THE EFFECT OF CULTURE-SPECIFIC DIFFERENCES IN URBAN STREETSCAPES ON THE INFERENCE ACCURACY OF DEEP LEARNING MODELS

T. Inoue[1,*], R. Manabe[1], A. Murayama[1], H. Koizumi[1]

[1] Department of Urban Engineering, School of Engineering, The University of Tokyo, Japan
(tak, rik, murayama)@up.t.u-tokyo.ac.jp, hide@cd.t.u-tokyo.ac.jp

Commision IV, WG IV/9

KEY WORDS: Streetscape, Deep Learning, Semantic Segmentation, Intersection over Union, Inference Accuracy.

ABSTRACT:

Owing to the increasing focus on places in urban planning and design, methods to evaluate the quality and value of urban places is crucially needed. Many studies use deep learning models to identify the proportion and composition of landscape elements in images for evaluation. The accuracy of semantic segmentation achieved with such models is often validated using Cityscapes, a street-level image dataset taken from German cities. However, few studies have quantitatively revealed the inference accuracy decrease caused by culture-specific characteristics of streetscapes.

In this study, we calculated by-class intersection over union (IoU) and newly-defined indices of false inferences to demonstrate how and to what extent deep learning models can infer each landscape element falsely when applied to Japanese street-level images. Our analysis revealed that certain landscape elements are more difficult to infer correctly based on specific causes, such as their appearances in images and unique characteristics of the fixed physical configuration of Japanese streets. By applying the false inference categorization framework presented in this study, researchers can adjust their approaches considering two aspects: a decrease in inference accuracies of deep learning models and the impact of culture-specific characteristics of streetscapes on people's perception and valuation of urban places. Based on the results and analyses, a future research direction is to develop and implement more accurate image recognition models considering culture-specific characteristics to understand people's perceptions of urban spaces and assess the value of urban places by using the big data including street-level images.

1. INTRODUCTION

1.1 Growing Attention to Places

During the 20th and 21st centuries, the world has been experiencing an era of urbanization (Friedmann, 2010). According to the United Nations, the urban population, which was only 270 million in 1900, surpassed 750 million in 1950 and exceeded 50% of the entire world population in 2007. Today, the number of urban dwellers is estimated to be 4.2 billion (Ritchie & Roser, 2018; United Nations, 2018). Looking at the two centuries of urbanization macroscopically, it is noticeable that we have already passed halfway through the significant shift.

Supposing that urban planning was born to address the problems arising from urbanization, its aim and scope would differ between the first and second halves of the urbanization swell. During the first half, the main focuses were improving public health, enhancing urban aesthetics, and properly arranging urban functions (Corburn, 2009; Ward, 2004). After Patrick Geddes advocated civics and Ebenezer Howard the garden city, in Europe, North America, and later in Asia, cities experienced the advent of the modern urban planning by the 1910s (Hall, 2013; Home, 1990; Peterson, 2009; Sorensen, 2005). From then on, rational, valid, objective, scientific, and comprehensive planning based primarily on functionality defined the urban physical context of the 20th century (Faludi, 2013; Hudson et al., 1979).

In contrast, the focus of urban planning in the latter half of urbanization shifted onto the "place," which acquired an academic definition in the 1970s by humanitarian geographers such as Yi-fu Tuan and Edward Relph as spatial locations that have been become meaningful through human experiences (Manzo & Perkins, 2006; cf. Relph, 1976; Tuan, 1979). Places became an academic focal point for researchers to understand the cities through the lens of their psychological relationships with the people. Urban planning and design have gradually shifted their focus since the 1960s, when activists and philosophers became increasingly critical of rational comprehensive urban planning for prioritizing efficiency and simultaneously marginalizing people (Irving, 1993).

Meanwhile, urban design practitioners and policymakers have maintained a theoretically strained relationship with the theories of place that were discussed and developed in human geography or environmental psychology. Subsequently, they narrowed their argument on places down to the urban regeneration of creating lively public spaces (Aravot, 2002). In 1980s, Bryant Park in New York City was transformed into a vibrant and inviting plaza, symbolizing the revival of a city in disrepair (Francis, 1989; Madden, 2010). The New Urbanism movement flourished inspired by those practices, resulting in its principles articulating in the form of the Ahwahnee Principles and the New Urbanism Charter (Grant, 2005; Katz, 1994). Later, Danish architect Jan Gehl insisted that planners prioritize people's lives before space and architecture. He also argued that enhancing public life in public spaces was essential for achieving democratic and fulfilling lives (Gehl, 2013).

* Corresponding author

Moreover, British urban planner Patsy Healey specified that urban planning projects should aim to enhance the quality of places (Healey, 2010). Partly as an antagonism to dominant objectivity and rationality adopted in the past, urban planning now means not just the planning of spaces but the making of places as well.

1.2 Evaluating Urban Places Based on Streetscapes

Owing to the increasing academic attention paid to urban places, there has been prominent needs to assess their quality and value. Initially, information about places was gathered through workshops and interviews. However, these methods lack comparability and scalability (Zhang et al., 2018). Numerous research has attempted to solve these problems using information technologies and big data. As Tim Cresswell pointed out, the concept of place is both epistemological and ontological (Cresswell, 2014). Conceptual frameworks, such as place identity and landscape value, have been established to evaluate the ontological aspects of urban places ("what kind of values they have"; Proshansky et al., 1983; Brown, 2004). The geographic distributions of place values have been visualized using participatory methods of public participation geographic information systems (PPGIS; Brown & Weber, 2012).

Meanwhile, streetscape has often been used to assess epistemological aspects of places. Evaluating urban spaces using visual information, which accounts for most of our sensory stimuli, is a longstanding practice, including the notable studies by Kevin Lynch (Lynch, 1964, 1984). Then, a new trend has been emerging recently that applies computer vision technologies to recognize what is in images and videos. Ordonez & Berg (2014) conducted a survey in which respondents were presented with street images and asked to rate the "safety," "uniqueness," and "wealth" they felt about each image. Then, they proposed a regression model that estimates the level of those attributes using the image features as parameters. Zhang et al. (2018) conducted similar regression analyses for six indicators: safety, liveliness, beauty, wealth, depression, and boredom. These studies provide notable examples of how to elucidate people's perceptions using big data of streetscape images.

1.3 The Use of Deep Learning Models and Their Limitation

These studies use deep learning models to identify the proportion and location of landscape elements in an image. For example, the model used by Zhang et al. (2018) is the pyramid scene parsing network (PSPNet; Zhao et al., 2017). Such deep learning models use Cityscapes, a dataset containing streetscape images taken from 50 German cities (Cordts et al., 2016), to test the accuracy of pixel-level semantic segmentation; that is, the task of inferring which landscape elements each pixel of the image expresses. The benchmark indices shown on the Cityscapes dataset website indicate how accurately each model was able to conduct the labelling task regarding each landscape element (Cityscapes Dataset, 2022).

However, because urban streetscape datasets, such as Cityscapes, were created to help develop autonomous vehicles, most landscape elements annotated are related to transportation and mobility. In contrast, the use and form of buildings or street furniture, which may seriously affect people's evaluation of a city, are not classified on the dataset. In addition, the sidewalks along German roads would be different in shape from those on Asian roads, and German buildings would be different in material and form from those in Africa. Likewise, the colourful and vivid billboards that can be seen in large cities in Asian countries would not be present in Germany. While it may be essential to understand the impact of these cultural differences in urban streetscapes on deep learning models, only few research have been conducted considering this aspect.

Different cities and countries have different cultures. In turn, cultural differences have considerable influences on streetscapes through laws and regulations that produce specific types of urban spaces, various measures and techniques that affect the physical configuration of spaces, and differences in the shape and size of products used. Consequently, the difference in streetscapes among cultures may affect the research on the perception of urban places based on image recognition. Although Ordonez & Berg (2014) studied four U.S. cities (New York, Boston, Chicago, and Baltimore) and Zhang et al. (2018) studied two Chinese cities (Beijing and Shanghai), it has not been confirmed whether the deep learning models are as accurate as expected when applied to cities in these countries. It is necessary to specify the impact of differences in urban streetscapes on deep learning models.

1.4 Research Objective

The purpose of this study is to understand the culture-specific effects on the accuracy of deep learning models. This was primarily done by comparing the segmentation accuracy of each landscape element validated on Cityscapes with those when using Japanese streetscape images. Deep learning models used were DeepLab v3 plus (Chen et al., 2017) and PSPNet (Zhao et al., 2017). Specifically, we compared the published benchmark values for each landscape element with those obtained using Japanese streetscape images and counted the number of incorrectly inferred pixels between different landscape element classes. Moreover, we examined the factors that cause such differences in accuracy (culture-specific landscape elements and differences in design and materials in the same element). In doing so, since not all differences in inference accuracy can be attributed to culture-specific characteristics, of course, we set up a framework for the types of false inferences and attempted to distinguish culture-specific effects from others. This trial provides insights into the use of big data including urban streetscape images for evaluating places in cities worldwide.

2. INDICES AND METHODS

2.1 Definitions of Indices

An index to evaluate the accuracy of semantic segmentation achieved by deep learning models is intersection over union (IoU). This index is defined as the number of "true positive" pixels divided by the total number of "true positive," "false positive," and "false negative" pixels; the maximum value is 1 when the model makes completely accurate inferences, while the minimum is 0 for the utterly inaccurate inferences. The IoU for each class of landscape elements is denoted as the by-class IoU, and the mean value for all classes is denoted as the mean IoU (mIoU).

IoU is an evaluation index based on inference accuracy on the same class. However, the purpose of this study is to clarify how the different streetscape characteristics due to the cultural background affect the inference accuracy achieved by deep learning models. Therefore, in addition to this index, two novel indices, namely, the "Number of Falsely Inferred Pixels" (NFIP) and "Rate of Falsely Inferred Pixels" (RFIP), are introduced.

They are indicators of how each landscape element was inferred as other landscape elements. We defined $NFIP_{tf}$ as the number of landscape element T pixels per image that have been incorrectly inferred as landscape element F. We also defined $RFIP_{tf}$ as the percentage of landscape element T pixels that were incorrectly inferred as landscape element F.

2.2 Deep Learning Models

Deep learning models tested in this study are DeepLab v3 plus and PSPNet. DeepLab v3 plus is a model that combines a spatial pyramid pooling module with an encoder-decoder structure to refine the segmentation results, especially at object boundaries, by leveraging both (Chen et al., 2017). It has achieved an 82.1% mIoU by classes for the pixel-level semantic labeling task as a benchmark on the Cityscapes dataset (Cityscapes Dataset, 2022). PSPNet is a model that uses a convolutional neural network (CNN) to create feature maps, followed by a pyramid module to obtain both local and global information (Zhao et al., 2017). The mIoU by classes for the pixel-level semantic labeling task in the Cityscapes dataset was 81.2% (Cityscapes Dataset, 2022). In this study, we used the models that were pre-trained with the Cityscapes dataset.

2.3 Landscape Elements

In the Cityscapes dataset, 30 classes of landscape elements are defined as shown in Table 1. IoU values are calculated for 19 of these landscape elements for each deep learning model. The elements included are road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, sky, person, rider, car, truck, bus, train, motorcycle, and bicycle (Chen et al., 2017; Zhao et al., 2017). In this study, we attempted to verify the difference in accuracy for these 19 classes.

Category	Classes	
	IoU Evaluation Targets	Non-IoU Evaluation Targets
Flat	Road, Sidewalk	Parking, Rail track
Human	Person, Rider	-
Vehicle	Car, Truck, Bus, On rails, Motorcycle, Bicycle	Caravan, Trailer
Construction	Building, Wall, Fence	Guard rail, Bridge, Tunnel
Object	Pole, Traffic sign, Traffic light	Pole group
Nature	Vegetation, Terrain	-
Sky	Sky	-
Void	-	Ground, Dynamic, Static

Table 1. Landscape elements defined in Cityscapes dataset.

2.4 Images of Japanese Streetscape

In this study, the target of comparison with the inference accuracy achieved on German urban streetscape imagery was that achieved on the Japanese counterparts. We created a dataset comprising 1,990 images. Of those images, 1,790 were acquired from Google Street View. Image acquisition was carried out by specifying latitudes and longitudes approximately every 100 m in the north-south direction and approximately every 200 m in the east-west direction in an area that roughly covers the southern half of the special wards of Tokyo, Japan (35.68°N - 35.72°N, 136.6°E - 136.8°E). The imaging direction was set as the following: due north at the initial point, due east at the next point, then due south, due west, and coming back to due north. After collecting the images, we manually excluded unsuitable ones for streetscape image recognition, such as heavily distorted or indoor images.

Moreover, one hundred images of Fujimidai neighborhood, Kunitachi City, Tokyo, where the authors' research group is conducting their joint research. Yet another one hundred Google Street View images obtained by the authors from different parts of the Tokyo metropolitan area were added. Note that, although using images acquired from Google Street View for analysis was not prohibited as of 2019 when we conducted this study, it is not permitted now according to their updated terms and conditions. In response to the change in their service, we opted to use images obtained from the open-source imagery platform Mapillary instead of Google Street View. The images used in this study are available on the Zenodo repository (doi: 10.5281/zenodo.6546479).

2.5 Calculations of By-class IoUs and Falsely Inferred Pixel Indices

Using the images and deep learning models described above, we calculated the by-class IoUs and falsely inferred pixel indices using the following procedure. First, semantic segmentation was performed on the prepared 1,990-image dataset using the two deep learning models to paint images into 20 area types, namely selected 19 landscape elements in the Cityscapes dataset and other areas. Second, manual annotation was applied to each image to define the ground truth for the regions of the landscape elements. Third, we used an online platform for computer vision named Supervisely (https://supervise.ly/) to recognize the ground truth and inferred landscape element class on a pixel-by-pixel basis between the ground truth and inference results. Then, by-class IoUs were calculated based on the number of pixels for "true positive," "false positive," "true negative," and "false negative." We compared these values to those published on the Cityscapes dataset website. In addition, we calculated how landscape elements were falsely inferred. Finally, for all the pairs of different classes, NFIP and RFIP were calculated.

2.6 Analysis of Factors That Cause Differences in the Inference Accuracy

Next, we examined the forms and causes of false inference that decrease the accuracy of semantic segmentation achieved by deep learning models. First, based on the IoU and falsely inferred pixel indices calculated by the method described in the previous section, we detected pairs with a large percentage of false inferences (combinations of landscape elements in ground truths and inference results). In addition, because the "void" area of the ground truth is thought to contain landscape elements that are not covered by the model but are likely to influence people's evaluation of places, we specifically extracted typical elements included in the "void" area. Then, we examined how they were inferred by DeepLab v3 plus.

Based on these results, we developed a framework for assessing the forms and causes of false inference. Next, taking Japanese streetscapes as an example, we discussed the impact of culture-specific street characteristics on the inference accuracy of the deep learning models. This discussion ranges across the landscape element classes set in the Cityscapes dataset and other landscape elements included in the void area.

3. RESULTS

3.1 Calculated By-class IoUs and Decrease in the Inference Accuracy

Figure 1 presents the calculated by-class IoUs for the two deep learning models. Note that the "train" class was excluded because not enough images taken in Tokyo contained trains. Therefore, the by-class IoUs include values for 18 types of landscape elements in the figure.

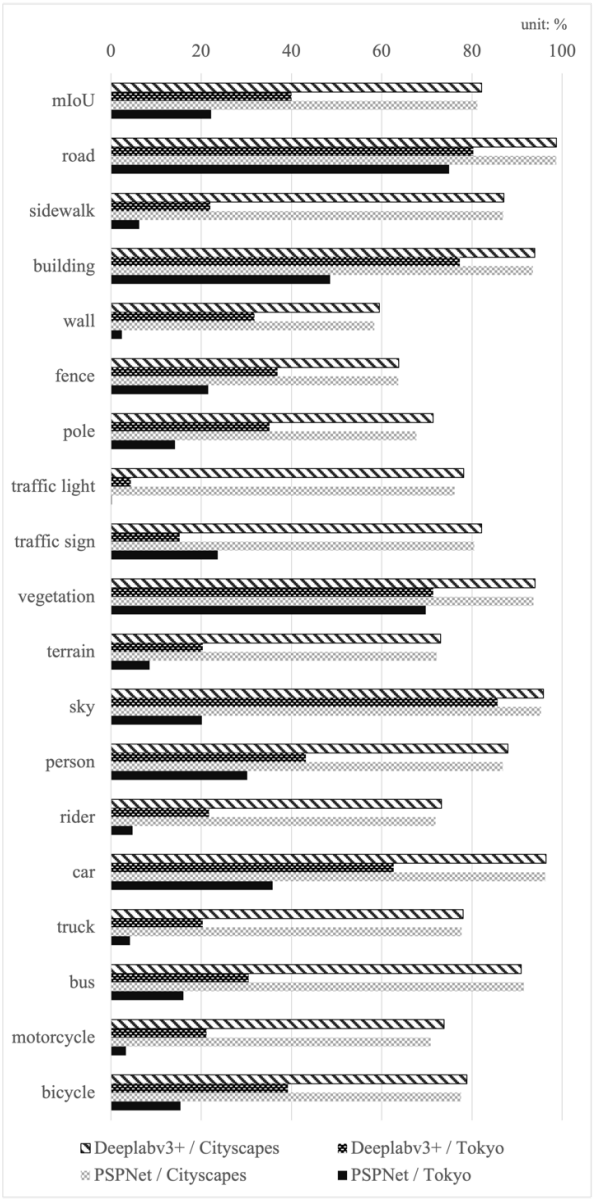


Figure 1. The mIoU and by-class IoUs for Germany and Japan.

The mean IoU was 82.1% using DeepLab v3 plus and 81.2% using PSPNet on the Cityscapes dataset (Cityscapes Dataset, 2022). However, the accuracy dropped to 40.0% and 22.2%, respectively, for the Tokyo dataset. The range of accuracy decreases were relatively small for the "road" (18.4 points for DeepLab v3 plus and 23.8 points for PNPNet) and "vegetation" classes (22.6 points and 23.9 points, respectively). On the other hand, it was the most significant for the "sidewalk" class, at 65.0 and 80.6 points, respectively. This was followed by the "traffic light" class results, at 73.7 and 76.9 points, respectively. There were also differences in how the accuracy dropped among the deep learning models. The "sky" class had an accuracy decrease of only 10.2 points for DeepLab v3 plus, while PSPNet experienced a 75.1-point drop.

3.2 Calculated Falsely Inferred Pixel Indices and Trends in False Inference

Falsely inferred pixel indices (NFIP and RFIP) defined in the methods section were calculated for DeepLab v3 plus. Table 2 reports the NFIPs for the selected landscape element pairs based on the ground truth and inference results. The resolution of Google Street View images used in this study is 640x640; in other words, each image consists of 409,600 pixels. The full table is available on the same Zenodo repository mentioned above. The rows of the table indicate the ground truth classes, which indicate landscape elements defined in the Cityscapes dataset, plus the "google" class, which is independently set up for the Google logo included in the Google Street View images. The number of rows is 29 because the three classes of "void" have been integrated. The columns show inference results that include 18 landscape element types, excluding the "train" class, out of the 19 landscape elements that deep learning models target for inference. The number in the cell, where the ground-truth "car" and inference result "car_dl" intersect, indicates that the area of where the actual car was correctly inferred to be a car was 3,521.04 pixels per image. Similarly, the value in the cell, where "sidewalk" and "road_dl" intersect, indicates that the area inferred by the model to be a road despite being a sidewalk was 15,345.46 pixels per image.

Ground Truth	Inference Result (unit: pixels per image)			
	road_dl	car_dl	building_dl	sky_dl
Road	108763.77	146.84	166.44	1.11
Sidewalk	15345.46	125.05	1415.95	6.85
Car	102.15	3521.04	150.03	0.02
Building	304.77	167.99	86227.49	502.00
Wall	1187.84	57.41	3278.10	6.82
Fence	230.80	52.36	2023.15	79.03
Vegetation	404.13	49.23	1011.47	57.78
Terrain	953.63	9.33	61.66	0.14
Sky	1.29	0.30	4017.82	67083.93

Table 2. NFIP for the selected pairs of landscape elements.

Table 3 reports the RFIPs that indicate what percentages of the ground truth for certain selected landscape elements were inferred as specific landscape elements. The full table is available on the Zenodo repository. For example, the value in the cell where "road" and "road_dl" intersect indicates that the model correctly inferred 98.09% of the roads as roads. On the other hand, the cell value where "traffic light" and "traffic_light_dl" intersect is 5.82%, indicating that traffic lights were rarely correctly identified.

There were 17 pairs for which more than 10% of the ground truth was incorrectly inferred as another specific landscape element (Table 4). The highest percentage of false inferences was found when traffic lights were inferred as buildings (60.72%), followed by sidewalks inferred as streets (60.33%). Generally, the pairs with the lowest accuracy included "sidewalk," "traffic light," or "traffic sign" classes. For the "sidewalk" class, more than 60% of the cases were classified as the "road" class. For the "traffic light" class, approximately 60% of the area were inferred as "building," and approximately 12% as "vegetation." For the "traffic sign" class, 53% of the areas were correctly identified as traffic signs; however,

approximately 23% were inferred as the area of the "building" class. Although the IoU for the "traffic sign" class was low, namely, 0.15, the fact that 53% of the ground truth areas were correctly recognized suggests that in many cases, areas that were not traffic signs were inferred falsely as traffic signs.

Ground Truth	Inference Result (unit: %)			
	road_dl	car_dl	traffic_light_dl	sky_dl
Road	98.09	0.13	0.00	0.00
Sidewalk	60.33	0.49	0.00	0.03
Car	2.52	86.98	0.00	0.00
Building	0.33	0.18	0.00	0.54
Traffic sign	0.32	6.25	0.24	2.71
Traffic light	0.12	0.00	5.82	7.68
Vegetation	0.82	0.10	0.00	0.12
Terrain	25.89	0.25	0.00	0.00
Sky	0.00	0.00	0.00	86.87

Table 3. RFIP for the selected pairs of landscape elements.

Ground Truth	Inference Result	RFIP (%)
Sidewalk	Road	60.33
Rider	Person	33.33
Truck	Car	23.46
Bus	Car	38.64
Motorcycle	Car	12.86
Motorcycle	Bicycle	11.00
Motorcycle	Building	13.79
Wall	Building	25.30
Fence	Building	21.23
Fence	Vegetation	11.51
Pole	Building	26.85
Pole	Vegetation	18.67
Traffic sign	Building	23.41
Traffic light	Building	60.72
Traffic light	Vegetation	11.96
Terrain	Road	25.89
Terrain	Vegetation	25.37

Table 4. List of the pairs with RFIP values higher than 10%.

3.3 Inference Trends for Landscape Elements Not Defined as Classes

We have mentioned the analysis of landscape elements defined as classes in the Cityscapes dataset. However, other landscape elements are also essential to people's evaluation of urban places. Therefore, we decided to count landscape elements that were not defined in the Cityscapes dataset but were annotated as being in the "void" region. Then, we identified to which classes each element was inferred to belong to by DeepLab v3 plus. Six hundred images were randomly selected out of the 1,990-image dataset prepared, and the presence or absence of each element, not the number of pixels, was assessed for each image.

Table 5 lists the landscape elements appearing in at least 30 of the 600 images. Streetlights were the most common, appearing in 138 images, followed by lightning rods/transformers (108 images), bollards (86 images), triangular cones (78 images), plantings (70 images), curbs (67 images), and flowerpots (63 images). As for signboards and billboards, because people's impressions might differ depending on their form, they were divided into eight types: back of traffic signs, rooftop billboards, commercial/non-commercial side signboards, free-standing commercial/non-commercial billboards, and standing commercial/non-commercial signboards.

Regarding these elements, we counted the classes each landscape element was often inferred as. Figure 2 demonstrates the case of streetlights. They were inferred as "building" in 73 images, "vegetation" in 42 images, and "sky" in 34 images. Similarly, lightning rods and transformers were most frequently inferred as "vegetation" in 54 images; bollards were often inferred as "pole," "road," or "sidewalk; triangular cones were frequently inferred as "road," "building," or "traffic sign."

Element	N	Element	N
Streetlights	138	Litter bins	44
Lightning rods/ transformers	108	Commercial side signboards	43
Bollards	86	Vending machines	43
Triangular cones	78	Antennas	37
Plantings	70	Terminal/ junction boxes	36
Curbs	67	Electric wire protection covers	34
Flowerpots	63	Standing commercial signboards	31
Free-standing non-commercial billboards	51	Switchboards	30
Free-standing commercial billboards	49		

Table 5. List of frequently appearing elements in “void” regions.

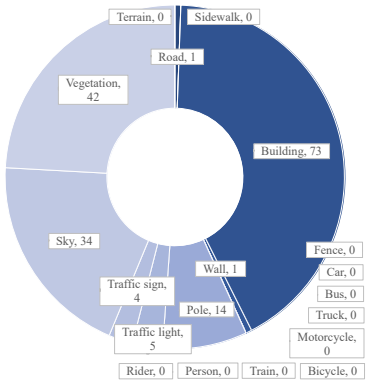


Figure 2. Classes the streetlights were inferred as.

4. DISCUSSION

4.1 Typology of False Inference

This study reveals deep learning models' tendency to misinterpret landscape elements in different countries by calculating IoUs and falsely inferred pixels indices. As represented in the results section, false inferences were more likely to occur among certain combinations of landscape elements. Reviewing the images in which false inferences occur suggests six types of false inferences, where each pair of landscape elements falls into one of them.

These six types are first classified along two axes: "similarity of appearance/proximity of location" (the forms of false inference) and "based on appearance in the image/based on culture-specific characteristics" (the causes of false inference). The "based on culture-specific characteristics" category is further divided based on the characteristics of the fixed physical configuration of the street space and characteristics of the mobile elements that appear temporarily in that space (Table 6).

In the sixfold typology, "similarity of appearance based on appearance in the image" (Type I) and "proximity of location based on the appearance in the image" (Type II) are related to issues of representation. They are not related to differences in the streetscape based on cultural differences. In addition, in the "based on culture-specific characteristics" category, those concerning the characteristics of mobile products (Type III and IV) are mainly related to the shape and color of mobility products. For example, we can observe examples where buses traveling on Japanese streets are smaller than their German counterparts, resulting in Japanese buses being falsely inferred as cars. However, as such differences regarding products are not within the scope of urban design and planning, a detailed analysis is omitted.

Cause		Form	
		Similarity of appearance	Proximity of location
Based on appearance in the image		Type I	Type II
Based on culture-specific characteristics	Mobile products that appear temporarily in the space	Type III	Type IV
	Fixed configuration of the street space	Type V	Type VI

Table 6. The typology of false inferences.

Based on this, we insist that the types of false inference that are affected by the differences in the streetscape characteristics in Germany and Japan are those "based on the characteristics of the streetscape," namely Type V and VI. Table 7 lists false inference types for each of the pairs presented in Table 2, which demonstrated RFIP values higher than 10%. The following section discusses the combinations prone to false inferences based on the culture-specific characteristics of the streetscape and factors that contribute to such inferences.

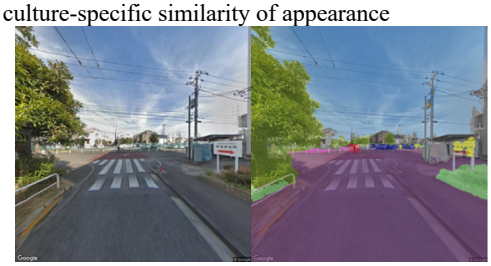
Type	Pair	
	Ground Truth	Inference Result
I	Rider	Person
	Terrain	Vegetation
II	Motorcycle	Building
	Fence	Vegetation
	Pole	Building
	Pole	Vegetation
	Traffic sign	Building
	Traffic light	Building
	Traffic light	Vegetation
III	Truck	Car
	Bus	Car
	Motorcycle	Car
	Motorcycle	Bicycle
IV	-	-
V	Sidewalk	Road
	Wall	Building
	Fence	Building
VI	Terrain	Road

Table 7. Types of false inference for pairs of ground truth elements and inferred classes of which RFIP>10%.

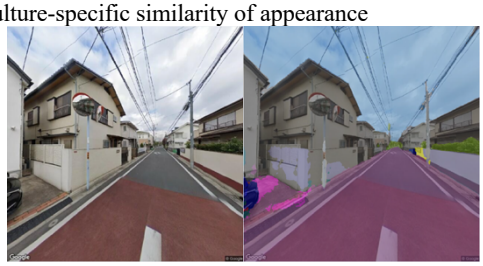
4.2 Characteristics of Japanese Streetscapes and Their Effects

Among the false inference types, we will address those in which the streetscape characteristics are considered a factor (Types V and VI). Three pairs of classes, namely, "sidewalk/road," "wall/building," and "fence/building," were placed under the category of "similarity of appearance based on culture-specific characteristics (fixed configuration of the street space)." The reason why sidewalks and roads are hard to distinguish seems to be that sidewalks and roadways are just separated by low curbs or simple white lines, with no difference in height between the two in many places, as this is permitted under the law due to the de facto difficulty in securing sufficient road space in the existing urban areas with high building density (Figure 3a).

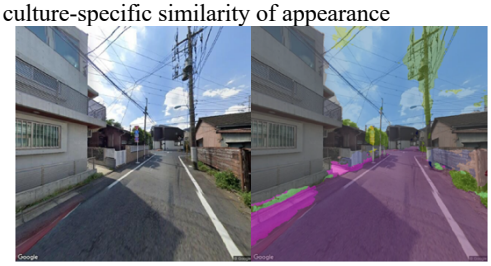
(a) An example of the false inference of sidewalk/road: a type of the culture-specific similarity of appearance



(b) An example of the false inference of wall/building: a type of the culture-specific similarity of appearance



(c) An example of the false inference of fence/building: a type of the culture-specific similarity of appearance



(d) An example of the false inference of terrain/road: a type of the culture-specific locational proximity

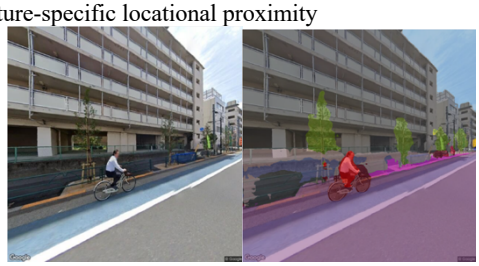


Figure 3. Examples of the false inference based on culture-specific fixed configuration of streets (image/inference result).

Several possible reasons for the "wall/building" (Figure 3b) and "fence/building" (Figure 3c) pairs are the following. Firstly, free-standing structures are often provided to block the street and private property in Japanese urban areas, which is ubiquitous, especially in residential areas. Secondly, structures

found on Japanese roads, such as free-standing stone walls and block walls, have appearances similar to walls of European buildings in terms of colour and material, contributing to the decrease in the inference accuracy. The "terrain/road" pair was determined to fall under the category of "proximity of location based on culture-specific characteristics (fixed configuration of the street space)." The scale of plantings along Japanese roads is usually tiny, so they are perceived as a continuous flat structure along the road (Figure 3d).

The abovementioned characteristics of Japanese streetscapes might influence the analysis of several issues related to people's streetscape evaluation. For example, the perception of safety based on the lack of clear distinction between sidewalks and roadways, the perception of comfort based on the placement of greenery along the street, and the perception of historical and cultural aspects based on Japanese elements such as stone walls, Japanese style hedges, and walls or fences of shrines and temples. When analysing image-based streetscape evaluation for a specific country, researchers should recognize the factors described in this section and consider improving the deep learning model via transfer learning, focusing on the relevant landscape elements.

4.3 Other Landscape Elements That May Influence People's Evaluation of Streetscapes

In this study, we also focused on landscape elements that are classified as "void" in the Cityscapes dataset owing to a lack of need to classify them to develop autonomous vehicles. However, they can potentially impact people's evaluation of streetscapes significantly. Table 8 lists all pairs of landscape element classes in which the top 15 most frequently occurring landscape elements in the void were inferred as another specific landscape element at a rate of 20% or more. They are organized along the same two axes as in Table 4: forms ("similarity of appearance/proximity of location") and causes ("based on appearance in the image/based on culture-specific characteristics").

The characteristics of streetscapes in Japanese cities seem affect the inference accuracy with the following points: As indicated in the previous section, the division of roadway, sidewalk, and greenery on the road cross-section is unclear; the colour and shape of free-standing structures are similar to those of buildings. In addition, there are many colourful billboards, signages, and vending machines in the city, especially on narrow streets. The everyday use of the streets by residents (including the placement of flowerpots) is prevalent as well. These may affect the evaluation of the quality and value of urban places, including safety, convenience, and the strength of community ties. We insist that it is necessary to identify landscape elements that often appear in the streetscape of the country under study, whose presence may indicate cultural characteristics, and create an image recognition model that considers peculiarities when analysing urban places.

5. CONCLUSIONS

In this study, two deep learning models (DeepLab v3 plus and PSPNet), whose accuracies were verified on a German urban streetscape image dataset (Cityscapes), were applied to Japanese streetscape images to verify the inference accuracy for each landscape element. The analysis was conducted based on comparing the values for the IoU metric. We also defined novel indices of the falsely inferred pixels to analyse the likelihood and factors for patterns of false inference that occur between the image recognition classes and landscape elements. The analysis revealed that the accuracy was significantly lower with Japanese streetscape images. Furthermore, the range of accuracy decrease varied depending on the class of landscape elements. Considering that Japan was the only target of this study, it may be necessary to verify whether the accuracy verified using images of urban streetscapes in other cultures is similar. Not all false inferences are the result of culture-specific characteristics, but if culture-specific landscape elements that tend to have significant impacts on the inference accuracy could be identified, image recognition models would be improved to be much more suitable for each country, making empirical studies of the perception and values of places more forcible.

Type	Pair Landscape Element	Inference Result
I	Bollards	Pole
	Triangular cones	Traffic sign
	Plantings	Vegetation
	Flowerpots	Vegetation
	Litter bins	Building
II	Streetlights	Building
	Streetlights	Vegetation
	Streetlights	Sky
	Lightning rods/transformers	Building
	Lightning rods/transformers	Vegetation
	Lightning rods/transformers	Sky
	Bollards	Road
	Bollards	Sidewalk
	Bollards	Building
	Triangular cones	Road
	Triangular cones	Building
	Curbs	Vegetation
	Flowerpots	Sidewalk
	Litter bins	Car
	Antennas	Building
	Antennas	Sky
	Terminal/junction boxes	Building
	Terminal/junction boxes	Vegetation
	Electric wire protection covers	Road
	Electric wire protection covers	Building
	Electric wire protection covers	Fence
	Electric wire protection covers	Vegetation
	Outdoor units of air conditioners	Building
III	-	-
IV	-	-
V	Billboards/signboards	Traffic sign
	Curbs	Road
	Curbs	Sidewalk
	Vending machines	Traffic sign
VI	Billboards/signboards	Building
	Plantings	Road
	Plantings	Sidewalk
	Flowerpots	Building
	Vending machines	Car
	Vending machines	Building
	Door gates	Building
	Door gates	Wall
	Door gates	Fence

Table 8. Pairs of elements and their corresponding false inference types (Those categorized into void areas).

The classification framework for the various false inference patterns introduced in this study will help identify culture-specific characteristics. The framework consists of two axes: the forms (divided into similarity in appearance and proximity of location) and causes (divided into appearance in the image and

culture-specific characteristics). Furthermore, it is possible to distinguish urban design features from others by categorizing culture-specific features into movable products and fixed physical configurations. Using this methodology, researchers can adjust deep learning models considering two aspects: the decrease in model accuracies and impact of culture-specific streetscape characteristics on people's perception and valuation of urban places. Based on the results and analyses presented in this study, a future research direction is to develop and implement more accurate and tailormade models to understand people's perceptions of urban spaces, as well as the value that places have, by using big data such as street-level images.

REFERENCES

Aravot, I. (2002). Back to phenomenological placemaking. Journal of Urban Design, 7(2), 201-212.

Brown, G. (2004). Mapping spatial attributes in survey research for natural resource management: methods and applications. *Society and natural resources*, 18(1), 17-39.

Brown, G., & Weber, D. (2012). Measuring change in place values using public participation GIS (PPGIS). *Applied Geography*, 34, 316-324.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834-848.

Cityscapes Dataset. (2022). Benchmark Suite. https://www.cityscapes-dataset.com/benchmarks/ Accessed Apr 29, 2022.

Corburn, J. (2009). Toward the healthy city: people, places, and the politics of urban planning. Mit Press.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213-3223).

Cresswell, T. (2014). Place: an introduction. John Wiley & Sons.

Faludi, A. (2013). A reader in planning theory (Vol. 5). Elsevier.

Francis, M. (1989). Control as a dimension of public-space quality. In Public places and spaces (pp. 147-172). Springer, Boston, MA.

Friedmann, J. (2010). Place and place-making in cities: A global perspective. *Planning Theory & Practice*, 11(2), 149-165.

Gehl, J. (2013). Cities for people. Island press.

Grant, J. (2005). Planning the good community: New urbanism in theory and practice. routledge.

Hall, P. (2014). Cities of tomorrow: An intellectual history of urban planning and design since 1880. John Wiley & Sons.

Healey, P. (2010). Making better places: The planning project in the twenty-first century. Macmillan International Higher Education.

Home, R. K. (1990). Town planning and garden cities in the British colonial empire 1910–1940. *Planning Perspective*, 5(1), 23-37.

Hudson, B. M., Galloway, T. D., & Kaufman, J. L. (1979). Comparison of current planning theories: Counterparts and contradictions. *Journal of the American planning association*, 45(4), 387-398.

Irving, A. (1993). The modern/postmodern divide and urban planning. *University of Toronto Quarterly*, 62(4), 474-487.

Katz, P. (1994). The New Urbanism. Toward an architecture of community.

Lynch, K. (1964). The image of the city. MIT press.

Lynch, K. (1984). Good city form. MIT press.

Madden, D. J. (2010). Revisiting the end of public space: Assembling the public in an urban park. *City & Community*, 9(2), 187-207.

Manzo, L. C., & Perkins, D. D. (2006). Finding common ground: The importance of place attachment to community participation and planning. *Journal of planning literature*, 20(4), 335-350.

Ordonez, V., & Berg, T. L. (2014, September). Learning high-level judgments of urban perception. In *European conference on computer vision* (pp. 494-510). Springer, Cham.

Peterson, J. A. (2009). The birth of organized city planning in the United States, 1909–1910. *Journal of the American Planning Association*, 75(2), 123-133.

Proshansky, H. M., Fabian, A. K., & Kaminoff, R. (1983). Place-identity: Physical world socialization of the self. *Journal of Environmental Psychology*, 3(1), 57-83.

Relph, E. (1976). Place and placelessness (Vol. 67). London: Pion.

Ritchie, H., & Roser, M. (2018). *Urbanization*. Our world in data.

Sorensen, A. (2005). The making of urban Japan: cities and planning from Edo to the twenty first century. Routledge.

Tuan, Y. F. (1979). Space and place: humanistic perspective. In *Philosophy in geography* (pp. 387-427). Springer, Dordrecht.

United Nations. (2018). Revision of world urbanization prospects. United Nations: New York, NY, USA, 799.

Ward, S. V. (2004). Planning and urban change.

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180, 148-160.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881-2890).