wc), was extracted. Any nodes bridging the above four nodes such as a corridor (corri) that were necessary to form a single connected subgraph that contained all four of the above nodes were also included depending on the floor plan layouts. If there were more than two such bridging nodes and only one node was sufficient to form a connected subgraph, the node that served a communal use or as circulation, such as corridor, was selected over a node use for a private purpose, such as a bedroom. In total, 162 unique subgraph types for water-related rooms were extracted. Out of all the floor plans, 76% belonged to the five types. 120 types of subgraph appeared only once and therefore they were discarded. As a result, 42 subgraph types were used in this study.

Two out of the five remaining types of subgraphs included three nodes that were directly linked: (wc)–(wash)–(bath). However, this is not a desirable layout, as its circulation paths for (wc) and (bath) crisscross at the washing room (Fig. 7). It forces one to enter (wash) while someone else is still present after using either room. The mean evaluation scores for Q5 (wet areas) for those two types were lower (0.171 and 0.208) than the mean scores from the other three types (0.330, 0.296, and 0.249). As a result, we extracted a 42-dimensional one-hot feature vector based on the presence of the 42 types of subgraphs for wet areas.

3) Feature Based on MCS Graph Similarity: The similarity between graphs of all 800 floor plans in the training set of dataset A was calculated based on the method described in [17], [83], [84] using the maximum common subgraph (MCS) as a graph similarity measure. The similarity was 1 when the two graphs perfectly matched, and 0 when there were no common parts. Any input graph could be expressed by an 800-dimensional vector that represents distances from each of the 800 graphs in the training set of the dataset A. We used this vector based on the MCS similarity to extract a feature that represented an entire (global) characteristic of each apartment, as opposed to a local sub-structure from a subgraph.

4) Feature Based on Metadata: In addition to the above three features extracted from graph structures, we listed areas and numbers of room types using the metadata for each floor plan. This resulted in 30-dimensional feature vectors that represented the areas and numbers of room types.

From the above, we obtained four feature vectors that represented structured data based on emergent subgraphs, subgraphs for wet areas, MCS graph similarity, and metadata. All feature vectors were standardized before using them for machine learning (see the next section) such that their distributions had a mean value of 0 and a standard deviation of 1.

B. Image Features

We also fed the network two types of images: floor plan images and consistently color-coded semantically segmented images (prepared for the automated graph generation in Section III). The inputs to the network were two sets of RGB images of resolution 224 × 224.

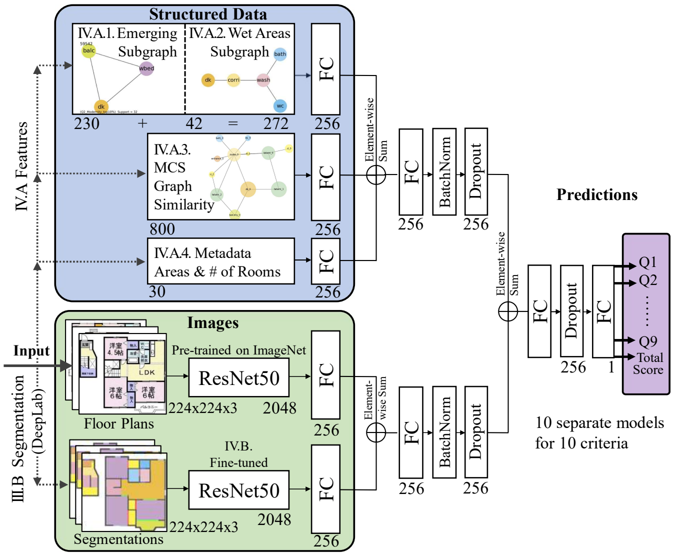


Fig. 8. Proposed network architecture.

First, for the floor plan images, using ResNet50 [80] pre-trained on ImageNet [85], a 2,048-dimensional vector of the pool5 layer was extracted from deep features of the images. Then, a new fully connected (FC) layer was added in place of the original one.

Second, for the segmented images, we used the same network architecture, but the network was pre-trained first using ImageNet and then fine-tuned for the multi-task classification task to predict the number of rooms of the 13 room types, excluding walls and unknown elements, from the 15 types defined in Section III-B. Similarly, a 2,048-dimensional vector of the pool5 layer was extracted.

Two features extracted from the two sets of images described above were also added in the same manner as above, followed by the FC, batch normalization (BN) [86], and dropout [87] layers. Finally, these two added features from the structured data and images were added again, followed by the FC and dropout layers. The final FC layer generated a value predicting the score (see Fig. 8 for details). For the FC layers, we used Leaky ReLu as an activation function.

C. Prediction Model

Figure 8 shows our proposed network architecture to predict the evaluation scores from Q1 to Q9 and the "Total Score" (ten separate models in total).

Two features based on emergent subgraphs (IV-A1) and subgraphs for wet areas (IV-A2) were concatenated into one 272-dimensional vector, which was reduced to 256 dimensions using a FC layer. The 800-dimensional feature vector from the MCS graph similarity (IV-A3) and the 30-dimensional feature based on metadata, such as areas and numbers of room types (IV-A4), were also reduced to 256 dimensions using FC layers. Then, the above three input features were added, followed by the FC, BN, and dropout layers. The latter two layers were added to improve generalization and to reduce overfitting.

V. Experiments

We used the dataset created in Section III and divided it into 800 floor plans for training, 100 for validation, and 100 for testing. We trained 10 separate models for predicting 10 scores using the network in Section IV-C because it is better than multi-task learning using a single model in our preliminary experiment. Regression models were created for each using the mean squared error (MSE) as a loss function, and we trained these networks using 800 training images. We applied the Momentum Stochastic Gradient Descent (SGD) algorithm to train models with a batch size of 20, a learning rate of 10^{-3}, a decay rate of 2.86×10^{-5}, and a momentum of 0.9 for the 35 epochs. The system was implemented using Keras [3]. We applied PCC to measure the correlation between the predicted values and the evaluation scores from the dataset as ground truths for 100 floor plans in the test data. We repeated the above process five times using a different randomly selected set of 800 training, 100 validation, and 100 test floor plans each time, and used the mean values of all five results as the final PCCs for the proposed as well as the following comparative baseline models, which were prepared to compare and verify the prediction results.

A. Baseline Methods

As baseline methods, we prepared the following seven models, of which five models use DNNs by subtracting one of five input features from the proposed network in Fig. 8:

- *Images Only: The model with only two input features based on floor plans and color-coded semantically segmented images.*
- *w/o SubG: The model without the input features based on subgraphs for emergent (IV-A1) and for wet areas (IV-A2).*
- *w/o MCS: The model without the input features based on MCS graph similarity (IV-A3).*
- *w/o Meta: The model without the input features based on metadata such as area and number of room types (IV-A4).*
- *w/o Img: The model without the features for floor plan images.*
- *w/o Segm: The model without the input features based on consistently color-coded semantically segmented images.*
- *Upper Bound: The model's network architecture is identical to the proposed model. However, the dataset based on GT segmented images and GT graphs was used for training and testing of the model as the best case, instead of using inferred segmented images and extracted graphs used in our proposed model explained in Sec. III-B)*

B. Results

1) Segmentation Performance: First, we discuss the accuracy of the semantic segmentation using dataset B introduced in III-B. The number of train/validation/test data were set to 3040/380/380, respectively, and we trained the segmentation prediction network model to learn the correspondence between the floor plan images and the ground truth (GT) label masks using the training and validation data. After the training, the test data was used for evaluation with a metric mean intersection over union (IoU) [88] defined by (1), where n_c is the number of classes, t_i is the total number of pixels belonging to class i, and n_j is the total number of predicted as class j belonging to class i.

$$\text{mean IoU} = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{n_{ii}}{t_i + \sum_{i'=1}^{n_c} n_{i'i} - n_{ii}}. \tag{1}$$

The mean IoU for the test data was 82.1%, and the average mean accuracy was 89.0% for the test set. Although the segmentation accuracy is not perfect, it is acceptable considering the prediction accuracy of functionality and comfort prediction as discussed below.

2) Performance Comparison: To evaluate the performance of each model, we used two metrics, namely Pearson Correlation Coefficient (PCC) and Root-Mean-Squared Error (RMSE). Table I shows the PCC results of both the proposed and baseline models using the test set of dataset A, and Table II shows the RMSE values. Note that a higher PCC and a lower RMSE mean better prediction performance. Our proposed model, which uses all features extracted from the structured data, floor plan images, and segmented images, outperformed all the comparative baseline models in terms of the mean value of PCCs from all 10 criteria (0.701) and recorded the highest PCC for 9 out of 10 criteria (Table I). While we used ResNet50, which is known for its high performance in a wide range of image-recognition tasks, the naive approach only using two sets of images, Images Only model, has a lower mean value of PCCS, 0.491. Therefore, it can be inferred that the network compensates for the weakness of each feature by combining them. The PCC values for our proposed model were over 0.7 for five criteria (Q1=0.721, Q2=0.793, Q3=0.776, Q6=0.751, and Total Score=0.794), and over 0.8 for the "privacy" criteria (Q9=0.816). Table II also shows that our proposed model outperforms baseline models with the lowest mean value of RMSEs from all 10 criteria.

We carried out the statistical significance tests to determine whether PCCs and RMSEs between the proposed and baseline models are statistically significant or not. Following the recommended procedure in Section 7.6.1 of ITU-T Rec. P.1401 [89], the statistical significance tests for PCC use statistics derived from Fisher's z-transformed correlation coefficients in each comparison, compared with the 95% two-tailed Student's t-test critical value. In Table I, we have used an asterisk to denote the case when 0.05 significance level of difference compared to the proposed model's PCC is found. Table III shows p-values from Images Only and Upper Bound models. Our multimodal network model outperformed the Images Only model with p-values>0.05 in 8 out of 10 criteria (all criteria except for Q5 and Q8) (Table III).

The statistical significance tests for the differences in RMSEs were also performed, following the recommended procedure in Section 7.6.4 of ITU-T Rec. P.1401 [89]. In Table II, we have used an asterisk to the RMSE values to denote the case when significant differences between the proposed model's RMSEs are found. The RMSEs from our proposed model and the Images Only model were significantly different

[3] https://keras.io

TABLE I
PREDICTION ACCURACY COMPARISON. VALUES ARE IN TERMS OF PEARSON CORRELATION COEFFICIENTS (PCC). BLACK CELLS REPRESENT PCC>0.7 AND GRAY CELLS INDICATE PCC<0.7. STATISTICAL SIGNIFICANCE BETWEEN PROPOSED AND BASELINE MODELS WERE DETERMINED BY A T-TEST (* DENOTES P-VALUE<0.05).

Model	Images Only	w/o SubG	w/o Meta	w/o MCS	w/o Img	w/o Segm	Proposed	Upper Bound
Q1.Spaciousness	**0.490***	0.669	0.707	0.730	0.637	0.648	0.721	0.794
Q2.Modernity	**0.600***	0.782	0.772	0.787	0.771	0.765	0.793	0.838
Q3.Luxuriance	**0.549***	0.762	0.751	0.753	0.749	0.747	0.776	0.805
Q4.Connectivity	**0.432***	0.542	0.592	0.620	0.546	0.571	0.637	0.698
Q5.Wet Areas	0.392	0.477	0.422	0.452	0.439	0.499	0.525	0.585
Q6.Storage	**0.518***	0.715	0.728	0.751	0.713	0.733	0.751	0.778
Q7.Airflow	**0.363***	0.494	0.573	0.514	0.446	0.592	0.607	0.657
Q8.Solar Exp.	0.402	0.520	0.531	0.571	0.525	0.564	0.591	0.680
Q9.Privacy	**0.567***	0.764	0.786	0.780	0.791	0.748	0.816	0.822
Total Score	**0.553***	0.727	0.763	0.740	0.712	0.755	0.794	0.809
Average	0.491	0.645	0.662	0.670	0.633	0.662	0.701	0.747

TABLE II
ROOT-MEAN-SQUARED ERROR (RMSE) COMPARISON.
(* DENOTES 0.05 SIGNIFICANCE LEVEL OF DIFFERENCE COMPARED TO THE PROPOSED MODEL'S RMSE)

Model	Images Only	w/o SubG	w/o Meta	w/o MCS	w/o Img	w/o Segm	Proposed	Upper Bound
Q1.Spaciousness	**0.338***	0.266	0.254	0.242	**0.285***	0.259	0.236	0.218
Q2.Modernity	**0.357***	0.257	0.266	0.255	0.272	0.267	0.253	0.202
Q3.Luxuriance	**0.281***	0.213	0.217	0.205	**0.235***	0.209	0.198	0.192
Q4.Connectivity	**0.342***	0.252	0.235	0.227	0.307	0.259	0.245	0.251
Q5.Wet Areas	0.253	0.215	0.225	0.209	**0.276***	0.203	0.216	0.238
Q6.Storage	**0.344***	0.282	0.274	0.259	0.265	0.282	0.249	0.228
Q7.Airflow	**0.332***	0.272	0.255	0.257	**0.318***	0.265	0.233	0.267
Q8.Solar Exp.	**0.303***	**0.255***	**0.253***	0.247	**0.255***	0.239	0.211	0.227
Q9.Privacy	**0.374***	0.242	0.242	0.248	0.235	0.265	0.239	0.218
Total Score	0.257	0.253	0.234	0.244	0.253	0.240	0.218	0.243
Average	0.318	0.251	0.246	0.239	0.270	0.249	0.230	0.228

TABLE III
SIGNIFICANCE OF THE DIFFERENCE IN PCCS COMPARED TO THE PROPOSED MODEL'S PCCS. VALUES ARE IN TERMS OF P-VALUES. (* DENOTES P-VALUE<0.05)

Model	Images Only	Upper Bound
Q1.Spaciousness	**0.009***	0.230
Q2.Modernity	**0.007***	0.350
Q3.Luxuriance	**0.004***	0.589
Q4.Connectivity	**0.042***	0.446
Q5.Wet Areas	0.240	0.544
Q6.Storage	**0.005***	0.647
Q7.Airflow	**0.024***	0.560
Q8.Solar Exp.	0.079	0.294
Q9.Privacy	**0.001***	0.895
Total Score	**0.001***	0.768

in 8 out of 10 criteria (all criteria except for Q5 and Total Score), indicating that the improvements using our multimodal network are statistically significant. Only Q5 (Wet Areas) does not show significant improvement using our model suggested by the differences in both PCC and RMSE values. As stated in Section IV-A2, subgraphs for wet areas have fewer pattern variations than other features, which may make prediction of the Q5 (Wet Areas) score more difficult. The test also suggests that our model is statistically significantly better than the model without the features for floor plan images, w/o Img, in 5 out of 10 criteria, including Q1, Q3, Q5, Q7, and Q8. The results indicate that the use of both features based on images and structured data extracted from graphs contributes to the improvement in prediction accuracy.

The differences in PCC and RMSE values between the proposed and Upper Bound models are not significant, indicating that the performance drops caused by the use of graphs extracted from inferred segmented images, instead of the GT graphs described in Section III-B, are not significant (average difference in PCCs from nine questions is 0.046). The inferred segmented images have some noise; therefore, the extracted graphs are slightly degraded from the best case using the grand truth segmented images. However, the performance drop is not significant. Our model can automatically generate all

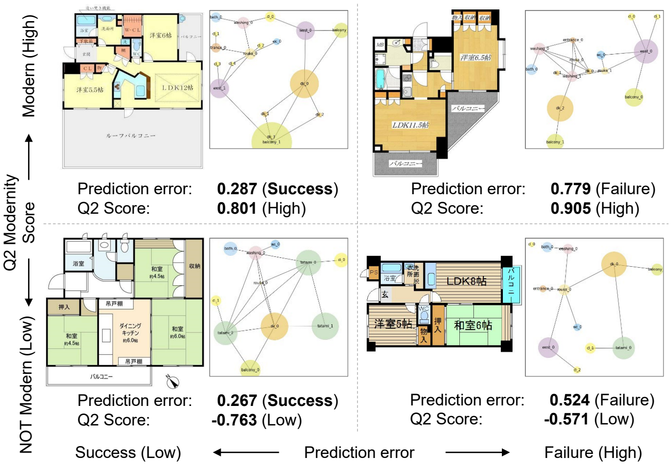


Fig. 9. Success and failure cases from our model's prediction results. While both floor plans in the top row have high ground-truth scores for Q2 Modernity in our dataset, the prediction error for the top-left was lower than the prediction error for the top-right, indicating a more successful prediction result.

input features from floor plan images alone as inputs using the processes in Section III-B to extract segmented images and graphs, allowing us to develop the search tool described in Section VI without the time-consuming and labor-intensive manual annotation process for segmented images. We also tested the prediction accuracy of our proposed network architecture trained and tested using a dataset based on raw values rated by the participants as evaluation scores, instead of using standardized evaluation scores in our dataset explained in Section III-A. We found no statistically significant difference in PCC and RMSE values between the two datasets (i.e., average PCC was 0.699 using raw score values and 0.701 using ours).

3) Success and Failure: Figure 9 depicts some success and failure examples from our prediction results for Q2 Modernity scores. Here, "success" means that the prediction error between the predicted score and the ground-truth score from the dataset is lower. It is inferred that if similar floor layouts are included in the proposed network's training data, prediction errors would likely be reduced. If the floor plans are rare or uncommon examples that do not exist in the training set, prediction errors could increase.

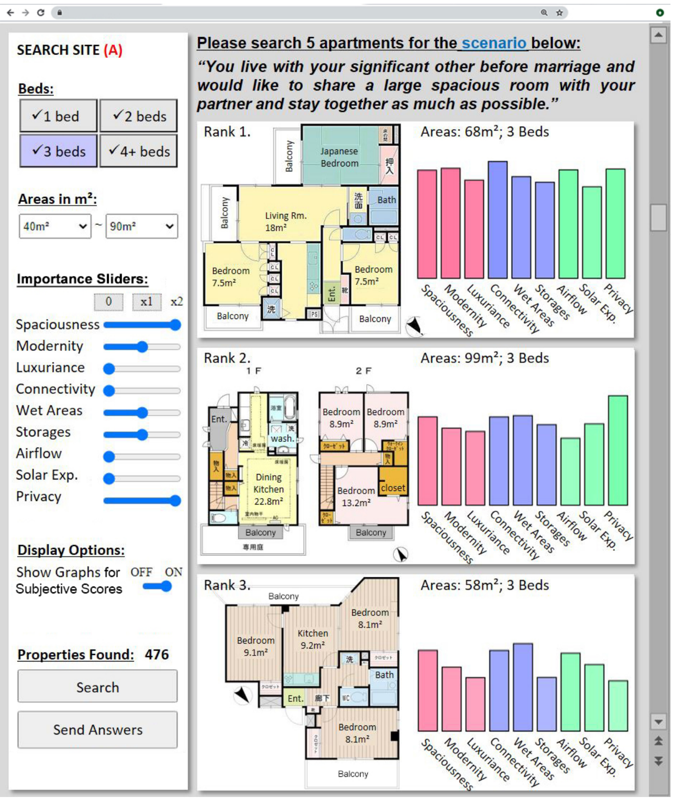
In conclusion, although the measurement of the functionality of floor plans is a very subjective problem, our proposed prediction models are able to achieve a very strong correlation with human evaluation.

VI. Usability Study

A. Implementation

We introduce a new interface for an apartment search tool that implements functionality and comfort as query items. In addition to a common search interface based on user selection of the number of bedrooms and a range of areas, our tool offers *importance sliders* with adjustable weights for importance in three levels: 0 (none); $\times 1$ (important); and $\times 2$ (very important), for nine subjective criteria (see Fig. 10). After a user presses the search button, the tool calculates the scores for all floor plans based on the weighted sum of the predicted subjective values using the information from the sliders, and displays the scores in ranked order. This feature allows a user to search apartments based on a controlled weighted priority for qualitative criteria (i.e., extremely spacious apartments with sufficient privacy and storage spaces).

For our proposed search tool's floor plan database, we prepared a new set of $1,535$ floor plan images that include apartments with one, two, three, and four or more bedrooms in balanced proportions (dataset C). These images were completely unbiased and were not previously used in this study. Using the pre-trained network using DeepLab v3+ and the rule-based method in Section III-B, we obtained the $1,535$ corresponding segmented images and graphs. Next, we executed the procedures outlined in Sections IV-A and IV-B to extract the five features from these images, which were then used to obtain the predicted attractiveness scores for the nine criteria for each of the $1,535$ floor plans using the best model (i.e., the one with the highest PCC) from the proposed network introduced in Section IV-C.



SEARCH SITE (A)

Please search 5 apartments for the scenario below:

"You live with your significant other before marriage and would like to share a large spacious room with your partner and stay together as much as possible."

Beds: ✓1 bed ✓2 beds [✓3 beds] ✓4+ beds

Areas in m²: 40m² ~ 90m²

Properties Found: 476

Rank 1. Areas: 68m²; 3 Beds

Rank 2. Areas: 99m²; 3 Beds

Rank 3. Areas: 58m²; 3 Beds

Fig. 10. The proposed apartment search tool.

B. Procedure

We evaluated our method through a large-scale user study. Among the 200 participants recruited through a crowdsourcing service, we removed three who chose the same rating for all questions (i.e., "straight-lining") and another six who did not complete the survey to obtain a final count of 191 participants with a wide variety of attributes. Out of the 191 participants, 81 identified themselves as male and 109 as female. The numbers of participants in their 20s, 30s, 40s, 50s, and over 60 were 30, 78, 53, 20, and 4, respectively. Of the total, 38 lived alone at the time of the experiment, while 81 participants were married. While 115 lived in urban areas, 60 lived in suburban areas, and 11 in rural areas. Participants were required to answer all questions to complete the task. Incomplete responses were not included in the study. We verified that all histograms of responses for 5-point Likert scale questions were approximately normally distributed, and that no isolated peaks with unusual values were found.

To make the study task realistic and provide internal motivation, we prepared hypothetical scenarios with different demands from five unique families and asked participants to search five apartments for each scenario. We set the scenarios based on five completely different family structures to avoid demographic background bias of the participants.

- *"You are a married couple, and your two children need their own rooms soon. You want a functional floor plan layout and don't want to pay extra for unnecessarily large spaces."*
- *"You are a family of five, living with one child and your spouse's parents. A well-functioning home with large stor-*

age spaces, kitchen, and wet areas are your top priority since you have a big family.”

- *“Due to the COVID-19 pandemic, you have been working from home and would like to have your own study room. You and your spouse want to have separate rooms to respect each other’s privacy.”*
- *“You live with your significant other before marriage and would like to share a large spacious room with your partner and stay together as much as possible.”*
- *“You and your partner are a young couple and hope to have at least one child in the future. You prefer to have a large balcony for your family to spend the weekend together.”*

We assigned these tasks using both our proposed tool and another baseline tool. The baseline tool represented commonly available real estate portal sites without our proposed features, featuring a search interface for a user to select the number of bedrooms and a range of areas. We studied major real estate portal sites [4][5][6][7], and found the above two items as common search features. The differences between the proposed and baseline tool are that the baseline tool has an identical interface except that it does not have functionality- and comfort-related options. As we wanted participants to focus on analyzing information only readable from floor plans to enable a fair comparison with our proposed tool, we excluded other common search features based on the location and cost of properties. The study employed a within-participant design in which participants used both tools (counterbalanced across participants) and provided feedback on them.

Participants were asked to search five apartments that met needs of each of the five scenarios using one tool, and then switch to the other tool for the same search (with the order of tool counterbalanced across participants). After completing both search tasks, participants completed a post-experiment survey. The survey asked participants to directly compare their experience with the proposed tool to the baseline tool in five Likert-scale questions (shown in Table IV). We also asked them to rate 50 selected floor plans (i.e., five plans × five scenarios × two tools) in a five-grade score on a scale of 1 (very unsatisfied) to 5 (very satisfied) (i.e., each participant gave a score of 1 to 5 for each retrieved floor plan image.).

C. Results

Overall, it can be observed that the participants showed a significant preference for our proposed tool in response to all five direct comparison questions in Table IV. For each question, 95% confident intervals of mean scores are indicated. In addition, a one-sample one-tailed t-test ($p<0.05$) was used to evaluate whether the mean responses are greater than 3 (i.e., a score of 3 indicates no preference) at the significance level. Participants appreciated the proposed tool as well. Compared to the baseline tool, it helped them find significantly more desirable floor plans (M = 4.03, 95% CI[3.90, 4.17], p=4.9×10^{-34} <0.001), and gave them a significantly more enjoyable experience (M = 3.97, 95% CI[3.83, 4.11], p=1.92×10^{-30} <0.001).

TABLE IV
THE DIRECT COMPARISON QUESTIONS WERE ASKED ON A 5-POINT LIKERT SCALE. A HIGHER SCORE INDICATED A PREFERENCE FOR OUR PROPOSED TOOL, WHILE A LOWER SCORE INDICATED A PREFERENCE FOR THE BASELINE TOOL. A SCORE OF 3 INDICATED NO PREFERENCE.

Question	Mean	CI	p-value
Which tool helped you find more desirable floor plans?	4.03*	[3.90, 4.17]	4.9×10^{-34} < 0.001
Which tool did you enjoy better while searching?	3.97*	[3.83, 4.11]	1.6×10^{-30} < 0.001
Which tool was faster for you to search floor plans?	3.84*	[3.67, 4.01]	1.1×10^{-18} < 0.001
Which tool was easier to search apartments?	3.76*	[3.60, 3.93]	2.5×10^{-16} < 0.001
Which tool was more intuitive for you to search floor plans?	3.43*	[3.25, 3.62]	7.0×10^{-6} < 0.001

* Significantly different based on 95% confidence interval.

TABLE V
THE 5-GRADE RATINGS OF ALL FLOOR PLANS SELECTED BY PARTICIPANTS USING TWO METHODS.

	Baseline	Proposed	p-value
Mean scores of the 5-grade ratings of all floor plans selected	3.75	4.01	1.67×10^{-40} < 0.001

An independent two-sample one-tailed t-test was used on two samples of scores of floor plans using our tool and the baseline tool.

Even though the participants spent a longer average time to complete a search task using the proposed tool (M = 158.0s, median = 131s, SD = 105.7s) than the baseline tool (M = 129.0s, median = 105s, SD = 97.4s), the result from the direct comparison question shows that participants felt that they found the desired floor plans faster using the proposed tool (M = 3.84, 95% CI[3.67, 4.01], p=1.1×10^{-18} <0.001). The number of clicks on buttons per search increased using the proposed tool (M = 18.3, median = 17, SD=7.6) compared to the baseline tool (M = 13.2, median = 12, SD = 5.9). These results show that our proposed tool required more time and mouse clicks for users due to additional features that are not included in the baseline tool (e.g., importance sliders). However, these additional features did not lead them to believe that the proposed system is harder to use. Participants felt that the proposed tool made the task significantly easier (M = 3.76, 95% CI[3.60, 3.93], p=2.5×10^{-16} <0.001) and more intuitive (M = 3.43, 95% CI[3.25, 3.62], p=7.0×10^{-6} <0.001). The longer search time and additional operations did not make them feel burdened with tasks, and they preferred the proposed tool over the baseline one.

In Table V, the 5-grade ratings of all floor plans selected by participants also indicated that they were significantly more satisfied with their selections using the proposed tool than the baseline one (M(proposed) = 4.01, M(baseline) = 3.75, p=1.67×10^{-40} <0.001). To compute the p-value, we used an independent two-sample one-tailed t-test, and the mean score for selected floor plans by participants using our proposed tool was found to be significantly higher than the mean score for selected floor plans using the baseline tool.

[4] https://lifull.com
[5] https://www.livable.co.jp
[6] https://suumo.jp
[7] https://www.redfin.com

VII. Limitations

One limitation of our work is that predicted functionality and comfort scores do not come with explanations. People may have different opinions and viewpoints on such subjective scores, and visualizing how the system evaluates floor plan images would be preferred.

Our proposed method has been applied only to apartment floor plans in Japan. As can be seen in the figures, the drawing styles employed in the floor plan images are very diverse; however, the style employed in other counties may be more distinct. Therefore, the prediction model will need to be trained for each country. We expect that our segmentation and functionality and comfort prediction models can be used as pre-trained models for fine-tuning, but this is left as a future work.

VIII. Conclusions

We created and analyzed a large-scale dataset based on subjective evaluation of real estate floor plan images using crowdsourcing. Our proposed methods for extracting features from graph structures and images of floor plans proved to be effective, as we obtained functionality and comfort prediction models with relatively high accuracy (PCC=0.701) for very subjective scores, which is essentially different from conventional image recognition tasks where the answer is apparent to all the evaluators. This study is the first work to propose a highly accurate prediction model for dwelling functionality and comfort using machine learning. We applied the results of the prediction model to our new apartment search tool using functionality and comfort as query items, and our user study showed that our tool could provide a better user experience.

References

[1] V. James, S. Wu, A. Gelfand, and C. Sirmans, "Apartment rent prediction using spatial modeling," *Journal of Real Estate Research*, vol. 27, no. 1, pp. 105–136, 2005.

[2] O. Poursaeed, T. Matera, and S. Belongie, "Vision-based real estate price estimation," *Machine Vision and Applications*, vol. 29, no. 4, pp. 667–676, 2018.

[3] M. Heidari, S. Zad, and S. Rafatirad, "Ensemble of supervised and unsupervised learning models to predict a profitable business decision," in *IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1–6.

[4] H. Seya and D. Shiroi, "A comparison of residential apartment rent price predictions using a large data set: Kriging versus deep neural network," *Geographical Analysis*, pp. 1–22, 2021.

[5] T. Yamasaki, J. Zhang, and Y. Takada, "Apartment structure estimation using fully convolutional networks and graph model," in *ACM Workshop on Multimedia for Real Estate Tech*, 2018, pp. 1–6.

[6] S. Law, B. Paige, and C. Russell, "Take a look around: Using street view and satellite images to estimate house prices," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 5, Sep. 2019.

[7] M. Zeppelzauer, M. Despotovic, M. Sakeena, D. Koch, and M. Döller, "Automatic prediction of building age from photographs," in *ACM International Conference on Multimedia Retrieval*, 2018, pp. 126–134.

[8] J. H. Bappy, J. R. Barr, N. Srinivasan, and A. K. Roy-Chowdhury, "Real estate image classification," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 373–381.

[9] X. Wang, Y. Takada, Y. Kado, and T. Yamasaki, "Predicting the attractiveness of real-estate images by pairwise comparison using deep learning," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2019, pp. 84–89.

[10] T. Hanazato, Y. Hirano, and M. Sasaki, "Syntactic analysis of large-size condominium units supplied in the tokyo metropolitan area," *Journal of Architecture and Planning*, no. 591, pp. 9–16, 2005.

[11] A. Takizawa, K. Yoshida, and N. Katoh, "Applying graph mining to rent analysis considering room layouts," *Journal of Environmental Engineering (Transaction of AIJ)*, vol. 73, no. 623, pp. 139–146, 2008.

[12] R. Hattori, K. Okamoto, and A. Shibata, "Visualizing the importance of floor-plan image features in rent-prediction models," in *Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS)*, 2020, pp. 1–3.

[13] ——, "Impact analysis of floor-plan images for rent-prediction model (in japanese)," *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, vol. 33, no. 2, pp. 640–650, 2021.

[14] K. Solovev and N. Pröllochs, "Integrating floor plans into hedonic models for rent price appraisal," in *Web Conference*, 2021, p. 2838–2847.

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

[16] S. Sirmans, D. Macpherson, , and E. Zietz, "The composition of hedonic pricing models," *Journal of Real Estate Literature*, vol. 13, no. 1, p. 1–44, 2005.

[17] Y. Takada, N. Inoue, T. Yamasaki, and K. Aizawa, "Similar floor plan retrieval featuring multi-task learning of layout type classification and room presence prediction," in *IEEE International Conference on Consumer Electronics (ICCE)*, 2018, pp. 1–6.

[18] K. Naoki, Y. Toshihiko, A. Kiyoharu, and O. Takemi, "Users' preference prediction of real estates featuring floor plan analysis using floornet," in *ACM Workshop on Multimedia for Real Estate Tech*, 2018, p. 7–11.

[19] N. Kato, T. Yamasaki, K. Aizawa, and T. Ohama, "Users' preference prediction of real estate properties based on floor plan analysis," *IEICE TRANSACTIONS on Information and Systems*, vol. E103-D, no. 2, pp. 398–405, 2020.

[20] M. Yamada, X. Wang, and T. Yamasaki, "Graph structure extraction from floor plan images and its application to similar property retrieval," in *IEEE International Conference on Consumer Electronics*, 2021.

[21] V. Trinh and R. Manduchi, "Semantic interior mapology: A toolbox for indoor scene description from architectural floor plans," *arXiv preprint arXiv:1911.11356*, 2019.

[22] N. Nauata, K.-H. Chang, C.-Y. Cheng, G. Mori, and Y. Furukawa, "House-gan: Relational generative adversarial networks for graph-constrained house layout generation," 2020.

[23] R. Hu, Z. Huang, Y. Tang, O. van Kaick, H. Zhang, and H. Huang, "Graph2plan: Learning floorplan generation from layout graphs," *ACM Transactions on Graphics*, 2020.

[24] S.-T. Yang, F.-E. Wang, C.-H. Peng, P. Wonka, M. Sun, and H.-K. Chu, "Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama," in *IEEE/CVF Computer Vision and Pattern Recognition*, 2019, pp. 3363–3372.

[25] C. Lin, C. Li, and W. Wang, "Floorplan priors for joint camera pose and room layout estimation," 2018.

[26] C. Liu, J. Wu, and Y. Furukawa, "Floornet: A unified framework for floorplan reconstruction from 3d scans," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 203–219.

[27] C. Lin, C. Li, and W. Wang, "Floorplan-jigsaw: Jointly estimating scene layout and aligning partial scans," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 5674–5683.

[28] J. Chen, C. Liu, J. Wu, and Y. Furukawa, "Floor-sp: Inverse cad for floorplans by sequential room-wise shortest path," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[29] Y. Cui, Q. Li, B. Yang, W. Xiao, C. Chen, and Z. Dong, "Automatic 3-d reconstruction of indoor environment with mobile laser scanning point clouds," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 3117–3130, 2019.

[30] A. Phalak, V. Badrinarayanan, and A. Rabinovich, "Scan2plan: Efficient floorplan generation from 3d scans of indoor scenes," *arXiv preprint arXiv:2003.07356*, 2020.

[31] K. Wang, Y.-A. Lin, B. Weissmann, M. Savva, A. X. Chang, and D. Ritchie, "Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks," *ACM Transactions on Graphics*, vol. 38, no. 4, 2019.

[32] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, no. 11, pp. 1–52, 2020.

[33] X. Min, K. Gu, G. Zhai, X. Yang, W. Zhang, P. Le Callet, and C. W. Chen, "Screen content quality assessment: overview, benchmark, and beyond," *ACM Computing Surveys (CSUR)*, vol. 54, no. 9, pp. 1–36, 2021.

[34] W. Zhou, X. Min, H. Li, and Q. Jiang, "A brief survey on adaptive video streaming quality assessment," *Journal of Visual Communication and Image Representation*, p. 103526, 2022.

[35] Q. Li, W. Lin, J. Xu, and Y. Fang, "Blind image quality assessment using statistical structural and luminance features," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2457–2469, 2016.

[36] Q. Wu, H. Li, Z. Wang, F. Meng, B. Luo, W. Li, and K. N. Ngan, "Blind image quality assessment based on rank-order regularized regression," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2490–2504, 2017.

[37] Z. Fan, T. Jiang, and T. Huang, "Active sampling exploiting reliable informativeness for subjective image quality assessment based on pairwise comparison," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2720–2735, 2017.

[38] H. Hofbauer, F. Autrusseau, and A. Uhl, "Low quality and recognition of image content," *IEEE Transactions on Multimedia*, 2021.

[39] Y. Li, S. Meng, X. Zhang, M. Wang, S. Wang, Y. Wang, and S. Ma, "User-generated video quality assessment: A subjective and objective study," *IEEE Transactions on Multimedia*, 2021.

[40] Y. Liu, J. Wu, A. Li, L. Li, W. Dong, G. Shi, and W. Lin, "Video quality assessment with serial dependence modeling," *IEEE Transactions on Multimedia*, 2021.

[41] J. Gutierrez, P. Perez, M. Orduna, A. Singla, C. Cortes, P. Mazumdar, I. Viola, K. Brunnstrom, F. Battisti, N. Cieplinska, D. Juszka, L. Janowski, M. I. Leszczuk, A. Adeyemi-Ejeye, Y. Hu, Z. Chen, G. Van Wallendael, P. Lambert, C. Diaz, J. Hedlund, O. Hamsis, S. Fremerey, F. Hofmeyer, A. Raake, P. Cesar, M. Carli, and N. Garcia, "Subjective evaluation of visual quality and simulator sickness of short 360 videos: Itu-t rec. p.919," *IEEE Transactions on Multimedia*, 2021.

[42] P. Lebreton and K. Yamagishi, "Predicting user quitting ratio in adaptive bitrate video streaming," *IEEE Transactions on Multimedia*, vol. 23, pp. 4526–4540, 2021.

[43] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.

[44] G. Guo, H. Wang, C. Shen, Y. Yan, and H.-Y. M. Liao, "Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2073–2085, 2018.

[45] V. Hosu, B. Goldlücke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9367–9375.

[46] P. Lv, J. Fan, X. Nie, W. Dong, X. Jiang, B. Zhou, M. Xu, and C. Xu, "User-guided personalized image aesthetic assessment based on deep reinforcement learning," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[47] H. Zhu, Y. Zhou, L. Li, Y. Li, and Y. Guo, "Learning personalized image aesthetics from subjective and objective attributes," *IEEE Transactions on Multimedia*, 2021.

[48] Y. Bai, Z. Zhu, G. Jiang, and H. Sun, "Blind quality assessment of screen content images via macro-micro modeling of tensor domain dictionary," *IEEE Transactions on Multimedia*, vol. 23, pp. 4259–4271, 2021.

[49] X. Lei, X. Qian, and G. Zhao, "Rating prediction based on social sentiment from textual reviews," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1910–1921, 2016.

[50] F. Chen, R. Ji, J. Su, D. Cao, and Y. Gao, "Predicting microblog sentiments via weakly supervised multimodal deep learning," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 997–1007, 2018.

[51] R. Ji, F. Chen, L. Cao, and Y. Gao, "Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1062–1075, 2019.

[52] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *ACMMM*, 2014, pp. 457–466.

[53] T. Li, B. Ni, M. Xu, M. Wang, Q. Gao, and S. Yan, "Data-driven affective filtering for images and videos," *IEEE TCYB*, 2015.

[54] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2513–2525, 2018.

[55] X. Yao, D. She, H. Zhang, J. Yang, M.-M. Cheng, and L. Wang, "Adaptive deep metric learning for affective image retrieval and classification," *IEEE Transactions on Multimedia*, vol. 23, pp. 1640–1653, 2021.

[56] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2513–2525, 2018.

[57] S. Ruan, K. Zhang, L. Wu, T. Xu, Q. Liu, and E. Chen, "Color enhanced cross correlation net for image sentiment analysis," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[58] Y. Su, W. Zhao, P. Jing, and L. Nie, "Exploiting low-rank latent gaussian graphical model estimation for visual sentiment distribution," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[59] H. Zhang and M. Xu, "Multiscale emotion representation learning for affective image recognition," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[60] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.

[61] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE TMM*, vol. 8, no. 3, pp. 564–574, June 2006.

[62] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE TMM*, vol. 12, no. 6, pp. 523–535, Oct 2010.

[63] R. Teixeira, T. Yamasaki, and K. Aizawa, "Affective determination of video clips by low level audiovisual features -a dimensional and categorial approach-," *Multimedia Tools and Applications*, 2011.

[64] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE TAC*, vol. 3, no. 2, pp. 211–223, 2012.

[65] T. Liu, J. Wan, X. Dai, F. Liu, Q. You, and J. Luo, "Sentiment recognition for short annotated gifs using visual-textual fusion," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1098–1110, 2020.

[66] Q. Fang, C. Xu, J. Sang, M. S. Hossain, and G. Muhammad, "Word-of-mouth understanding: Entity-centric multimodal aspect-opinion mining in social media," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2281–2296, 2015.

[67] W. Guo, Y. Zhang, X. Cai, L. Meng, J. Yang, and X. Yuan, "Ld-man: Layout-driven multimodal attention network for online news sentiment recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 1785–1798, 2021.

[68] W. Nie, R. Chang, M. Ren, Y. Su, and A. Liu, "I-gcn: Incremental graph convolution network for conversation emotion detection," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[69] D. Liu, Q. Li, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li, "Towards unified surgical skill assessment," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9517–9526.

[70] M. Redi, N. O'Hare, R. Schifanella, M. Trevisiol, and A. Jaimes, "6 seconds of sound and vision: Creativity in micro-videos," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4272–4279.

[71] B. Xia, X. Wang, T. Yamasaki, K. Aizawa, and H. Seshime, "Deep neural network-based click-through rate prediction using multimodal features of online banners," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 162–170.

[72] J. Ikeda, H. Seshime, X. Wang, and T. Yamasaki, "25th international conference on pattern recognition (icpr)," in *25th International Conference on Pattern Recognition (ICPR)*, 2995-3002, pp. 2995–3002.

[73] S. Oyama and T. Yamasaki, "Visual clarity analysis and improvement support for presentation slides," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 421–428.

[74] S. Yi, J. Matsugami, and T. Yamasaki, "Assessment system of presentation slide design using visual and structural features," *IEICE Transactions on Information and Systems*, vol. E105-D, no. 3, 2022.

[75] C. Liu, J. Wu, P. Kohli, and Y. Furukawa, "Raster-to-vector: Revisiting floorplan transformation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2214–2222.

[76] Z. Zeng, X. Li, Y. K. Yu, and C.-W. Fu, "Deep floor plan recognition using a multi-task network with room-boundary-guided attention," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[77] C. Mura, R. Pajarola, K. Schindler, and N. Mitra, "Walk2map: Extracting floor plans from indoor walk trajectories," in *Computer Graphics Forum*, vol. 40, no. 2, 2021, pp. 375–388.

[78] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[79] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[80] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[81] X. Yan and J. Han, "gspan: Graph-based substructure pattern mining," in *IEEE International Conference on Data Mining*, 2002, pp. 721–724.

[82] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in *fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 1999, pp. 43–52.

[83] K. Ohara, T. Yamasaki, and K. Aizawa, "An intuitive system for searching apartments using floor plans and areas of rooms," *78th national convention of IPSJ*, vol. 2016, no. 1, pp. 311–312, mar 2016.

[84] J. J. McGregor, "Backtrack search algorithms and the maximal common subgraph problem," *Software: Practice and Experience*, vol. 12, no. 1, pp. 23–34, 1982.

[85] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[86] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[87] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[88] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[89] P. ITU-T, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models (p. 1401)," http://handle.itu.int/11.1002/1000/14159., 2020.

Taro Narahara (Member, IEEE) is currently an Associate Professor with the Hillier College of Architecture and Design, New Jersey Institute of Technology (NJIT). He received the Doctor of Design degree from Harvard University and the M.S. degree in design and computation from Massachusetts Institute of Technology (MIT). He also holds an M.Arch. degree from Washington University and a B.S. degree in Mathematics from Waseda University. Between 2018 and 2020, he was a visiting scholar at ETH Zurich's Institute of Technology in Architecture (ITA) and the University of Tokyo's Graduate School of Information Science and Technology. He worked on award-winning projects such as the Mori Arts Center while associated with Gluckman Mayner Architects and Skidmore, Owings & Merrill LLP as a licensed architect. His research interests include architectural design and machine learning. Dr. Narahara is a member of ACM, SIGGRAPH, CAADRIA, and IEICE.

Toshihiko Yamasaki (Member, IEEE) received the B.S. degree in electronic engineering, the M.S. degree in information and communication engineering, and the Ph.D. degree from The University of Tokyo, Bunkyo City, Tokyo, in 1999, 2001, and 2004, respectively. He is currently a Professor with the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo. He was a JSPS Fellow for Research Abroad and a Visiting Scientist with Cornell University, Ithaca, NY, USA, from February 2011 to February 2013. His research interests include attractiveness computing based on multimedia Big Data analysis, pattern recognition, and machine learning. Dr. Yamasaki is a member of ACM, AAAI, IEICE, ITE, and IPSJ.