Omni-CityMood: Vision-based Urban Atmosphere Perception from Every Angle

Yuki Kubota*
Institute of Science Tokyo
Tokyo, Japan
kubota@miubiq.cs.titech.ac.jp

Kota Tsubouchi*
LY Corporation
Tokyo, Japan
ktsubouc@lycorp.co.jp

Soto Anno
Institute of Science Tokyo
Tokyo, Japan
anno@miubiq.cs.titech.ac.jp

Kaito Ide
Institute of Science Tokyo
Tokyo, Japan
ide@miubiq.cs.titech.ac.jp

Masamichi Shimosaka
Institute of Science Tokyo
Tokyo, Japan
simosaka@miubiq.cs.titech.ac.jp

ABSTRACT

Understanding how cities are perceived from on-site visitors' perspectives can provide valuable insights for urban planning and development applications. However, existing studies estimated people's perceptions by having them view photographed landscape images; the scores derived by these methods were thus merely quantified impressions of specific viewpoints that do not necessarily represent perceptions people would have were they at the site. To address this issue, we developed a framework, named **Omni-CityMood**, for quantifying people's on-site perceptions of urban atmospheres. Based on the idea that the viewpoint influences the perception of an urban landscape, the proposed framework identifies critical viewpoints of a location by using both visual-based features of landscape images and geographical characteristics of the site. In particular, Omni-CityMood enables the mood of a location to be evaluated from viewpoints over a range of 360 degrees by leveraging the techniques of neural recommendation systems. We evaluated Omni-CityMood on a dataset we built that includes perceived atmosphere experiences in various cities. Experiments and extensive analyses demonstrate the promising capability of modeling landscape viewpoints to quantify urban on-site atmospheres.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**; • **Computing methodologies** → *Perception.*

KEYWORDS

Urban perception, multi-modal fusion, neural recommendation

* These authors equally contributed to this work.

SIGSPATIAL '25, November 3–6, 2025, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2086-4/2025/11…$15.00
https://doi.org/10.1145/3748636.3762722

ACM Reference Format:
Yuki Kubota*, Kota Tsubouchi*, Soto Anno, Kaito Ide, and Masamichi Shimosaka. 2025. Omni-CityMood: Vision-based Urban Atmosphere Perception from Every Angle. In *The 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3748636.3762722

1 INTRODUCTION

With the growing volume of visual geo-tagged images, there has been an increasing effort to analyze various aspects of cities based on landscape images [13, 16, 41]. As typified by the broken window theory [42], the appearance of a landscape has a tremendous impact on people's psychological condition [20, 21]. Thus, quantifying impressions, or *perceptions*, of urban landscapes should play an important role in urban planning and development [9, 35, 36]. With recent advances in machine-learning techniques and the availability of large amounts of landscape images, researchers have attempted to quantify people's perceptions of landscapes by leveraging computer vision technologies [1, 16, 23, 29–31]. This field is known as *visual urban perception* and advanced network architectures have been developed for it [10, 25].

While existing studies have accurately estimated perceptions of locations using landscape images taken from pre-determined viewpoints, they have yet to capture the impressions visitors to the site perceive. Visitors do not judge a location from a single pre-determined view; instead, they judge the overall impression of the site from a 360-degree view of the surrounding area. However, estimating such *real* perceptions of a 360-degree landscape is a non-trivial problem because human perceptions are completely variable depending on the viewpoint direction, visual scale, and coherence [37, 38, 50].

Changing viewpoints, even in the same place, can give significantly different impressions of the landscape's atmosphere. Figure 1 shows examples of landscape images taken from different viewpoints at the same location. We set the reference viewpoint as 0 degree, corresponding to the heading parallel to the road facing the north side. The values in the figure represent the heading incremented by 90 degrees clockwise from the reference viewpoint. Figure 1(a) shows an example of an area where a distinctive temple is present. The e views show general residential areas, while the

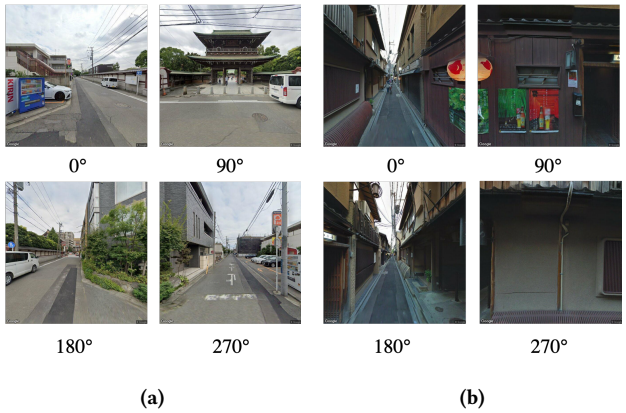


Figure 1: Examples of landscape images taken at the same location but from different camera viewpoints. (a) Example images taken nearby a temple. (b) Example landscape images taken at a historical location with a narrow alley.

90-degree view, perpendicular to a road, offers a conspicuous temple. Clearly, while the 0, 180, and 270-degree views would typically give the impression of an ordinary, mundane landscape, the view of the temple may provide an extraordinary impression that will dominate visitor perceptions of this location. In contrast, in the example shown in Figure 1(b) of a narrow road area, the viewpoint perpendicular to the road captures only the wall, providing limited information for understanding perceptions at this location.

A key assumption of this research is that identifying the dominant viewpoint is essential to quantifying the on-site perception, and the characteristics of the locations, such as road width and distribution of surrounding POIs, determine such views. The spaciousness of the street may give visitors a sense of openness and a positive impression of the place. At the same time, the appearance of the buildings in the surrounding POIs (e.g., temples, luxury boutiques, prison-like school buildings, cheap pubs) may attract visitors' attention. Existing studies in environmental psychology have demonstrated to some extent these associations by showing that the openness of the landscape affects the movement of the gaze [7] and that geographical features, such as surrounding buildings, also impact perception [48]. However, those studies used a uniform criterion for determining viewpoints across all locations [8, 24, 47]; none of them identified the dominant views affecting perception based on the characteristics of each site.

To address this issue, we propose a novel framework, called Omni-CityMood, to quantify on-site perceptions through two tasks: (1) modeling the complex interaction in different modalities, i.e., vision-based features of landscapes and geographic context features, and (2) selecting the crucial viewpoints from 360-degree landscapes on the basis of the interaction. In order to select the crucial viewpoints on the basis of the modeled interactions, we propose the new attention mechanism which can integrate information on the surrounding environment.

We demonstrate that our framework contributes to the analysis of the characteristics of urban areas by leveraging the both the atmosphere of the landscape and a human viewpoint perspective . The results of an evaluation show that the Omni-CityMood framework, which explicitly weights dominant viewpoints using attentions, is able to estimate the atmosphere of the landscape more closely to the on-site feeling than other visual urban perception methods.

Our contributions are summarized as follows: (1) We propose Omni-CityMood to quantify the quality of the atmosphere perceived on-site in urban areas by explicitly modeling the importance of each landscape viewpoint based on vision-based features. (2) We utilize the technology of neural recommender systems to bridge the gap between the perception of images and the 'real' environment and present a novel potential application of the existing techniques. (3) We validated the effectiveness of the Omni-CityMood framework on landscape data scored by users.

2 RELATED WORK

2.1 Visual Urban Perception

This research field aims to reveal the connection between the visible appearance of cities and human perceptual responses by using computer vision technologies [5, 19, 27]. Dubey et al. [6] created a large-scale dataset for this field, Place Pulse 2.0, that evaluates the six different perceptions of landscape images collected from Google Street View (GSV). This dataset includes perceptual scores acquired using an online crowdsourcing tool based on pairwise comparison. Here, researchers have converted the pairwise rankings into a single numerical score by using the Microsoft TrueSkill algorithm [14]. Many succeeded at accurately quantifying perceptions of images by using hand-crafted features from street-view images [28, 29] or by using more recently developed deep-learning methods [30, 46].

While previous efforts have provided highly accurate quantifications of landscape images, they still suffer when it comes to quantifying perceptions people have in urban environments, i.e., on-site perceptions. The main reason is that no matter how accurately quantified the perception of a landscape image is, if the viewpoint is inappropriate, the score will not match perceptions of visitors to the site. Despite this fact, visual urban perception studies have not placed much importance on considering the viewpoint of the landscape. Some studies obtained landscape images without setting clear criteria [6], while others simply aggregated landscape images from fixed viewpoints [24, 47].

2.2 Neural Recommender System

Research on recommendation has had a long history. For decades, linear methods such as collaborative filtering [33] and matrix factorization [18] were dominant, but more recently, a lot of effort has gone into developing neural network-based recommender systems [43]. Deep learning has made it possible to consider higher-level features and has resulted in methods aimed at learning user and item feature representations [22, 34, 44], as well as methods incorporating multimodal data such as images and texts as additional information [2, 11]. Although their approaches differ, we can say that all of these methods focus on capturing the relationships among different features.

One of the promising approaches is explicitly considering high-order interactions among multiple features. Cheng et al. [3] proposed Wide & Deep, which combines interactions based on shallow

and deep features, and He et al. [12] devised a method that combines deep learning with factorization machines (FM) [32] to enable learning complicated interactions of different features. The essential characteristic of these approaches is the capability of capturing interactions among different types of features.

In this paper, we attempt to use these neural recommendation approaches to capture interactions between perceptions of landscape images and urban geographical characteristics. Attempts to apply neural recommendation techniques to modeling real-world geospatial features are still few in number, but we believe our efforts presented here will bridge the gap between the two fields.

3 OMNI-CITYMOOD FRAMEWORK

3.1 Overview of Omni-CityMood

A schematic diagram of the Omni-CityMood framework is shown in Figure 2. The key idea is to estimate the atmosphere at a target location not only from a fixed viewpoint, but also from a 360-degree perspective. Specifically, we determine an 'on-site urban atmosphere score' by determining the contribution of each of the multiple viewpoints to the atmosphere.

The atmosphere score feature is based on multiple viewpoints that capture the entire appearance of the target location. Here, images multiple viewpoints are generated from Google Street View by changing the heading of the camera. Then, a CNN-based model is trained by having it calculate the atmosphere score feature for an arbitrary number of images. In addition, geographical context features of the target area, such as the number of surrounding buildings and road width, are obtained from open data sources.

The resulting Omni-CityMood framework uses the atmosphere scores of individual landscape images and the geographical characteristics of the target location to identify critical viewpoints to perceptions and scores the on-site atmosphere. Omni-CityMood uses deep neural recommender techniques to calculate the contribution of each viewpoint by considering interactions between the atmosphere score feature and the context feature. As a result, it can make an evaluation of the atmosphere of a location that close to perceptions that people on-site have.

3.2 Problem Formulation

Our goal is to quantify the on-site atmosphere score at each location in a city. We denote the total number of target locations as L, with each location represented as $l = 1, \ldots, L$. We assume that N landscape images are associated with each location l and define a vector of atmosphere scores of all landscape images corresponding to l, $\boldsymbol{x}_l \in \mathbb{R}^N$. We utilize a CNN model, which is fine-tuned to predict image atmosphere scores, to obtain the score vector $\boldsymbol{x}_l$ at arbitrary locations. Denoting the ground-truth atmosphere score at location l as y_l, the problem of quantifying on-site urban atmospheres can be formulated as a function $f(\cdot)$ that takes $\boldsymbol{x}_l$ as input and infers y_l as follows:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{l=1}^{L} \mathcal{L}(y_l,\ f(\boldsymbol{x}_l; \boldsymbol{\theta})), \tag{1}$$

where $\mathcal{L}$ and $\boldsymbol{\theta}$ are respectively the loss function and parameters used to learn the function $f(\cdot)$.

3.3 Image Atmosphere Score Feature

We train a model for scoring the atmospheres of landscape images by using annotated data. Then, we apply the trained model to individual landscape images to calculate a score. This process is necessary in order to prepare the atmosphere score feature vector $\boldsymbol{x}_l$, which is the input of Omni-CityMood. Note that the method described in this section is not the only one that can be used to evaluate the atmosphere of a single landscape image.

3.3.1 Image Atmosphere Score Dataset. We gathered landscape images from various locations and created a dataset of 2, 305 images that could be used to train the image atmosphere scoring model.

Next, we used an on-line questionnaire platform to annotate each image with an atmosphere score. In particular, we gathered more than 100 evaluations of each landscape image from an unspecified number of users. Specifically, the users had been shown the landscape image and were asked to rate the location on a 5-point scale: "very bad," "bad," "neutral," "good," or "very good". Each user responded with the option that best fit his/her impression in a single-choice format. We converted these user ratings into numerical values by assigning scores ranging from −2 to +2 in order of lowest (i.e., "very bad") to highest rating (i.e., "very good"). Score annotations were made by calculating the user's response rate for each evaluation grade. We obtained a five-dimensional ground-truth score distribution vector that represented the quality of the atmosphere for each of the 2, 305 landscape images in the dataset.

The annotation process was conducted using the crowdsourcing platform, where 100 ordinary participants from the general public evaluated the impression scores for each image. During the evaluation, each participant was assigned a masked ID to ensure that no personal information was disclosed.

3.3.2 Image Atmosphere Scoring Model Development. To obtain atmosphere scores for arbitrary street view images, we constructed a model to quantify the atmosphere of a landscape image and trained it on our dataset. Specifically, we utilized MobileNet [15] pre-trained by ImageNet [4] as a feature extractor. Next, the extracted image features were passed to two fully-connected layers to infer the atmosphere scores of the images. The number of neurons in the final layer was set to five, the same as in the crowdsourcing evaluation, and a softmax function was used to turn the score into a distribution of the score from −2 to +2.

To evaluate the performance of the scoring model, we used two standard evaluation metrics, mean absolute error (MAE) and Pearson linear correlation coefficient (PLCC). Since both the ground truth and the estimated value are distributed over a five-point scale, we converted each of them into scalar values by taking the weighted average of the evaluation scores (i.e., $-2 \sim +2$) multiplied by the voting rate of the users. We evaluated the model's performance by using 5-fold cross-validation, resulting in an MAE of 0.255 ± 0.013 and a PLCC of 0.755 ± 0.016. We concluded that the image atmosphere scoring model performs well enough to use it for estimating the atmosphere of landscape images.

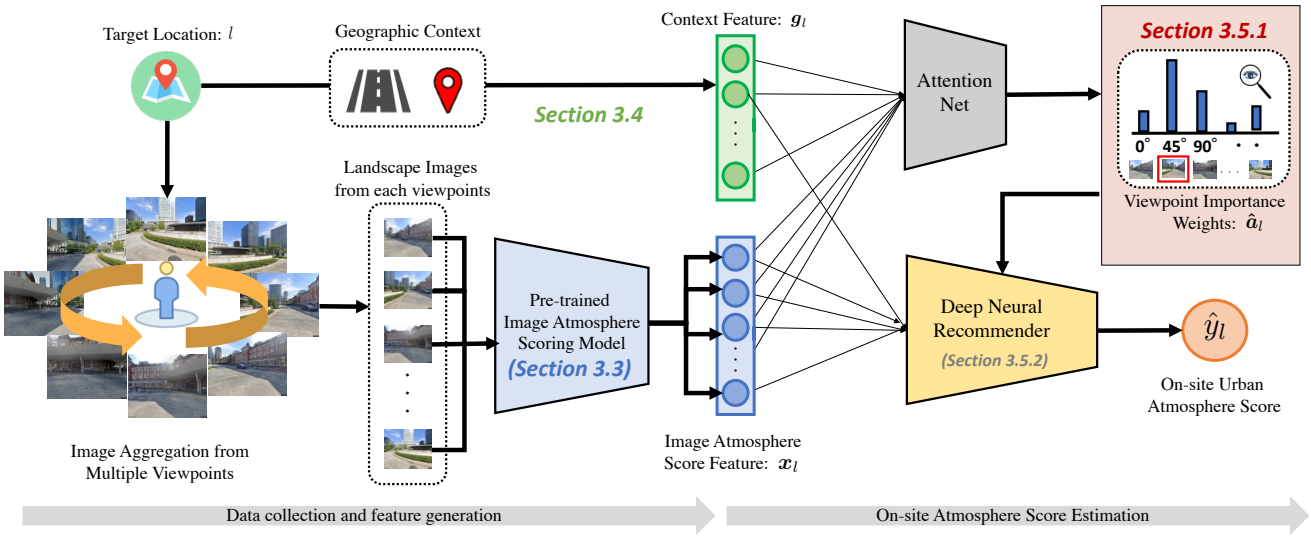


Figure 2: Overview of the Omni-CityMood framework. Omni-CityMood generates two types of features. Then, it gives atmosphere scores toeach viewpoint and geographical context by using an attention mechanism to identify the essential views. After that, it infers an on-site atmosphere score from the two features and essential viewpoint weights.

3.4 Geographical Context Feature

In addition to the atmosphere score vector $\boldsymbol{x}_l$, Omni-CityMood requires the geographical context feature of the location l. Here, we used open-source road centerline data [1] and POI data provided by Foursquare[2]. We assumed that contextual information plays an important role in viewpoint selection and landscape evaluation.

We generated geographic features based on two different aspects: road width and distribution of neighborhood POIs. As for the road feature, we assumed that there are C_r predefined road-width categories. We defined a road feature vectors of the location l as $\boldsymbol{g}_l^{(r)} \in \mathbb{R}^{C_r}$ based on one-hot encoding by assigning 1 to the road-width category to which the target location belongs. As for the POI feature, we generated a feature vector based on the category ratio of the POIs in the vicinity of the target location l. Specifically, we defined the total number of POI categories as C_p and represented the POI vector $\boldsymbol{g}_l^{(p)} \in \mathbb{R}^{C_p}$ on the basis of the POI category ratio within a certain distance from the target location l. Finally, we created a geographical context vector for each location l by concatenating the two different aspect vectors: $\boldsymbol{g}_l = [\boldsymbol{g}_l^{(r)}, \boldsymbol{g}_l^{(p)}] \in \mathbb{R}^C$, where $C = C_r + C_p$ represents the total number of dimensions of the geographical context feature vector.

3.5 On-site Atmosphere Score Estimation

Omni-CityMood framework simultaneously identifies critical viewpoints that significantly impact human perception and the inference of on-site atmosphere scores via multi-task learning schema. Specifically, Omni-CityMood takes the atmosphere score of each viewpoint and the geographical context of the location as input and infers the importance of each view based on an attention mechanism [39].

[1] https://cyberjapandata.gsi.go.jp/
[2] https://sites.google.com/site/yangdingqi/home/foursquare-dataset

At the same time, the interaction between the atmosphere scores and geographic context is modeled using a neural factorization machine (NFM) [12], a recently developed neural recommendation method. The input features of the NFM are adjusted using the attention weights to explicitly account for each landscape viewpoint's importance.

3.5.1 Viewpoint Importance Weight Calculation. In this section, we describe the module of Omni-CityMood for learning the importance of each viewpoint of a landscape. To explicitly learn the importance of viewpoints, we first create ground-truth labels $\boldsymbol{a}_l \in \mathbb{R}^N$ that represent the degree to which each viewpoint should be attended. In general, it is a non-trivial problem to identify a single correct viewpoint, and it is challenging to create clear criteria to assess the quality of each view. For these reasons, we determined the importance of each viewpoint on the basis of the closeness of the on-site atmosphere score and the image-based atmosphere score of each viewpoint. In particular, each element $a_{l,i}$ of the vector $\boldsymbol{a}_l$ is defined as

$$a_{l,i} = \frac{\exp(-s_{l,i})}{\sum_{j=1}^{N} \exp(-s_{l,j})}, \quad \text{where } s_{l,i} = \frac{|y_l - x_{l,i}|}{\lambda}. \tag{2}$$

Here, λ denotes the temperature parameter of the softmax function. Hence, the index of a viewpoint whose score is close to (far from) the on-site atmosphere y_l is given a larger (smaller) value.

In Omni-CityMood, $\boldsymbol{a}_l$ is inferred by an attention module using a geographic context vector $\boldsymbol{g}_l$ and an atmosphere score vector $\boldsymbol{x}_l$ as input. Specifically, each element of $\boldsymbol{g}_l$ is embedded into a dense vector and concatenated with the atmosphere score vector $\boldsymbol{x}_l$. The inference vector $\hat{\boldsymbol{a}}_l$ is output after two fully-connected layers. By

incorporating a context-aware attention mechanism based on a geographical feature vector, our method can infer critical viewpoints relevant to human perception for each location.

3.5.2 Neural Recommendation-based On-site Atmosphere Quantification. We describe the module of Omni-CityMood for estimating the atmosphere score in urban environments. Omni-CityMood multiplies the inferred attention weight vector $\hat{\boldsymbol{a}}_l$ of each viewpoint with the input atmosphere score vector $\boldsymbol{x}_l$. This operation enables the importance of each viewpoint to be incorporated into the inference procedure.

To consider the relations between image-based atmospheres and the geographical context, Omni-CityMood captures the interaction between the weighted atmosphere score vector and the geographic context based on NFM [12]. To capture high-order interactions between input features, NFM combines factorization machines [32] and a multilayer perceptron (MLP). Here, we define a new feature vector combining the weighted atmosphere score vector and geographic context: $\boldsymbol{v}_l = [\boldsymbol{g}_l,\ \boldsymbol{x}_l \odot \hat{\boldsymbol{a}}_l] \in \mathbb{R}^{C+N}$, where $\odot$ denotes the element-wise product of vectors. The inference of the on-site atmosphere score based on the NFM using this joint vector is formulated as follows:

$$\hat{y}_l = \boldsymbol{w}^\top \boldsymbol{v}_l + b + h(\boldsymbol{v}_l), \tag{3}$$

where $\boldsymbol{w} \in \mathbb{R}^{N+C}$ denotes the parameter vector and b the bias term of the model. Note that the first two terms of Equation 3 are identical to linear regression of the input feature vector, and the function $h(\cdot)$ accounts for the interaction between features. The interaction is modeled in $h(\cdot)$ by applying an embedding layer to each input feature and calculating the element-wise product between pairs of embedded vectors . Specifically, we define the embedding vector for the i-th element of the vector $\boldsymbol{v}_l$ as $\boldsymbol{e}_{l,i} \in \mathbb{R}^k$. Accordingly, the feature interaction operation is performed as follows:

$$h(\boldsymbol{v}_l) = f_{\mathrm{MLP}}\left(\sum_{i=1}^{N+C}\sum_{j=i+1}^{N+C} v_{l,i}\,\boldsymbol{e}_{l,i} \odot v_{l,j}\,\boldsymbol{e}_{l,j}\right), \tag{4}$$

where $f_{\mathrm{MLP}}(\cdot)$ represents the MLP module, which is applied to the feature vector obtained through the FM module. In the element product calculation, the embedded vector $\boldsymbol{e}_{l,i}$ is multiplied by the corresponding element value $v_{l,i}$ to explicitly consider the real values of the input features [32].

Our method captures the interaction between the image-based scores and the geographical characteristics by concatenating the atmosphere score vector with the context vector in the FM module. Furthermore, it attempts to automatically learn effective ways to utilize geographical context in the neural recommendation through the embedding layer. While some advanced techniques for region embeddings have been proposed [40, 49], our method integrates geographical context in a suitable manner for neural recommendation by automatically learning embeddings through interactions between features.

3.5.3 Model Training Schema. Omni-CityMood attempts to simultaneously infer $\hat{y}_l$ and $\hat{\boldsymbol{a}}_l$ by using a multi-task learning framework. The training with a multi-task schema is formulated as follows:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{L}\sum_{l=1}^{L}\left\{\ \alpha\mathcal{L}_y(y_l,\ \hat{y}_l) + (1-\alpha)\mathcal{L}_a(\boldsymbol{a}_l,\ \hat{\boldsymbol{a}}_l)\right\}, \tag{5}$$

where $\mathcal{L}_y(\cdot)$ represents the loss function used to infer $\hat{y}_l$ and $\mathcal{L}_a(\cdot)$ denotes the loss function used to infer $\hat{\boldsymbol{a}}_l$, respectively. $\alpha \in [0, 1]$ is a hyperparameter, which adjusts the weights given to the loss function in different tasks.

4 DATASETS

4.1 Two types of Feature Data Creation

We developed a dataset to train Omni-CityMood. Here, we selected $L = 120$ target locations from various cities in Japan. We acquired a total of $N = 8$ landscape images from street views of each location (acquired from Google Street View) by changing the viewpoint in steps of 45 degrees. We set the starting viewpoint to be parallel to the road and facing north, the field of view to 90 degrees, and the pitch to horizontal (0 degrees). By applying the image atmosphere scoring model described in Section 3.3.2 to each landscape image, we created an input atmosphere score feature vector $\boldsymbol{x}_l \in \mathbb{R}^8$ corresponding to each location. As for the geographical context feature, we used open-source road centerline data and POI data provided by Foursquare. The road centerline data included a total of $C_r = 5$ categories according to the road width, and the POI data referenced route categories consisting of $C_p = 10$ categories.

4.2 Creation of Ground Truth Labels

Our goal is to evaluate the perceived quality of the atmosphere in an on-site situation; however, collecting large amounts of perception data on-site is costly. Therefore, we looked for a relatively simple approach that could adequately evaluate the perceptual experience on-site in order to acquire a ground-truth perception label for each location. Here, we could not accept any gaps between the evaluation scores of on-site landscapes and those of the simple approach.

To determine a simple but adequate approach, we first collected an atmosphere score dataset of urban landscapes by using three different approaches: fieldwork, a 360-degree image viewer, and VR goggles; then, we quantified the quality of the dataset obtained by each method and compared them. Here, a fieldwork survey evaluates the perceived atmosphere by having persons visit the location, whereas the image viewer and VR goggles attempt to convey visual information about the on-site experience by utilizing 360-degree panoramic images. Image viewers allow users to freely change their viewpoints by dragging a mouse, while VR goggles allow users to change their viewpoints by moving their heads.

We recruited five subjects and selected a total of 50 locations among several cities in Tokyo. Each participant attended a survey lecture that presented the three approaches in the following order: image viewer-based, VR goggles-based, and field survey. This was done to mitigate the influence from impressions formed during one survey on the results of other surveys. The subjects were required to evaluate the perceived quality of the atmosphere of the landscape on a five-point scale, in the same manner as the annotation of landscape images described in Section 3.3. The atmosphere scores were calculated by taking a weighted sum of the responses from all subjects for each rating on a five-point scale from $-2 \sim +2$.

Table 1: Results of no-correlation test.

Combination	PLCC	p
Image viewer & VR	0.846	1.08×10^{-14}
Image viewer & Field survey	0.743	6.54×10^{-10}
VR & Field survey	0.726	2.37×10^{-9}

Table 2: Results of two-way ANOVA.

Source of validation	SS	df	MS	F	p
Factor locations	256.19	49	5.228	7.666	0.000
Factor approaches	0.275	2	0.137	0.201	0.817
Interaction	47.992	98	0.490	0.718	0.979
Error	409.2	600	0.682	-	-
Total	713.655	749	-	-	-

We conducted a statistical test of no correlation and an analysis of variance (ANOVA) on the scores for 50 landscape atmosphere ratings obtained by the three different methods. In the no correlation test, we calculates the correlation coefficient between the scores of the 50 locations in the three different approaches. The two-way ANOVA considered two factors: the location of the data and the approaches used to collect the data.

Table 1 and Table 2 show the results. The no-correlation test results show high correlations for all combinations of image viewers, VR goggles, and field surveys. The correlation hypothesis is rejected at the 0.01 level of significance. This means that we can conclude that a correlation exists in scores aggregated through different approaches. Table 2 shows that there is a significant difference in the means with respect to the factor of location ($p < 0.01$). In contrast, there are no significant differences in the data-collection approaches or the interaction of the two factors. These results tell us that there are significant differences in the atmosphere scores depending on the location, but no significant difference in the atmosphere score that is due to the data aggregation approach.

Accordingly, we concluded that a survey using an image viewer or VR goggles with 360-degree images would provide good quality data sufficiently similar to on-site sensations. We chose to collect data with an image viewer, which has the lowest labor cost of the three methods. Finally, we attached a ground-truth atmosphere score y_l to each location in our dataset by administering a questionnaire using 360-degree images. As in the procedure above, the score y_l was calculated as a weighted average of vote percentages on a five-point scale.

5 EXPERIMENTS

We conducted evaluation experiments to answer the following research questions regarding the proposed method.

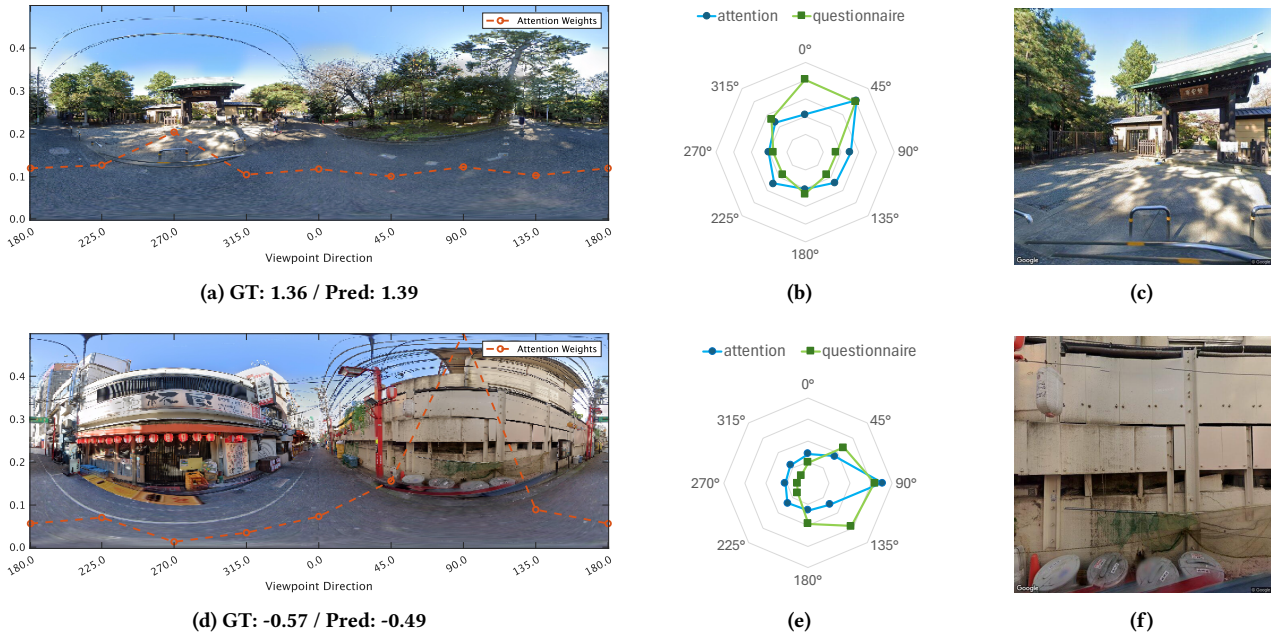
- RQ1: Can the Omni-CityMood framework properly evaluate on-site atmospheres?
- RQ2: Does the Omni-CityMood framework properly select the dominant viewpoints?
- RQ3: How important is the number of viewpoints when evaluating landscapes?
- RQ4: Do components of the Omni-CityMood framework contribute to overall performance improvement?

5.1 Experimental Settings

5.1.1 Compared Methods. We compared Omni-CityMood with baseline methods that can be categorized into two groups: (1) inference methods based on landscape viewpoint selection criteria and that have been used in studies in the visual urban perception domain and (2) the recently proposed neural recommender systems.

- **Random [6].** It follows the procedure used in Dubey et al. [6] that obtains a landscape image from each location without any specific criteria to decide the viewpoint. In the experiment, we randomly selected one viewpoint and referred to the corresponding score as the predicted value.
- **Average All [26, 47].** It predicts the atmosphere by taking the average of the scores of all viewpoints. This procedure is based on ones in studies that calculate the average scores in specific areas, such as road networks [47] and administrative divisions [26].
- **Perpendicular [8].** It predicts the atmosphere by taking the average scores corresponding to the viewpoints perpendicular to a street, i.e., at 90 and 270 degrees. This procedure was used in [8], which estimated the chances of crimes occurring in neighborhoods from their landscape images. The study suggested that images taken perpendicular to the street level are of help in understanding the characteristics of the landscape because buildings will show up often in them.
- **Wide&Deep [3].** This is a recommendation system that combines the strengths of linear models and neural networks. The "wide" component is a linear model that allows the model to learn explicit feature interactions, while the "deep" component is a neural network that learns implicit feature interactions. We only used raw features for the wide part, referring to the evaluation protocol used in the work of He et al. [12].
- **NFM [12].** Since Omni-CityMood incorporates the NFM as a module, we can interpret that NFM is the same as Omni-CityMood without the mechanism for learning weights to be assigned to each viewpoint.
- **Attentional Factorization Machines (AFM) [45].** This method combines the attention mechanism with NFM to consider the importance of each feature interaction. It should be noted that in contrast to AFM, Omni-CityMood explicitly learns weights for viewpoints rather than each interacted feature.

5.1.2 Model training settings. We used Adam [17] with a learning rate of 1e-3 and a learning rate decay of 1e-5 to optimize the neural recommendation method. We trained the models through early-stopping schema with a maximum of 100 epochs. The comparison baseline methods used loss functions based on the mean squared error, while Omni-CityMood used a loss function based on both the mean squared error,$\mathcal{L}_y(\cdot)$, and the Kullback Leibler (KL) divergence $\mathcal{L}_a(\cdot)$. The hyperparameter that adjusts the weights of the two loss functions was set to $\alpha = 0.75$. The temperature parameter was set to $\lambda = 0.2$.



(a) GT: 1.36 / Pred: 1.39 (b) (c)

(d) GT: -0.57 / Pred: -0.49 (e) (f)

Figure 3: Visualization of weights for each viewpoint of the landscape. Left (a, d): predicted attention weights overlaid on the panoramic image. Center (b, e): radar charts displaying the weights acquired from attention and questionnaire for each viewpoint. Right (c, f): landscape image at the viewpoint where the predicted weight shows the largest value.

5.1.3 Evaluation Protocol. As in the evaluation for the image atmosphere scoring model, each method was evaluated with the mean absolute error (MAE). The entire dataset was divided up into training and validation data by using 5-fold cross-validation. The models were compared on the basis of each fold's mean and standard deviation.

5.2 RQ1: Can the Omni-CityMood framework properly evaluate on-site atmospheres?

Table 3: Performance of atmosphere evaluation in each method.

Method	MAE
Random [6]	0.739 ± 0.061
Average All [26, 47]	0.719 ± 0.146
Perpendicular [8]	0.746 ± 0.055
Wide & Deep [3]	0.332 ± 0.044
NFM [12]	0.334 ± 0.026
AFM [45]	0.326 ± 0.053
Omni-CityMood (ours)	**0.320 ± 0.045**

Table 3 compares the overall performance of Omni-CityMood with those of the baseline methods in the visual urban perception domain and neural recommendation field.

The results show that viewpoint selection methods in the visual urban perception domain performed poorly. This indicates that approaches relying on random selection or fixed viewpoints at any location do not adequately capture on-site perceptions. The selection criteria for landscape viewpoints significantly impacted the final inference accuracy.

The neural recommendation-based methods outperformed the simple viewpoint selection criteria. This result is believed to be due to these methods' considering higher-order interactions between features.

Omni-CityMood outperformed all of the baselines. This result suggests that it is important to consider each viewpoint in addition to the interactions between features towards on-site human perception prediction. Omni-CityMood's superior performance to even NFM and AFM highlights the importance of explicitly modeling the importance of viewpoints in the landscape.

5.3 RQ2: Does the Omni-CityMood framework properly select the dominant viewpoints?

Omni-CityMood explicitly learns the importance of each viewpoint by using an attention mechanism. Here, we evaluate the appropriateness of the estimated viewpoints.

Figure 3(a) and Figure 3(d) show the results of visualizing the weights inferred by Omni-CityMood by overlaying them on a panoramic image.

Figure 3(b) and Figure 3(e) compare the calculated viewpoint importance weight and the results of a questionnaire on the importance of viewpoints. The online questionnaire was administered to 100 subjects. The subjects first viewed a 360-degree image at each location via Google Street View. We provided the URL of Google Street View[3] to the subjects, and asked them to click it and see

[3]E.g., https://maps.app.goo.gl/qksLDYPrfAKA8M1w8.

Figure 4: Example of a 360-degree image on Google Street View. The place of the picture is the same as the one shown in Figure 3(a).

(a) **(b)**

Figure 5: Examples of randomly selected viewpoints from among eight different directions. (a) and (b) show the views from two different viewpoints in the same place as Figure 4.

all directions of the scrollable image as shown in Figure 4. Next, we asked the subjects to see the view from a randomly selected viewpoint from among eight different directions in the same place as the 360-degree image, as shown in Figure 5. Finally, the subjects rated on a 4-point scale whether the scenery from the displayed viewpoint was representative of the atmosphere at that location. The average score is shown in the figure.

Figure 3(c) and Figure 3(f) depict the landscape image at the viewpoint where the model has the maximum weight. The upper figure shows an example where the atmosphere score is significantly positive, while the lower figure shows an example where the score is significantly negative.

These figures confirm that a significant weight is given to viewpoints with good (bad) scenery in a place with a good (bad) atmosphere (Figure 3(c)). For instance, the viewpoint shown in Figure 3(f) includes garbage, and we can clearly understand why the atmosphere is negative. Moreover, the calculated viewpoint importance is similar in trend to the viewpoint importance scores obtained from the questionnaires of 100 subjects.

These results confirm that Omni-CityMood contributes to selecting appropriate viewpoints for each location based on the city's characteristics. Moreover, when intuitively satisfactory viewpoints are selected, the model's inference accuracy is good, indicating the importance of different viewpoints in understanding on-site perceptions correctly.

5.4 RQ3: How important is the number of viewpoints when evaluating landscapes?

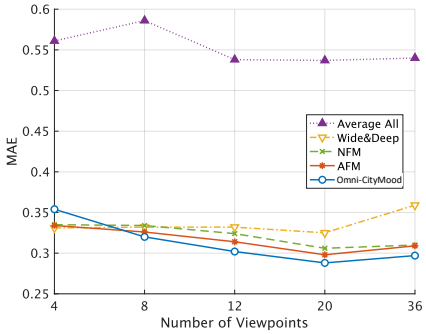


Figure 6: Performance comparison of each method *w.r.t.* different number of viewpoints.

The fundamental idea behind Omni-CityMood is to acquire landscape images from multiple viewpoints and perform inference by considering the importance of each viewpoint. Figure 6 shows how the performance of each method changes with the number of input viewpoints (4, 8, 12, 20, and 36 directions).

Overall, all methods tend to perform better with more viewpoints. This is because more viewpoints provide a more detailed understanding of the overall urban landscape, which supports the hypothesis that human perceptions about a place cannot be evaluated from a single landscape image.

On the other hand, performance leveled off at 12 or more viewpoints. One possible reason for this is the duplication of image content across multiple viewpoints, as well as the increased difficulty in identifying important viewpoints. From this result, we can understand that a model's performance will be highest when the state of the landscape in all directions can be grasped, but that the viewpoints should not have much overlap.

5.5 RQ4: Do components of the Omni-CityMood framework contribute to overall performance improvement?

The Omni-CityMood framework primarily comprises three modules: geographical context feature (Section 3.4), viewpoint importance weight calculation with an attention mechanism (Section 3.5.1), and neural recommendation-based on-site atmosphere quantification (Section 3.5.2). We evaluated the contributions of these modules via an ablation study, comparing full Omni-CityMood to its variants without each component as follows:

- **w/o GCF**: This model did not use the geographical context feature (GCF) $\boldsymbol{g_l}$ (Section 3.4). In this case, the model identified the dominant viewpoints and predicted the on-site atmosphere using only landscape images from each viewpoint.
- **w/o VIWC**: This model did not use the viewpoint importance weight calculation (VIWC) based on an attention mechanism

(Section 3.5.1). In this case, the model ignored the dominance of specific viewpoints in on-site atmosphere quantification.
- **w/o NFM**: This model did not employ a recommender-system-based model (i.e., NFM) for on-site atmosphere quantification (Section 3.5.2). Instead of the NFM, the model used a simple two-layer perceptron with ReLU activation.

Table 4: Ablation study of Omni-CityMood.

Method	MAE
Omni-CityMood	**0.320 ± 0.045**
w/o GCF	0.330 ± 0.044
w/o VIWC	0.334 ± 0.026
w/o NFM	0.330 ± 0.057

Table 4 compares the performance of Omni-CityMood with that of its variants. Specifically, the full Omni-CityMood model achieves the lowest MAE. By removing the geographical context feature (*w/o GCF*), MAE increases substantially compared to the full model, confirming that contextual information obtained from geographical data plays an important role in landscape evaluation.

Similarly, excluding the viewpoint importance weight calculation via an attention mechanism (*w/o VIWC*) yields degraded performance. This result indicates that the model without *VIWC* cannot properly identify dominant viewpoints, resulting in inaccurate atmosphere quantification.

Finally, the *w/o NFM* variant, where the recommender-system-based NFM is replaced with a simple two-layer perceptron with ReLU activation, also shows increased MAE, indicating that NFM effectively captures complex feature interactions necessary for on-site atmosphere quantification. From these results, we can conclude that all the components of Omni-CityMood contribute to performance improvements.

6 DISCUSSION

6.1 Applications of Omni-CityMood Framework

Since the Omni-CityMood is based on street view images for analysis, this framework makes it possible to conduct analyses at specific locations and over entire urban areas.

Figure 7 shows an example of applying Omni-CityMood to a substantially large region of Tokyo, Japan. Figure 7(a) visualizes the predicted atmosphere scores on a map. The blue circles indicate locations with positive scores, while the red circles indicate locations with negative scores, and the size of each circle corresponds to the absolute value of the predicted score. Figure 7(b) presents the direction of the viewpoint for which the weight inferred by the model shows the maximum value at each location. Figure 7(c) and Figure 7(d) are images of the landscape from the selected viewpoints with the maximum and minimum atmosphere scores.

From these results, we can confirm that Omni-CityMood selects visually appealing images that convey the atmosphere of a location. As this example demonstrates, our framework can easily examine the atmosphere of the entire city and viewpoints that are believed to influence human perception. Regardless of whether the atmosphere is good or bad, Omni-CityMood can score the on-site atmosphere by selecting the dominant viewpoints of the atmosphere of a location. Therefore, various applications can be realized using the Omni-CityMood framework, such as the discovery of new spots with good atmosphere, a new tourist information service based on the on-site atmosphere evaluation, and the evaluation of landscapes in urban development.

6.2 Limitation and Future Work

The limitation of our framework is that we have not been able to verify that the landscape scoring model (see Section 3.3) works in various contexts. The scoring model uses the evaluation results for landscape images in sunny urban areas as training data. Therefore, it may not be able to score appropriately for landscape images in non-standard contexts, such as night scenes or snowfall, for example. Also, it will not be able to perform well for street views inside facilities such as museums. To solve this problem, it is necessary to collect landscape images in various contexts as training data to increase the applicability of the scoring model.

7 CONCLUSION

We proposed the Omni-CityMood framework for quantifying on-site urban atmospheres by taking landscape images from multiple viewpoints and considering the importance of each viewpoint. By using domain knowledge of a neural recommender system, our method directly incorporates the importance of viewpoints into the learning process through an attention mechanism. Experiments on a quality-assured dataset showed that Omni-CityMood outperformed existing viewpoint selection strategies of urban perception and neural recommendation methods.

REFERENCES

[1] Sean M. Arietta, Alexei A. Efros, Ravi Ramamoorthi, and Maneesh Agrawala. 2014. City Forensics: Using Visual Elements to Predict Non-Visual City Attributes. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2624–2633.
[2] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized Fashion Recommendation with Visual Explanations Based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.
[3] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & Deep Learning for Recommender Systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. 7–10.
[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[5] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. 2012. What Makes Paris Look like Paris? *ACM Transactions on Graphics* 31, 4 (2012), 101:1–101:9.
[6] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A. Hidalgo. 2016. Deep Learning the City: Quantifying Urban Perception at a Global Scale. In *Proceedings of the European Conference on Computer Vision*.
[7] Lien Dupont, Marc Antrop, and Veerle Van Eetvelde. 2014. Eye-tracking Analysis in Landscape Perception Research: Influence of Photograph Properties and Landscape Characteristics. *Landscape Research* 39, 4 (2014), 417–432.
[8] Kaiqun Fu, Zhiqian Chen, and Chang-Tien Lu. 2018. StreetNet: Preference Learning with Convolutional Neural Network on Urban Crime Perception. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.
[9] Arjan Gosal and Guy Ziv. 2020. Landscape aesthetics: Spatial modelling and mapping using social media images and machine learning. *Ecological Indicators* 117 (2020), 106638.

(a) (b) (c) (d)

Figure 7: (a) Visualization of predicted atmosphere score on a map. (b) Visualization of viewpoint direction with the largest weights on a map. (c) Landscape image of the viewpoint with the maximum weight at the location where the atmosphere score is the largest. (d) Landscape image of the viewpoint with the maximum weight at the location where the atmosphere score is the lowest.

[10] Weili Guan, Zhaozheng Chen, Fuli Feng, Weifeng Liu, and Liqiang Nie. 2021. Urban Perception: Sensing Cities via a Deep Interactive Multi-Task Learning Framework. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17 (2021), 1–20.

[11] Ruining He and Julian McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 144–150.

[12] Xiangnan He and Tat-Seng Chua. 2017. Neural Factorization Machines for Sparse Predictive Analytics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 355–364.

[13] Zhiyuan He and Su Yang. 2018. Multi-View Commercial Hotness Prediction Using Context-Aware Neural Network Ensemble. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018).

[14] Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*, Vol. 19.

[15] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861* (2017).

[16] Aditya Khosla, Byoungkwon An, Joseph J. Lim, and Antonio Torralba. 2014. Looking Beyond the Visible Scene. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), 3710–3717.

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

[18] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.

[19] Stephen Law, Brooks Paige, and Chris Russell. 2019. Take a Look Around: Using Street View and Satellite Images to Estimate House Prices. *ACM Transactions on Intelligent Systems and Technology* 5 (2019), 1–19.

[20] Eva Leslie and Ester Cerin. 2008. Are perceptions of the local environment related to neighbourhood satisfaction and mental health in adults? *Preventive medicine* 47 (2008), 273–8.

[21] Fuzhong Li, K Fisher, Ross Brownson, and Mark Bosworth. 2005. Multi-level modeling of built environment characteristics related to neighborhood walking activity in older adults. *Journal of epidemiology and community health* 59 (2005), 558–64.

[22] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference*. 689–698.

[23] Lun Liu, Elisabete A. Silva, Chunyang Wu, and Hui Wang. 2017. A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. *Computers, Environment and Urban Systems* 65 (2017), 113–125.

[24] Xiaobai Liu, Qi Chen, Lei Zhu, Yuanlu Xu, and Liang Lin. 2017. Place-Centric Visual Urban Perception with Deep Multi-Instance Regression. In *Proceedings of the 25th ACM International Conference on Multimedia*. 19–27.

[25] Weiqing Min, Shuhuan Mei, Linhu Liu, Yi Wang, and Shuqiang Jiang. 2020. Multi-Task Deep Relative Attribute Learning for Visual Urban Perception. *IEEE Transactions on Image Processing* 29 (2020), 657–669.

[26] Marco De Nadai, Radu L. Vieriu, Gloria Zen, Stefan Dragicevic, Nikhil Naik, Michele Caraviello, César Augusto Hidalgo, Nicu Sebe, and Bruno Lepri. 2016. Are Safer Looking Neighborhoods More Lively?: A Multimodal Investigation into Urban Life. *Proceedings of the 24th ACM international conference on Multimedia* (2016).

[27] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L. Glaeser, and César A. Hidalgo. 2017. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences* 114, 29 (2017), 7571–7576.

[28] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César A. Hidalgo. 2014. Streetscore – Predicting the Perceived Safety of One Million Streetscapes. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2014), 793–799.

[29] Vicente Ordonez and Tamara L. Berg. 2014. Learning High-Level Judgments of Urban Perception. In *European Conference on Computer Vision*. 494–510.

[30] Lorenzo Porzi, Samuel Rota Bulò, Bruno Lepri, and Elisa Ricci. 2015. Predicting and Understanding Urban Perception with Convolutional Neural Networks. In *Proceedings of the 23rd ACM International Conference on Multimedia*. 139–148.

[31] Daniele Quercia, Neil Keith O'Hare, and Henriette Cramer. 2014. Aesthetic Capital: What Makes London Look Beautiful, Quiet, and Happy?. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work And Social Computing*. 945–955.

[32] Steffen Rendle. 2010. Factorization Machines. In *2010 IEEE International Conference on Data Mining*. 995–1000.

[33] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. *Collaborative Filtering Recommender Systems*. 291–324.

[34] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. 2015. AutoRec: Autoencoders Meet Collaborative Filtering. In *Proceedings of the 24th International Conference on World Wide Web*. 111–112.

[35] Chanuki Illushka Seresinhe, Helen Susannah Moat, and Tobias Preis. 2018. Quantifying scenic areas using crowdsourced data. *Environment and Planning B: Urban Analytics and City Science* 45, 3 (2018), 567–582.

[36] Yuhan Shao, Yuting Yin, Zhenying Xue, and Dongbo Ma. 2023. Assessing and Comparing the Visual Comfort of Streets across Four Chinese Megacities Using AI-Based Image Analysis and the Perceptive Evaluation Method. *Land* 12, 4 (2023).

[37] Mari Sundli Tveit. 2009. Indicators of visual scale as predictors of landscape preference; a comparison between groups. *Journal of Environmental Management* 90, 9 (2009), 2882–2888.

[38] Mari S. Tveit, Åsa Ode, and Gary Fry. 2006. Key concepts in a framework for analysing visual landscape character. *Landscape Research* 31, 3 (2006), 229–255.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.

[40] Hongjian Wang and Zhenhui Li. 2017. Region Representation Learning via Mobility Flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 237–246.

[41] Wenshan Wang, Su Yang, Zhiyuan He, Minjie Wang, Jiulong Zhang, and Weishan Zhang. 2018. Urban Perception of Commercial Activeness from Satellite Images and Streetscapes. In *Companion Proceedings of the The Web Conference 2018*. 647–654.

[42] James Q. Wilson and George L. Kelling. 1982. *Broken windows*. Critical issues in policing: Contemporary readings. 395–407 pages.

[43] Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. 2023. A Survey on Accuracy-Oriented Neural Recommendation: From Collaborative Filtering to Information-Rich Recommendation. *IEEE Transactions on Knowledge & Data Engineering* 35, 05 (2023), 4425–4445.

[44] Yao Wu, Christopher DuBois, Alice X. Zheng, and Martin Ester. 2016. Collaborative Denoising Auto-Encoders for Top-N Recommender Systems. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 153–162.

[45] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua. 2017. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks. In *Proceedings of the 26th International Joint*

Conference on Artificial Intelligence. 3119–3125.

[46] Yongchao Xu, Qizheng Yang, C. Cui, Cheng Shi, Guangle Song, Xiaohui Han, and Yilong Yin. 2019. Visual Urban Perception with Deep Semantic-Aware Network. In *MultiMedia Modeling*. 28–40.

[47] Fan Zhang, Bolei Zhou, Liu Liu, Yu Liu, Helene H. Fung, Hui Lin, and Carlo Ratti. 2018. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning* 180 (2018), 148–160.

[48] Lemin Zhang, Ruoxi Zhang, and Biao Yin. 2021. The impact of the built-up environment of streets on pedestrian activities in the historical area. *Alexandria Engineering Journal* 60, 1 (2021), 285–300.

[49] Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. 2021. Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*. 4431–4437.

[50] Åsa Ode, Mari S. Tveit, and Gary Fry. 2008. Capturing Landscape Visual Character Using Indicators: Touching Base with Landscape Aesthetic Theory. *Landscape Research* 33, 1 (2008), 89–117.