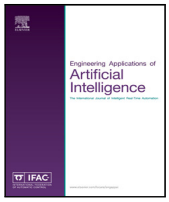
Engineering Applications of Artificial Intelligence 123 (2023) 106294

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Quantitative land price analysis via computer vision from street view images

Chenbo Zhao [a,*], Yoshiki Ogawa [b], Shenglong Chen [a], Takuya Oki [c], Yoshihide Sekimoto [b]

[a] Department of Civil Engineering, the University of Tokyo, 153-8505, Tokyo, Japan
[b] Center for Spatial Information Science (CSIS), the University of Tokyo, 153-8505, Tokyo, Japan
[c] School of Environment and Society, Tokyo Institute of Technology, 152-8550, Tokyo, Japan

ARTICLE INFO

Keywords:
Deep learning
Street view images
Land price estimation
Streetscape impact analysis
Spatial feature distribution

ABSTRACT

Land price is an important economic factor in producing meaningful references for regional planners by assisting them in urban planning, economic decision-making, and land resource allocation. However, related studies in land price analysis were mainly focused on the factors of site area and plot ratio, analysis of the potential impact of streetscape factors and human subjective perception on the land price has been lacking, regardless of the impact on supply and demand relationship. Therefore, this study developed a new approach for estimating and analyzing land prices through deep learning that considered the streetscape and human subjective perception factors. In the estimation part, we developed a fine-grained end-to-end deep learning model, the input is street view images, and the output is land prices. In the analysis part, we extracted the semantic segmentation results and human subjective perception scores from the images and combined them with the results of land price estimation. We then introduced a combination of quantitative analysis using the gradient-weighted class activation mapping and L1-based sparse linear regression to model the relationship between the streetscape and the human subjective perception quantitatively. The gradient-weighted class activation mapping was used to determine which categories of pixels deep learning relied on to output the results of land price estimation quantitatively. Combined with segmentation results, we implemented L1-based sparse linear regression and quantitatively determined the importance of the streetscape factors for land prices. Overall, our deep learning model achieved 77.99 % accuracy by only using street views in estimating the land price, and we illustrated the impacts of the streetscape and perception on the land price by showing that perception scores are more important than streetscape factors such as road and mountain in the streetscape. Comfortable, lived-in feel in perception have the most important impact on land price estimation.

1. Introduction

Land price, which has been considered an important indicator of the extent of regional development and macroeconomic fluctuations (Liu et al., 2013), can provide important information for urban planning, economic decision-making, and land resource allocation (Hu et al., 2016). Due to the promising use and importance of this kind of information, many researchers have developed various approaches to estimate or analyze land prices. Representative approaches include particle swarm optimization (Alfiyatin et al., 2017), linear regression based on hedonic pricing (Ghodsi et al., 2010), and random forest (Hau et al., 2018). Most models have been developed based on geographic information system (GIS) datasets of facilities and land use characteristics. Although several studies have reported associations between land prices and urban landscapes (Gao and Asami, 2007), entrepreneurship, social vitality, environmental quality (Nakamura, 2019), and land policy (Du et al., 2011), none of them has examined the land price associations with urban landscapes' objective information and human subjective perceptions from street views. Especially, perceptions of the streetscape, such as 'desirable for living' and 'safe' that considerably affect the supply and demand relationship, will be considered an important factor for land price estimation. This ranking of the most desirable places in Japan, this relationship affects the land price. However, related studies on land price analysis were mainly focused on the factors of site area and plot ratio, they did not consider that the streetscape and human perception would have an impact on the supply and demand relationship.

This study aimed to propose a fine-grained end-to-end deep learning approach for the estimation of land prices based on street view images and to develop a method to test the relationship between streetscape factors and human subjective perception scores and land prices, in which fine-grained indicates densely covered estimation that includes non-building areas, end-to-end means inputting an image and outputting a land price, without additional data or processing. To analyze the impact of factors on the potential influence on the land

* Correspondence to: Department of Civil Engineering, the University of Tokyo, 4-6-1 Komaba, Meguro, Tokyo 153-8505, Japan.
E-mail address: cbzhao@iis.u-tokyo.ac.jp (C. Zhao).

https://doi.org/10.1016/j.engappai.2023.106294
Received 22 November 2022; Received in revised form 3 April 2023; Accepted 4 April 2023
Available online 17 April 2023

price quantitatively, we proposed the deep learning method to extract people's subjective evaluations and developed L1-based sparse modeling that combines gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017) and a semantic segmentation method. This approach analyzed the scene impact that captures objective pixel class and human subjective perception score distribution characteristics for different land value levels, namely machine teaching. This approach illustrated the mechanism that constitutes the land price by revealing the relationship between land price, streetscape factors, and human subjective perception factors. Land price estimation is more difficult than housing price estimation because all streetscape elements that comprise the urban environment other than buildings affect the land price. The land price analysis should consider areas ranging from rural, non-built areas to urban, densely built areas as the study area. The primary contributions of this study are as follows:

1. We proposed a deep learning approach for end-to-end land price estimation using 2.5 m interval fine-grained street view images. This approach enabled a micro-estimation based on the overall streetscape which is related to the land price constitution.
2. We proposed the human subjective perception score extraction approach and designed a method to evaluate and quantitatively compare the impact of streetscape and perception factors on the land price.

The remainder of this paper is organized as follows: Section 2 reviews related studies; Section 3 describes the source data used in this study; Section 4 describes the methodology; Section 5 describes the experiments and results. Section 6 presents our conclusions.

2. Related work

2.1. Land price estimation and analysis

Thanks to the recent development of machine learning and statistics, researchers have been able to perform more efficient quantitative analysis of the parameters that have a considerable impact on land prices, which has increased the accuracy of land price estimation. Xiaozhu and Ling-wei (2013) proved in their study that the location determines the land price; more specifically, easy access to public (Kisilevich et al., 2013), and recreational facilities (Jim and Chen, 2009) are the location elements. Moreover, studies (Gao and Asami, 2007; Nakamura, 2019; Du et al., 2011) have shown that in addition to location factors, attributes such as urban landscapes, entrepreneurship, social vitality, environmental quality, and land policy should not be ignored when estimating land prices.

Therefore, after manually exploring the possible major factors influencing the land price, and demonstrating the efficiency of these factors, studies were conducted to estimate or predict land prices. Anand et al. (2021) used a regression model based on hedonic price theory to determine a price; Derdouri and Murayama (2020) implemented regression kriging and machine learning algorithms to conduct a comparative study on land price estimation; Alvarez et al. (2022) mined the data which may influence the land price from traffic count, core transit corridors, and parking lot entrances datasets for their tree ensemble machine learning model to estimate the property price; Jiang et al. (2021) proposed to embed sequential temporal features using a transformer and combine them with non-temporal features for real estate modeling; Viana and Barbosa (2021), proposed an attention layer based on a radial basis function kernel for house price prediction. Kang et al. (2021), considered human dynamics and perceptions to create a place-oriented hedonic pricing model, and Qiu et al. (2023) selected subjective and objective factors for a hedonic pricing model.

However, the aforementioned studies mainly focused on facilities or accessibility; landscape and human subjective perception were not considered. Likewise, most studies used text and GIS data and their approaches are difficult to use in many countries, especially in developing countries. Furthermore, in similar housing price studies based mainly on sparse housing transaction data, experiments were generally conducted on sparse individual houses, or they only estimated housing prices on average in large areas. Additionally, analyzing the non-building area is difficult for housing price studies, resulting in the difficulty of a dense analysis covering the entire city. As street view data are available in many countries, an image-based land price model and analysis approach that can densely cover the entire city is urgently required.

2.2. Related deep learning algorithms

In recent years, deep learning algorithms have shown excellent performance in general computer vision (CV) tasks, such as classification (He et al., 2016), segmentation (Huang et al., 2019), and detection (He et al., 2017; Ge et al., 2021), which is far better than traditional algorithms. The backbone of classification networks can be considered as feature extractors that can be transferred or generalized to other domains or downstream tasks such as economics (Yeh et al., 2020), security (Al-Garadi et al., 2020), or even physics (Degrave et al., 2022), etc. Implementing a deep learning algorithm for the land price estimation task also showed promise. In terms of feature extractors, convolutional neural networks (CNNs) have dominated almost all image processing tasks for a long time, far better than multilayer perceptions (MLP) (Botalb et al., 2018; Driss et al., 2017). Since 2020, a new type of deep learning is emerging to replace the CNN regime, namely vision transformers (ViTs) (Dosovitskiy et al., 2020). Accordingly, we will use the representative CNNs and ViTs for our land price estimation task, and make a comparison between them.

Many related studies on housing price estimation exist, Afonso et al. (2019) used text attributes to estimate the housing price. Naser et al. (2020) used interior images, exterior images, satellite images, or a combination of two types of images to estimate the value of a house. Nouriani and Lemke (2022) combined the text attributes with the image information to train the estimation model. Although they showed great performance, their backbone networks are mainly AlexNet (Krizhevsky et al., 2012), ResNet (He et al., 2016), and GoogLeNet (Szegedy et al., 2015), which are limited in the CNN regime from more than five years ago, and they mainly focused on the individual house price, not the land price. The land price and housing price are similar but different things because they depend on different factors, and their importance is also different, such as preliminary engineering cost, civil and erection cost, management cost, expenses of taxation, and interest rate for housing price. Particularly, building design, built year and structure affect housing price a lot; additionally, parcel size, floor area ratio, and land use type affect land price a lot. Thus, the approaches that worked on the housing price may not be relevant to the land price. Yamada et al. (2020) used satellite imagery for estimation, which was also based on ResNet (He et al., 2016). However, the spatial resolution of Yamada's study was 2 m, with 256×256 pixels for each image tile. This means that the prediction results would be mean values for 512×512 m^2 areas, a rather rough result. Moreover, these deep learning-based estimates were mostly limited to the general deep learning workflow for the individual images or data, namely training/validation/testing. They did not consider the information that can be obtained from deep learning models and the deep semantic information in the images. Further, studies analyzing the quantitative relationship between human subjective perception of the input images and land prices are limited. Therefore, a more fine-grained, end-to-end deep learning approach with the quantitative association analysis of street view images and land prices via vision and perception is urgently needed.

3. Source data

This study was conducted in Kōchi, Japan (133.47°E 33.49°N to 133.62°E 33.59°N, The World Geodetic System 1984 (WGS84)) in an

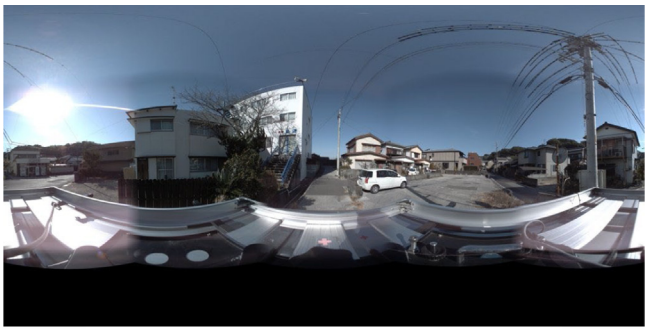


Fig. 1. Example of street view image in this study.

area of $16.96 \times 14.06 = 238.45$ km^2. Kōchi is a city in the south-central Shikoku region of Japan. Since Kōchi city located on the southern coast of Kōchi Prefecture, the street view contained a wide range of scenes from the ocean to the mountains, and from urban to rural area, which is better to conduct this study than metropolis such as Tokyo.

3.1. Street view images data

For this study, 853,268 street view images of Kōchi, Japan, obtained from the Zenrin Corporation (2012 and 2013) were used. Zenrin Corporation collected street view images from all over Japan and provided us part of their image data for this study. The images were taken at 2.5 m intervals using a 360° camera mounted on the roof of a vehicle as it drove along the streets. Each image was annotated and geotagged with textual information such as latitude and longitude, measured using the GPS, vehicle azimuth, and time of capture. The original images were panoramic in the jpeg tar format with 2700 (height) × 5400 (width) pixels, and the resolution was higher than that of Google street view (GSV) images. Since the resolution of Zenrin's images were better than GSV and the data collection time was in specific time, they can better reflect the street feature of that time for this study. The bottom of the images included the roof of the vehicle. Fig. 1 shows an example of the panoramic street view image.

3.2. Road Link-Based land price data

The Road Link-Based land price raw data (1000 JPY/m^2), which was used to estimate the land price adjacent to the road, used in this study was downloaded from the National Tax Agency of Japan.[1] The Road Link-Based land price raw data in Japan can easily to evaluate land price by multiplying road link-based land price by area. Fig. 2 shows several examples of the land price raw data. The key information in Fig. 2 is the numbers in the map, such as '75' in the '75E' indicated 75,000 JPY/m^2. The land price calculation examples are shown in Fig. 2(b) and (c), where the largest road price side was set as front, the front scale is 1.0, and the side and back scale is 0.08 and 0.03, respectively.

4. Methodology

The workflow of the study is shown in Fig. 3, and it can be divided into three parts: dataset creation, land price estimation, and analysis of land price characteristics.

4.1. Dataset creation

The first component of this study was the creation of the dataset, shown in Fig. 3(a).

[1] https://www.rosenka.nta.go.jp/.

Table 1
Summary and memory and compute consumption of land price estimation algorithms.

Network	Params (M)	FLOPs (G)
ResNet-101	44.55	7.85
Swin transformer small	49.61	8.52
Swin transformer base	87.77	15.14
ConvNeXt base	88.59	15.36

As described above, we have two types of source data: street view images, and land price raw data. In general, if we want to use a supervised deep learning approach to estimate the land price based on street view images, we should know the land price ground truth for each image. Thus, we should assign the land price ground truth to the images.

As shown in Fig. 3a, the land price data was initially converted to shapefile format; each road link has a land price property as discussed in Section 3.2. And as discussed in Section 3.1, these street view images also have their coordinates. Therefore, we assigned the land price property of the road link to the images based on the distance between the roads and images, the ground truth of the images was the nearest road link land price property. Furthermore, considering the width of the roads and crossings, the ground truth of the image should be the value of the nearest road unit within 20 m.

4.2. Deep learning land price estimation

The second component was the deep learning land price estimation, shown in Fig. 3(b).

CNN is a trainable feed-forward neural network with a deep structure and convolutional computation, which is a representative algorithm of deep learning (Goodfellow et al., 2016). The CNN feature extraction stage includes at least three types of layers: convolution, pooling, and a non-linear activation function (LeCun et al., 2015). In classification tasks such as in the current study, fully connected (FC) layers act as MLPs connected to the feature extraction stages to perform the final classification task.

Another type of deep learning algorithm, ViT (Dosovitskiy et al., 2020), proposed in 2020, works effectively in many CV tasks and may become better than CNN in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). ViT reached the state-of-the-art, with up to 91.00% accuracy (Yu et al., 2022).

These classification algorithms were trained on over 1 million images labeled across 1000 classes (ILSVRC), they learned the texture and semantic features of this large dataset, which covers most general scenes and objects in our daily life. Therefore, the backbone networks of the classification algorithm could be considered as general feature extractors and transferred to other downstream vision tasks, such as object detection and semantic segmentation. Considering these transfer processes, we proposed an end-to-end approach for estimating land prices by transferring the classification backbone models to directly represent the input street scene and the output land price. Representative state-of-the-art backbones were selected from CNNs and ViTs separately: the base version of ConvNeXt (Liu et al., 2022) (ConvNeXt-B), the small and basic versions of the Swin Transformer (Liu et al., 2021) (Swin-S, Swin-B), along with the set of the old best backbone ResNet as a comparison. The CNNs and ViTs implemented in this study are summarized in Table 1, where the memory and computational requirements were calculated when the input image size was 224×224.

The architectures of Swin transformer and ConvNeXt are shown in Fig. 4.

Swin Transformer (Fig. 4a) first splits an input image $X \epsilon \mathbb{R}^{H \times W \times C}$ into non-overlapping patches using a patch-splitting module, which is similar to the original ViT. Each patch is treated as a "token" and

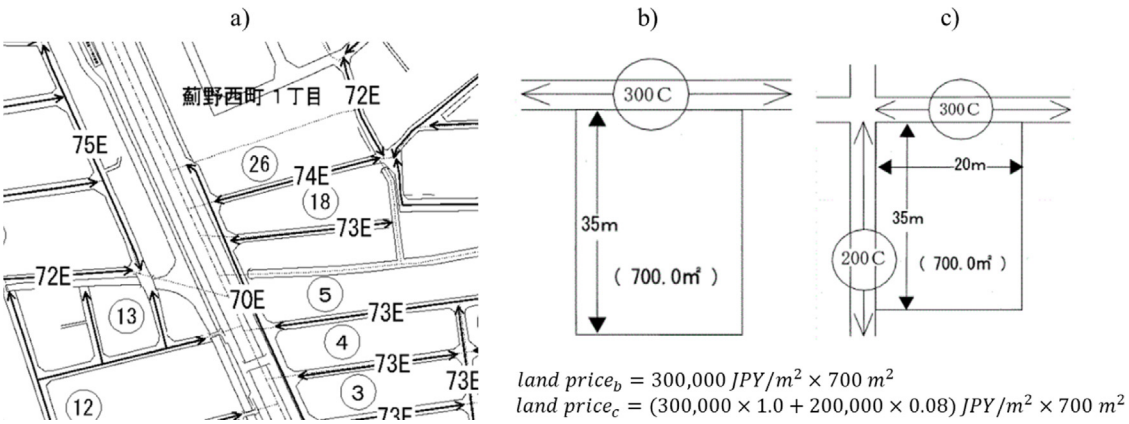


Fig. 2. (a) Examples of the land price raw data: Road link-based data and the number indicated the land price (1000 JPY/m^2), (b) and (c) are the land price calculation examples for one and two sides road.

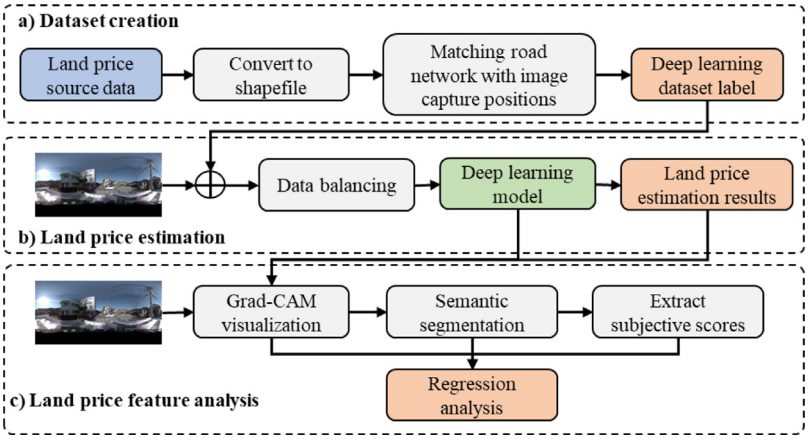


Fig. 3. Workflow of the study: (a) dataset creation; (b) land price estimation; (c) land price characteristic analysis.

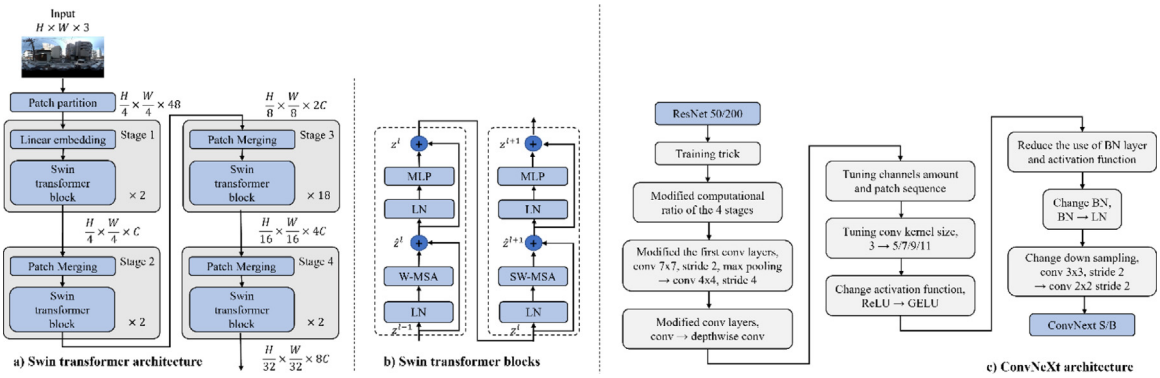


Fig. 4. Schematic of Swin transformer and ConvNeXt.

its feature is set as a concatenation of the raw pixel RGB values. Then, several transformer blocks (Fig. 4b) were responsible for feature extraction. Finally, the extracted feature will be input to the fully connection layers, the output is the land price results.

ConvNeXt is a variant of ResNet, as shown in Fig. 4(c). Considering the same memory and compute consumption, ConvNeXt small and base version are the improvement of ResNet 50 and 200. With the strategies such as tuning convolution kernel size, changed the activation function, changed batch normalization to layer normalization, as illustrated in Fig. 4(c), ConvNeXt achieved the best performance in this study.

4.3. Land price feature analysis

The third component was the Deep Learning Land Feature Analysis, shown in Fig. 3(c).

In this part, we implemented and improved an approach to extract human subjective perceptual values related street views, and applied this approach to the images to provide ancillary data for the land price feature analysis. Then, we applied a semantic segmentation algorithm on the street view image to additionally obtain the pixel level category information, and used the result as ancillary data for analyzing which

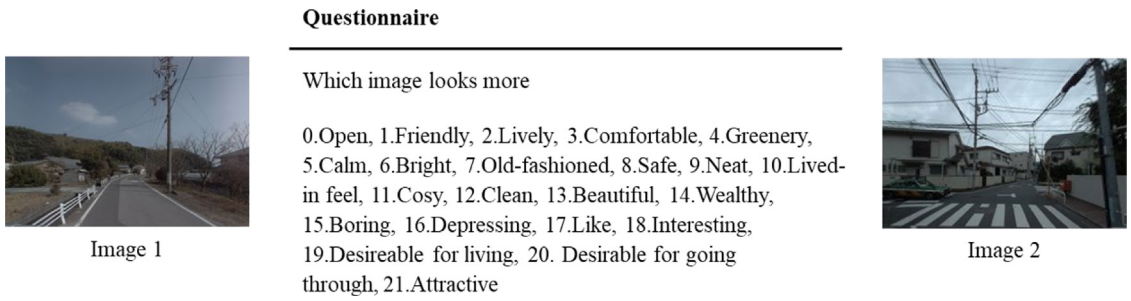


Fig. 5. Survey to construct the subjective score extraction dataset.

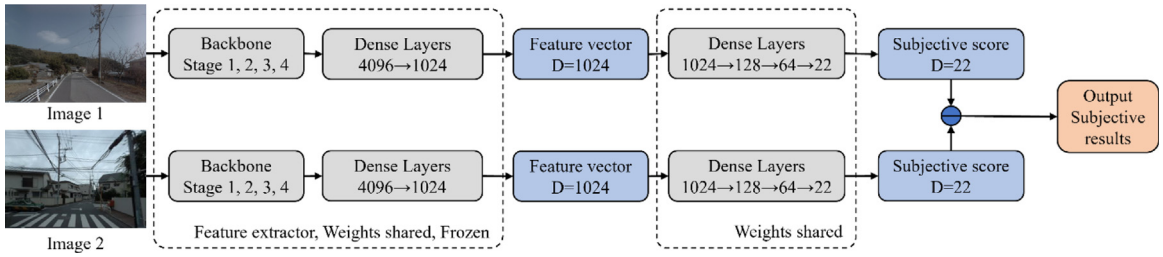


Fig. 6. Schematic of subjective score extraction model.

kind of pixels have a greater impact on the deep learning land price estimation model, based on Grad-CAM.

4.3.1. Extraction of human subjective perception scores

In many studies, subjective scores make an important contribution to the analysis of landscape (Levering et al., 2021), attractiveness (Oki and Kizawa, 2021), and house prices (Xu et al., 2022), which shows the prospects of subjective scores. In this study, we extracted subjective scores to conduct the land price feature analysis from the subjective side.

Human subjective scores are the quantitative emotional responses to street view images. However, the emotional response is difficult to quantify. For the detailed analysis of the deep learning explanation component and improving the performance of the subjective score extraction, our dataset was marginally different from the study of Dubey et al. (2016). We selected 1500 representative panoramic street view images and constructed 14,950 pairs. In each pair, we surveyed all 22 perceptual attributes, shown in Fig. 5, rather than surveying randomly selected items from the six perceptual attributes. We surveyed 40 volunteers for each perceptual attribute of each image pair, using Web base questionnaire survey (Rakuten CO., LTD.), totally 38,525 volunteers were surveyed in this study. Therefore, the total volume of our dataset was $14\,950 \times 22 \times 40$=13,156,000. We divided the 14,950 pairs into training, validation, and test datasets with proportions of 60%, 20%, and 20%, respectively.

14,950 image pairs were gathered, covering all 22 perceptual attributes for each pair. Then, the human subjective score extraction task from one logit value output were converted to a multi-labeled classification. The architecture is shown in Fig. 6.

4.3.2. Semantic segmentation

To obtain detailed information for the model, we implemented the semantic segmentation branch of the unified perceptual parsing network (UperNet) (Xiao et al., 2018) on the street view images to provide pixel-level category information for further analysis.

Since the segmentation model pre-trained on the open street view datasets worked effectively on our dataset without fine-tuning and we need more detailed information about pixel-level categories for further analysis, we chose a state-of-the-art model that was pre-trained on the ADE20K (Zhou et al., 2017) open dataset, which contains 150 segmentation classes, for which Swin-B served as the backbone.

4.3.3. Grad-CAM mechanism and usage in land price analysis

Originally, the Grad-CAM is a method for visualizing the attention of deep learning networks. The output of the Grad-CAM score shows the attention of each pixel, i.e., the information it can provide with respect to the image that has the greatest impact on the results. Therefore, we implemented Grad-CAM and combined the output with our ancillary data on pixel categories to quantitatively analyze the pixel categories that have the greatest influence on land price estimation.

4.3.4. L1-based sparse linear regression

To quantitatively compare and evaluate the impact of the streetscape and human subjective perception factors on the land price estimation, we implemented an L1-based sparse linear regression model on that factors. The process of the regression model is

$$f(x) = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_m x_m + b,$$

$$\omega = \underset{\omega}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \left\| f(X_i) - Y_i \right\|_2^2 + \alpha \|\omega\|_1 \quad (1)$$

where ω and b are the weights and bias of the linear regression model, x is the input 150 classes streetscape pixel proportion and 22 human subjective perception scores, so m is 172. n is the amount of the images, X_i is the ith set of $\{x_1, x_2, \ldots, x_m\}$, and Y_i is the ith land price ground truth α is the constant that multiplies the L1 term.

5. Experiments and results

Our experiment was conducted on the mdx cloud platform (Suzumura et al., 2022), which provides users with flexibility, high security, and the ability to interface with supercomputers and edge devices, through high-performance networks. The proposed deep learning architectures were mainly processed on four NVIDIA A100 40G.

5.1. Dataset construction

As described in Section 4.1, we assigned the land price property of the road link to the images based on the distance between the roads and images, the ground truth of the images was the nearest road link land price property. Thus, 827,343 images were selected from 853,268 images that had the ground truth attached. The rest are unattached images whose distances to the nearest road were more than 20 m. The attached examples and the ground truth distribution over the whole study area are shown in Fig. 7.

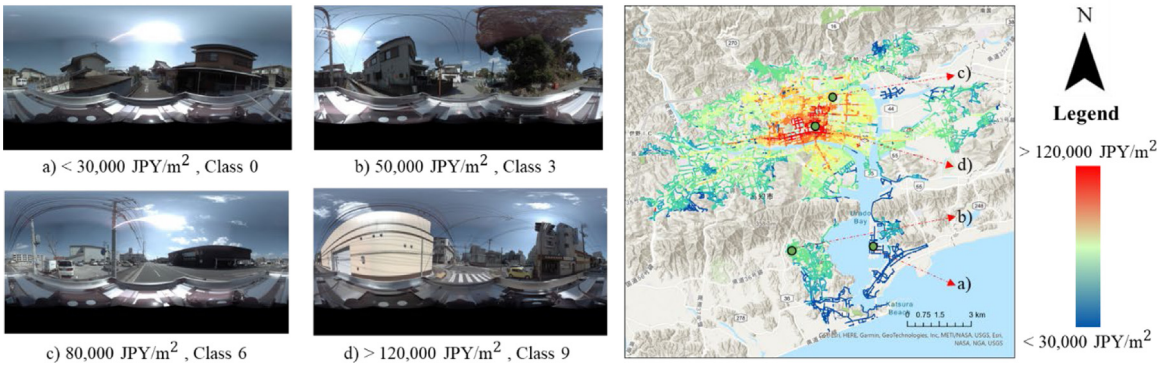


Fig. 7. Examples of street view images with ground truth and ground truth distribution.

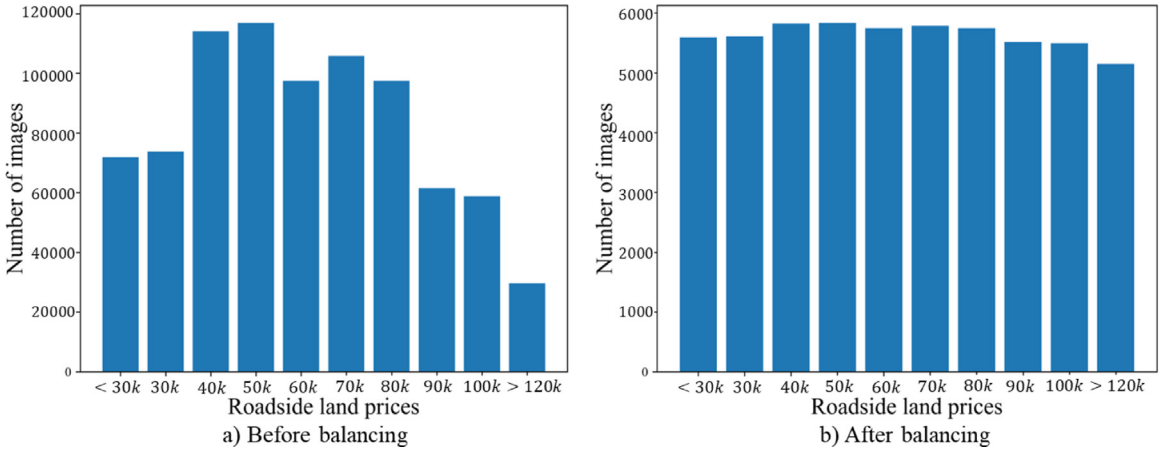


Fig. 8. Land price ground truth distribution: (a) before balancing; (b) after balancing.

5.2. Deep learning land price estimation

To increase the difference between the different land price truths and thus improve the fitting performance of deep learning, we convert the estimation regression task into a classification task by mapping the ground truth to 10 classes. We based the classes 0 to 9 on the ranges of: $(0, 30k], (30k, 40k], \ldots (90k, 100k], (100k, 120k], (120k, \infty)$ JPY/m^2, respectively.

The distribution of the ground truth is shown in Fig. 8(a), the difference between categories could be up to four times, which can be detrimental to deep learning training. Therefore, we performed land price class balancing by random selection and additionally kept the numerical relationship of the categories to ensure the robustness of the model (Fig. 8b). Finally, we selected 56,273 images (6.8% of the total 827,343 images) that formed the deep learning dataset. The 10 land price classes are shown in Fig. 8b, which were further divided into training, validation, and test datasets with proportions of 60%, 20%, and 20%, respectively.

In the deep learning processing, the raw images were first resized to 224 × 224 prior to input. We then applied data augmentation to improve the generalization ability of the model. We also normalized the images to prevent gradient explosion and used random flip (prob = 0.5), random erasing (prob = 0.25 and area = 0.02–0.33) (Zhong et al., 2020), and CutMix and MixUp (prob = 0.5) (Yun et al., 2019) to expand the training dataset. The training strategies were optimized using the AdamW (lr = 0.0005, decay = 0.05, and beta = (0.9, 0.999)) (Loshchilov and Hutter, 2017) optimizer to adaptively optimize the learning rate and prevent overfitting; cosine annealing (warmup_iters = 1000) (Smith and Topin, 2019) was used as the learning rate scheduler. The loss function was adopted from the general cross-entropy loss. Our training was based on the models pre-trained on the ImageNet-1K dataset, rather than starting it from scratch.

Table 2
Land price classification performance of the different experiments.

	Approach	R^2	mF1	Accuracy
Test data	ResNet-101	0.6459	0.4951	0.4934
	Swin-S	0.8346	0.7173	0.7148
	Swin-B	0.8566	0.7672	0.7648
	ConvNeXt-B	**0.8645**	**0.7820**	**0.7799**
Study area	Swin-B	0.8191	0.7648	0.7527
	ConvNeXt-B	**0.8331**	**0.7820**	**0.7706**

We implemented ResNet-101, Swin-S, Swin-B, and ConvNeXt-B to compare the legacy CNN benchmarks with the ViT and CNN state-of-the-art architectures. The models were implemented for only a portion (6.8%) of the total dataset, and precision estimation was performed for the test set. We then selected Swin-B and ConvNeXt-B to perform the comparison against the total data of approximately 0.8 million images. We selected precision, recall, mF1, and accuracy as our evaluation index, and the results are summarized in Table 2 and Fig. 9.

The confusion matrices show that the deep learning model successfully extracts the high-level semantic features and maps the content of the street view images to the land price since the prediction is mostly done on or near the diagonals of the confusion matrices. In addition, we visualized the overall result of the 0.8 million inferences (Fig. 10(a)) and compared it with the ground truth by $(prediction - ground\ truth)$ (Fig. 10(b)).

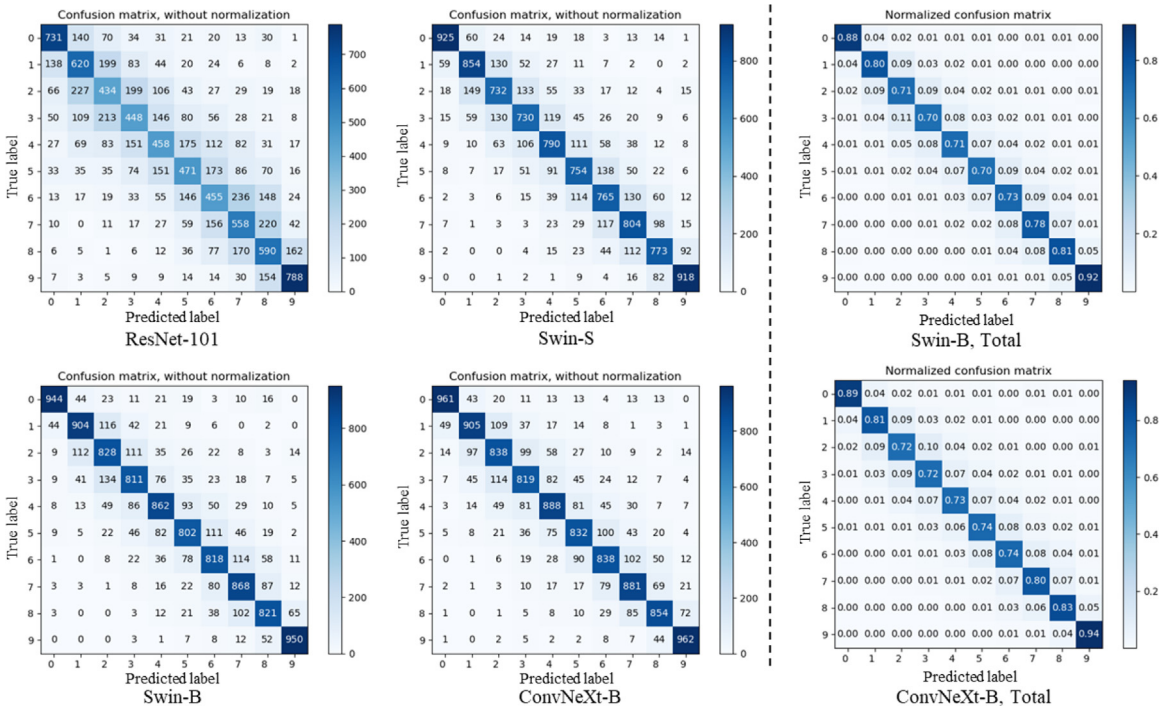


Fig. 9. Confusion matrices of the different land price estimation algorithms.

Table 3
Human subjective perception score extraction siamese network performance comparison.

Setting	Open	Friendly	Lively	Comfortable	Greenery	Calm	Bright	Old-fash.	Safe	Neat	Lived-in f.
VGG-16	0.8763	0.8373	0.8798	0.8455	0.8803	0.8518	0.8623	0.8633	0.8448	0.8675	0.8685
ConvNeXt-B	0.9060	0.8618	0.8957	0.8652	0.9000	0.8815	0.8892	0.8873	0.8615	0.8947	0.8940
ConvNeXt-B, Multi	**0.9112**	**0.8712**	**0.9020**	**0.8842**	**0.9062**	**0.8995**	**0.9082**	**0.9057**	**0.8630**	**0.9137**	**0.9038**
Setting	Cosy	Clean	Beautiful	Wealthy	Boring	Depressing	Like	Interesting	Des. living	Des. go thr.	Attractive
VGG-16	0.8373	0.8603	0.8563	0.8398	0.8105	0.8313	0.8230	0.7590	0.8385	0.8240	0.8233
ConvNeXt-B	0.8585	0.8848	0.8925	0.8608	0.8522	0.8648	0.8587	0.7947	0.8635	0.8665	0.8595
ConvNeXt-B, Multi	**0.8783**	**0.9000**	**0.8928**	**0.8640**	**0.8540**	**0.8730**	**0.8692**	**0.7958**	**0.8725**	**0.8765**	**0.8660**

5.3. Land price feature analysis

5.3.1. Semantic segmentation, Grad-CAM, and human subjective perception score extraction results

In this section, we implemented the subjective perceptive score extraction algorithm discussed in Section 4.3.1. In addition, we implemented the UperNet and Grad-CAM methods for further land price feature analysis. The examples of visualization results are shown in Fig. 11.

The validation result of human subjective perception score extraction is shown in Table 3. We compared three different architectures and used the ConvNeXt backbone with the multi-label head in further analysis. The numbers indicate the accuracy of the output of each branch shown in Fig. 6.

To determine the pixel distribution and human perception scores features of the different land price classes, we calculated the mean value of the top 20 of the ADE20k 150-pixel class proportions and the 22 perception scores corresponding to the 10 land price classes. The normalized results are shown in Fig. 12.

From Fig. 12, we can see the objective and subjective characteristics of each land price class. It can be observed that the areas with very low and high land prices have many natural pixels such as sky and trees and few artificial pixel classes such as buildings and roads. This phenomenon is in contrast to the middle land price classes such as class 5, which is similar to class 9. It can reflect the range of rural areas → concentrated residential areas → senior residential areas. That is, the streetscape influences the land price and vice versa.

Furthermore, the subjective perception scores show a stronger dependence on the land price classes. Essentially, positive perception scores increase when the land price increases; meanwhile, the land price decreases when negative perception scores increase. In particular, Greenery and Calm perceptions can reflect the urbanization degree; thus, an increase in their scores correspond to price decreases. In addition, it is worth noting that the pixel classes that occupy the largest proportion such as sky and building change rapidly between land price classes 4 and 5; thus, the subjective scores have a gap between the corresponding classes 4 and 5.

Next, we analyzed the spatial characteristics of the perception scores in combination with the spatial distribution of the land price in Fig. 7. We calculated the correlation coefficients (Table 4) between land price and perception classes and selected top-3 positively and negatively related scores, and two non-spatially related scores to visualize (Fig. 13). Obviously, the positive perception classes with high scores, such as 9. Neat, 14. Wealthy, and 3. Comfortable overlap with the high land price area, while the negative perceptions with low score like 17. Like, 15. Boring, and 16. Depressing overlap with the high land price area. There are also some perception classes such as 0. Open and 5. Calm, which have little spatial relationship with the land price. Therefore, we can find out which perceptions would have a great influence on the land price, which is a clue for further marketing and urban planning.

Lastly, with the combined analysis of semantic segmentation and human subjective perception score based on the results of land price estimation, we can obtain much information as described above. Conversely, by quantitatively analyzing the results of Grad-CAM, we can

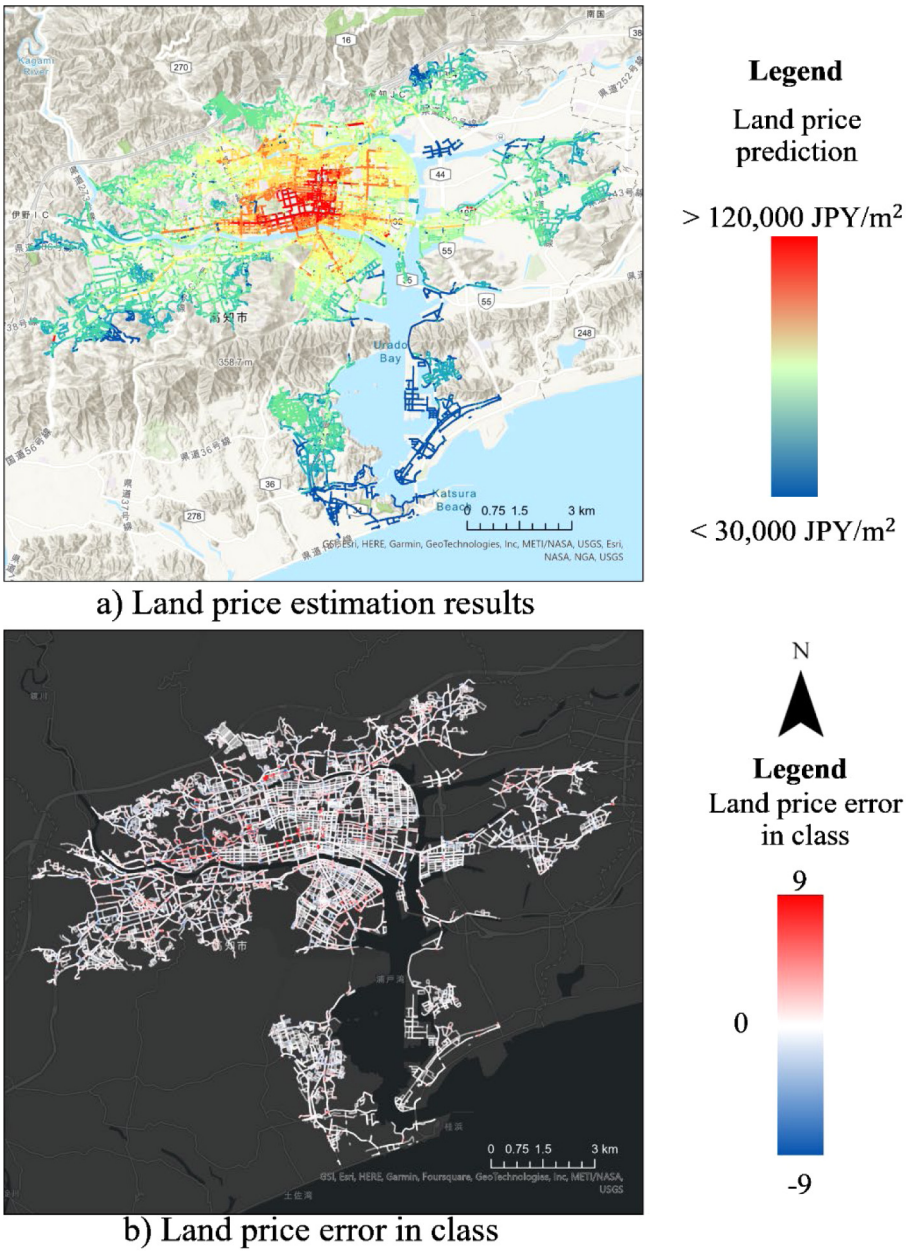


Fig. 10. Visualization of the (a) land price estimation result (b) land price error in land price class.

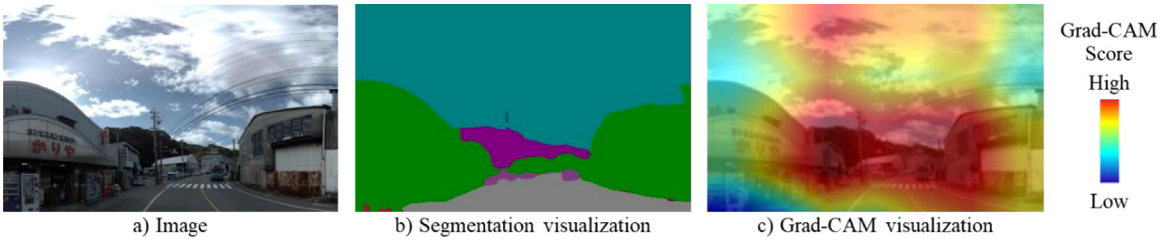


Fig. 11. (a) Street view image; (b) semantic segmentation visualization; (c) Grad-CAM visualization.

Table 4
Correlation coefficients between human subjective perception scores and land prices.

0. Open	1. Friendly	2. Lively	3. Comfortable	4. Greenery	5. Calm	6. Bright	7. Old-fash.	8. Safe	9. Neat	10. Lived-in f.
0.01	0.09	0.36	0.54	−0.25	−0.02	0.12	−0.44	0.12	0.67	−0.02
11. Cosy	12. Clean	13. Beautiful	14. Wealthy	15. Boring	16. Depressing	17. Like	18. Interesting	19. Des. living	20. Des. go thr.	21. Attractive
0.30	0.36	0.09	0.60	−0.47	−0.54	−0.57	0.16	0.37	0.36	0.40

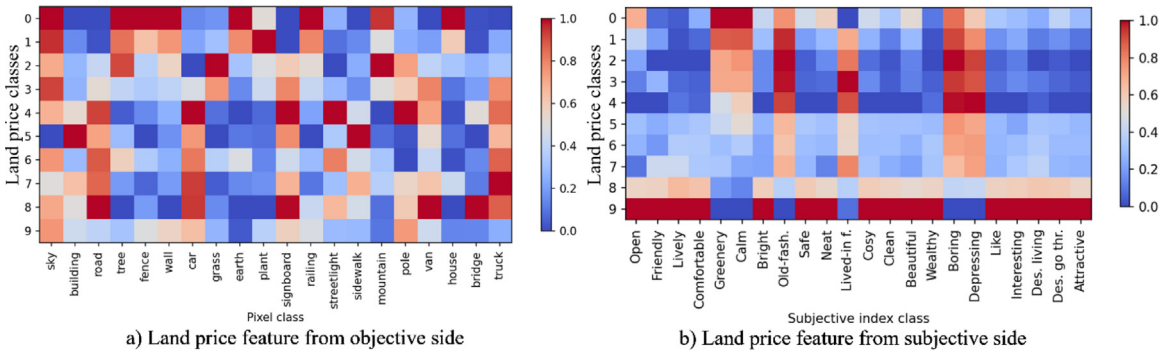


Fig. 12. Visualization of land price feature analysis.

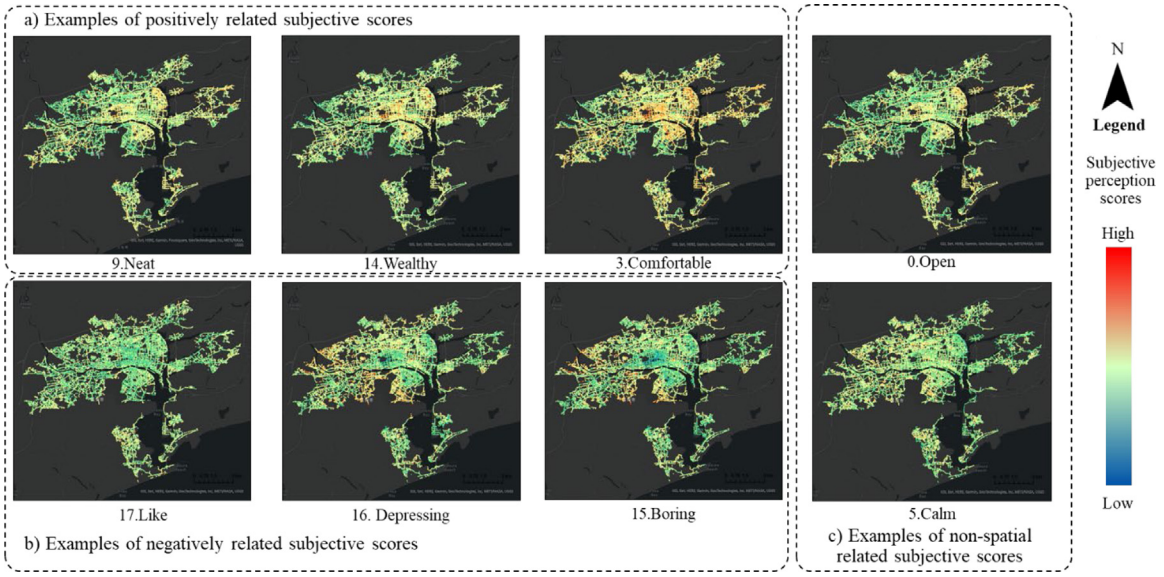


Fig. 13. Visualization of human subjective perception scores for spatial features: (a) examples of positively related subjective scores; (b) examples of negatively related subjective scores; (c) examples of non-spatially related subjective scores.

find out which factors in the street view images have the most influence on the land estimation model. Since we know that the Grad-CAM output scores denote the pixels that have a great influence on the estimation results, we can take advantage of the segmentation result to perform the quantitative analysis.

We calculated the Grad-CAM scores mean value of each segmentation pixel class, the mean values would indicate the importance degree for the land price estimation. The importance ranking was shown in Fig. 14.

5.3.2. Linear regression analysis results

To compare the impact of streetscape and perception, we used the pixel proportion and the quantitative perception scores to conduct a linear regression model. The input of the regression model for each image was normalized 150-pixel proportions and 22 proportion scores. After removing the weight whose P-value was larger than 0.01, the weight visualization is shown in Fig. 15.

Fig. 15 illustrates the impact analysis from Grad-CAM (Fig. 14). Perception spatial analysis has the same trend as Fig. 15. Comfortable and Wealthy perception factors play positive roles while Like and Boring show negative impacts. Then streetscape factors such as road, car, and sky in the 150-pixel classes showed their importance both in Figs. 15 and 16. Observing from the absolute values in Fig. 16, the perception scores have more impact on land price estimation than the streetscape factors, which fit our intuition well. As discussed above,

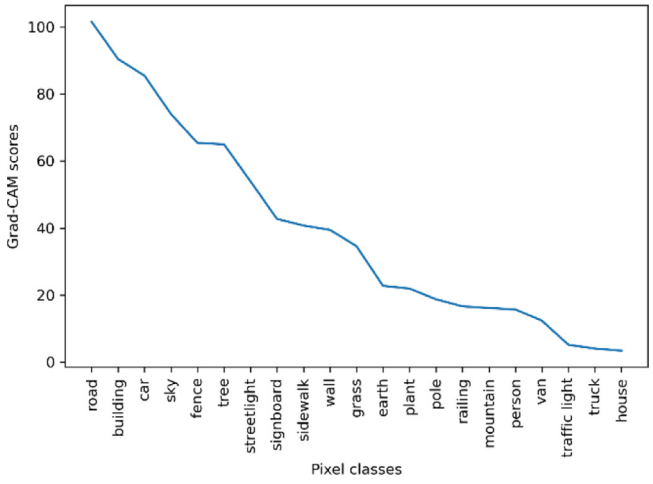


Fig. 14. Ranking of the streetscape factors' importance to the land price estimation model and results.

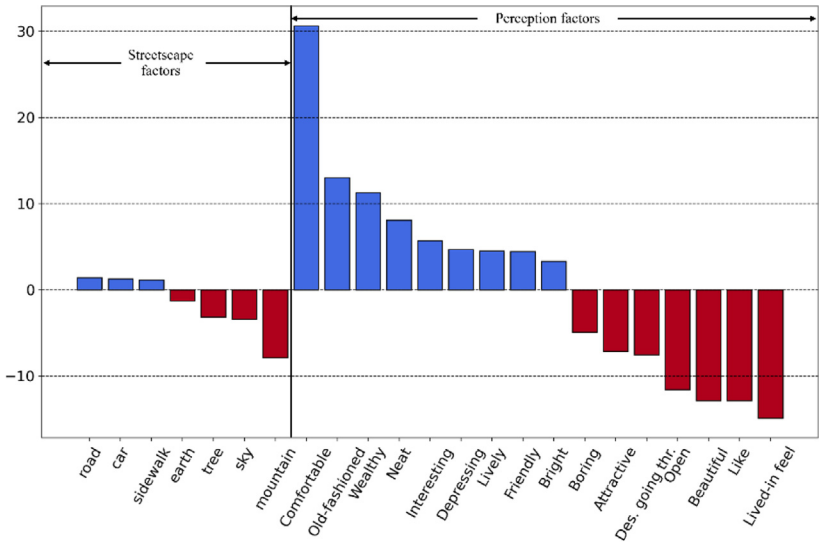


Fig. 15. Regression model weights visualization.

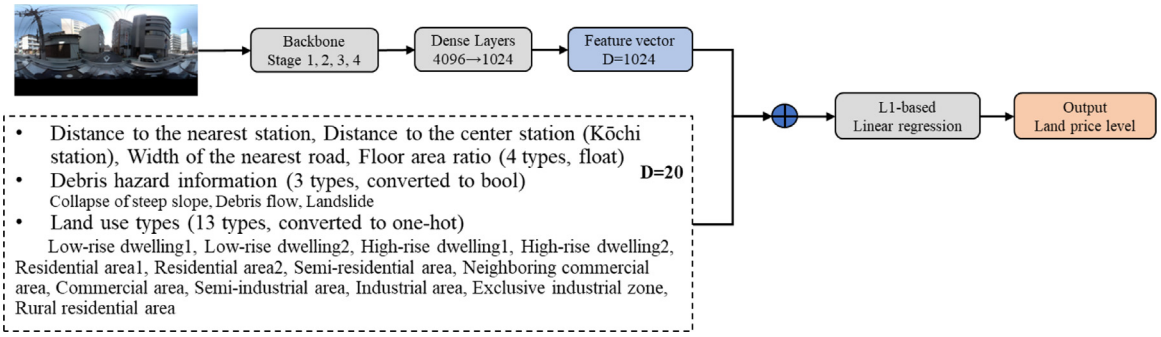


Fig. 16. Schematic of adding non-image factors processing.

we can conclude that streetscape and perception factors influence land prices positively as well as negatively. Further, human perception factors are more important than streetscape factors in land price estimation.

This study makes an important contribution to existing research on land prices by demonstrating the superiority of a newly developed model that accounts for landscape and human subjective perceptions from street view images, although it has some limitations. The land price model was calculated only for one city, as mentioned earlier. Additionally, the results cannot be generalized to other cities. Future studies should verify the results in other Japanese cities and even in other parts of the world.

5.4. Non-image factors consideration

As discussed in Section 2.1, many studies selected some structural factors that appear to have significant impact on the land price or housing price, such as location, public, and recreation facilities, to build the model (Xiao-zhu and Ling-wei, 2013; Kisilevich et al., 2013; Jim and Chen, 2009). We also considered adding some available non-image factors (N.I. factors in Table 5) to our models' pipeline to further compare and discuss how the structural factors affect our model.

Thus, considering the factors that were used in related studies of Japan, such as floor area ratio, width of the nearest road, and land use type, we constructed a ConvNeXt+N.I factors model (Fig. 16). Furthermore, we compared the state-of-the-art linear regression model based on only non-image factors.

The ancillary data, and the processing workflow are shown in Fig. 16. We first extracted the feature from the images, then concatenated these image features with the non-image features, and lastly implemented L1-based linear regression to output the land price level. We also only used the non-image features to do the regression analysis for the comparison. The linear regression processing is shown in Eq. (1). The results comparison is presented in Table 5.

The non-image factors based model's performance is better than that of ResNet-101, with the development of the computer vision algorithms, and ViTs and CNNs in the 2020s outperformed the non-image factors based model. Furthermore, the ancillary non-image factors improved the performance of the land price estimation model, but less than 1%, which is not remarkable. The main reason can be considered as: (i) compared to the 1024-D feature extracted from images, the information of the 20-D non-image feature is not sufficient for the land price estimation, and it is difficult to obtain other type of non-image data because of the data accessibility; (ii) the deep learning algorithm can learn the ancillary data information from the image data, such as road width and land use type, and the floor area ratio and parcel size limitation is strongly related to land use type in Japan, so the duplicated of ancillary data and extracted image feature is one of the reason.

5.5. Human subjective perception score clustering analysis

In the 22 types of human subjective perceptions, some of them are similar intuitively such as 15. Boring and 16. Depressing, 17. Like and 21. Attractive. From this assumption, we can divide the 22

Table 5
Results comparison of only images and both non-image and image factors.

	Setting	R^2	mF1	Accuracy
Test data	N.I. factors	0.7218	0.5622	0.5608
	ConvNeXt-B	0.8645	0.7820	0.7799
	ConvNeXt-B + N.I. factors	**0.8713**	**0.7862**	**0.7843**
Study area	N.I. factors	0.7127	0.5616	0.5565
	ConvNeXt-B	0.8331	0.7820	0.7706
	ConvNeXt-B + N.I. factors	**0.8386**	**0.7843**	**0.7745**

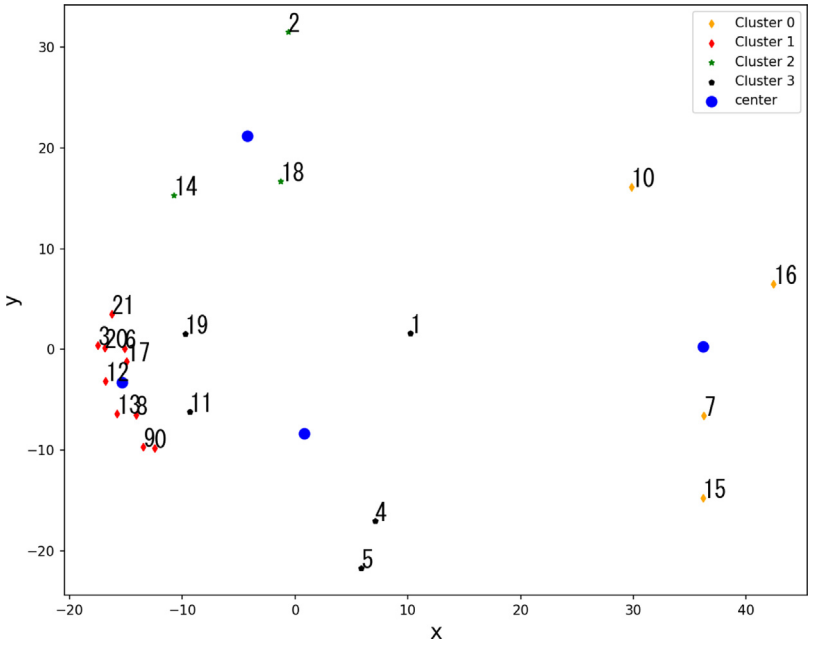


Fig. 17. Human subjective perception score clustering result.

attributes into several categories, and observe the relationship between the clustering results and land price distribution.

We extracted the human subjective perception scores as discussed in Sections 4.3.1 and 5.3.1; the result is an $(n, 22)$ shaped array, where n is the images' number. Then, we implemented k-means clustering on this array, to divide the 22 types of perception into 4 categories; the clustering results is shown in Fig. 17. We implemented principal component analysis (PCA) for visualization.

Then, we compared the clustering result with the correlation coefficients shown in Table 4 and noticed that the cluster 0 (Fig. 17) is strongly related to the land price; additionally, cluster 3 is basically negatively related to the land price, and the correlation coefficient is not remarkable between cluster 2 and land price. However, both strongly related and irrelevant items appeared in cluster 1. This result reminded us that although many perceptions interact with land prices, intuitively similar perceptions are not necessarily statistically similar, and furthermore, statistically similar perceptions do not necessarily have a similar effect on land prices.

6. Conclusions

In this study, we linked street views with land prices and proposed a deep learning approach for estimating land prices. This achieved an accuracy of 77.99% in the test set and 77.06% in the dataset with a total of approximately 0.8 million images. The experiments showed that with the development of the computer vision algorithms, ViTs and CNNs in the 2020s outperformed the non-image factors based model in many related studies. A deep learning algorithm can extract some important information such as road width, parcel size, floor area ratio, and land use type which are strongly related to the land price. Additionally, the street view images are easy to access in all countries, and our model is easier to be generalized to other countries. Thus, a vision-based, high accuracy dense land price model can be implemented in the study area, and we will generalize our workflow to other counties.

Furthermore, because the related land price estimation and causal analysis primarily focused on non-image factors, the vision-based study was still in absence. We proposed the human subjective perception score extraction approach, and designed a method to evaluate and quantitatively compare the impact of streetscape and perception factors on the land price. The experiment results showed positively related human subjective perceptions such as 9. Neat, 14. Wealthy, and 3. Comfortable, and negatively related perceptions such as 17. Like, 15. Boring, and 16. Depressing. In addition, the human perception contributes more to the causality of land price than objective streetscape factors. The conclusion in this study will provide a reference for the future studies of urban planning.

Regarding future work, we will have a more detailed analysis between the correlation of human subjective perception and land price. We will analyze the changes of the land price and the objective streetscape, as well as subjective perception, in a continuous time sequence and determine the causality of those changes. Furthermore, as our study is densely estimation and is based on scene level street view overall image information, non-building areas can also be considered for predicting the future prices in undeveloped areas. Combined with the human subjective perception, further causal analysis and validation can be deployed on those undeveloped areas. With the time sequence data, not only land price estimation but also prediction can be implemented in the future work, especially the area that can be well developed but has not been developed yet, such as new towns or urban peripheries. Analysis in those areas could focus on which factors may

cause significant changes in land prices, as well as how objective and subjective factors may change in those areas.

CRediT authorship contribution statement

Chenbo Zhao: Conceptualization, Methodology, Software, Writing – original draft. **Yoshiki Ogawa:** Supervision, Methodology, Writing – review & editing. **Shenglong Chen:** Methodology, Writing – review & editing. **Takuya Oki:** Supervision, Methodology, Writing – review & editing. **Yoshihide Sekimoto:** Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was conducted in collaboration with CSIS, and the University of Tokyo (No. 1013) and used road network data provided by Sumitomo Denko Co., Ltd.

References

Afonso, B., Melo, L., Oliveira, W., Sousa, S., Berton, L., 2019. Housing prices prediction with a deep learning and random forest ensemble. In: Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional. SBC, pp. 389–400.

Al-Garadi, M.A., Mohamed, A., Al-Ali, A.K., Du, X., Ali, I., Guizani, M., 2020. A survey of machine and deep learning methods for internet of things (IoT) security. IEEE Commun. Surv. Tutor. 22 (3), 1646–1685.

Alfiyatin, A.N., Febrita, R.E., Taufiq, H., Mahmudy, W.F., 2017. Modeling house price prediction using regression analysis and particle swarm optimization. Int. J. Adv. Comput. Sci. Appl. 8 (10), 323–326.

Alvarez, F., Roman-Rangel, E., Montiel, L.V., 2022. Incremental learning for property price estimation using location-based services and open data. Eng. Appl. Artif. Intell. 107, 104513.

Anand, S., Yadav, P., Gaur, A., Kashyap, I., 2021. Real estate price prediction model. In: 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). IEEE, pp. 541–543.

Botalb, A., Moinuddin, M., Al-Saggaf, U.M., Ali, S.S., 2018. Contrasting convolutional neural network (CNN) with multi-layer perceptron (MLP) for big data analysis. In: 2018 International Conference on Intelligent and Advanced System. ICIAS, IEEE, pp. 1–5.

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese …, F., Riedmiller, M., 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. Nature 602 (7897), 414–419.

Derdouri, A., Murayama, Y., 2020. A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across Fukushima prefecture, Japan. J. Geogr. Sci. 30 (5), 794–822.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner ., T., Houlsby, N., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Driss, S.B., Soua, M., Kachouri, R., Akil, M., 2017. A comparison study between MLP and convolutional neural network models for character recognition. In: Real-Time Image and Video Processing 2017, Vol. 10223. SPIE, pp. 32–42.

Du, H., Ma, Y., An, Y., 2011. The impact of land policy on the relation between housing and land prices: Evidence from China. Q. Rev. Econ. Finance 51 (1), 19–27.

Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A., 2016. Deep learning the city: Quantifying urban perception at a global scale. In: European Conference on Computer Vision. Springer, Cham, pp. 196–212.

Gao, X., Asami, Y., 2007. Effect of urban landscapes on land prices in two Japanese cities. Landsc. Urban Plan. 81 (1–2), 155–166.

Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430.

Ghodsi, R., Boostani, A., Faghihi, F., 2010. Estimation of housing prices by fuzzy regression and artificial neural network. In: 2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation. IEEE, pp. 81–86.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press;
Hayashi, Y., Suzuki, Y., Sato, S., Tsukahara, K., 2016. Disaster Resilient Cities: Concepts and Practical Examples. Butterworth-Heinemann.

Hau, A., Zhang, P., Zheng, Z., Zhu, M., He, Y., Li ., Q., Li, J., 2018. Land price prediction based on random forest. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 2948–2951.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hu, S., Yang, S., Li, W., Zhang, C., Xu, F., 2016. Spatially non-stationary relationships between urban residential land price and impact factors in Wuhan city, China. Appl. Geogr. 68, 48–56.

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W., 2019. Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 603–612.

Jiang, C., Li, J., Wang, W., Ku, W.S., 2021. Modeling real estate dynamics using temporal encoding. In: Proceedings of the 29th International Conference on Advances in Geographic Information Systems. pp. 516–525.

Jim, C.Y., Chen, W.Y., 2009. Value of scenic views: Hedonic assessment of private housing in Hong Kong. Landsc. Urban Plan. 91 (4), 226–234.

Kang, Y., Zhang, F., Gao, S., Peng, W., Ratti, C., 2021. Human settlement value assessment from a place perspective: Considering human dynamics and perceptions in house price modeling. Cities 118, 103333.

Kisilevich, S., Keim, D., Rokach, L., 2013. A GIS-based decision support system for hotel room rate estimation and temporal price prediction: The hotel brokers' context. Decis. Support Syst. 54 (2), 1119–1133.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Levering, A., Marcos, D., Tuia, D., 2021. On the relation between landscape beauty and land cover: A case study in the UK at sentinel-2 resolution with interpretable AI. ISPRS J. Photogramm. Remote Sens. 177, 194–203.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang ., Z., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. arXiv preprint arXiv:2201.03545.

Liu, Z., Wang, P., Zha, T., 2013. Land-price dynamics and macroeconomic fluctuations. Econometrica 81 (3), 1147–1184.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Nakamura, H., 2019. Relationship among land price, entrepreneurship, the environment, economics, and social factors in the value assessment of Japanese cities. J. Clean. Prod. 217, 144–152.

Naser, N., Serte, S., Al-Turjman, F., 2020. From traditional house price appraisal to computer vision-based: A survey. In: International Conference on Forthcoming Networks and Sustainability in the IoT Era. Springer, Cham, pp. 1–10.

Nouriani, A., Lemke, L., 2022. Vision-based housing price estimation using interior, exterior & satellite images. Intell. Syst. Appl. 200081.

Oki, T., Kizawa, S., 2021. Evaluating visual impressions based on gaze analysis and deep learning: a case study of attractiveness evaluation of streets in densely built-up Wooden Residential Area. Int. Arch. Photogram. Rem. Sens. Spat. Inf. Sci. 43, 887–894.

Qiu, W., Li, W., Liu, X., Zhang, Z., Li, X., Huang, X., 2023. Subjective and objective measures of streetscape perceptions: Relationships with property value in shanghai. Cities 132, 104037.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Smith, L.N., Topin, N., 2019. Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications (Vol. 11006, p. 1100612). International Society for Optics and Photonics.

Suzumura, T., Sugiki, A., Takizawa, H., Imakura, A., Nakamura, H., Taura ., K., Uchibayashi, T., Research Collaborations, 2022. mdx: A cloud platform for supporting data science and cross-disciplinary. arXiv preprint arXiv:2203.14188.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov ., D., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Viana, D., Barbosa, L., 2021. Attention-based spatial interpolation for house price prediction. In: Proceedings of the 29th International Conference on Advances in Geographic Information Systems. pp. 540–549.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 418–434.

Xiao-zhu, D., Ling-wei, K., 2013. The land prices and housing prices—Empirical research based on panel data of 11 provinces and municipalities in eastern China. In: 2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings. IEEE, pp. 2118–2123.

Xu, X., Qiu, W., Li, W., Liu, X., Zhang, Z., Li, X., Luo, D., 2022. Associations between street-view perceptions and housing prices: Subjective vs. Objective measures using computer vision and machine learning techniques. Remote Sens. 14 (4), 891.

Yamada, S., Yamasaki, S., Okuno, T., Harada, K., Sasaki, Y., Onizuka, M., 2020. Are satellite images effective for estimating land prices on deep neural network models? In: 2020 21st IEEE International Conference on Mobile Data Management (MDM). IEEE, pp. 304–309.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell ., D., Burke, M., 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. Nature Commun. 11 (1), 1–11.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y., 2022. CoCa: Contrastive captioners are image-text foundation models. arXiv preprint arXiv: 2205.01917.

Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032.

Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2020. Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. pp. 13001–13008, 07.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2017. Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 633–641.