Who Sees What? Examining Urban Impressions in Global South Cities

Luis Emmanuel Medina Rios, Salvador Ruiz-Correa, Darshan Santani
and Daniel Gatica-Perez

Abstract While the study of how people perceive urban environments according to subjective dimensions like beauty and danger is a classic topic in environmental psychology, only in recent years computing research has addressed these questions with an interest in machine recognition of such kinds of urban impressions. In this chapter, we describe some of the basic concepts related to this emerging domain, which is at the intersection of computing and human perception. Then, with a specific focus on Global South cities, we present a study on perception of urban scenes by people and machines. The study is conducted on a dataset of 1,200 images collected in three Mexican cities of different characteristics and six urban impression variables. First, we conduct a comparative analysis of crowdsourced impressions generated by local and non-local observers. The analysis shows that non-local observers reach higher inter-observer agreement compared to local ones, and that these two groups tend to disagree on what they report to perceive with respect to dimensions like danger and interest. Second, we show that visual cues extracted from deep neural networks and trained with impression labels provided by both local and non-local observers result in systems for which non-local judgments are better inferred that local ones. Based on these findings, the chapter discusses implications for the design of systems that use crowdsourced subjective labels for machine learning and inference on urban environments.

Luis Emmanuel Medina Rios
EPFL, Switzerland, e-mail: luis.medinarios@epfl.ch

Salvador Ruiz-Correa
IPICYT, Mexico e-mail: salvador.ruiz@ipicyt.edu.mx

Darshan Santani
Pepper Cloud, Singapore e-mail: dsantani@peppercloud.com

Daniel Gatica-Perez
Idiap Research Institute and EPFL, Switzerland e-mail: gatica@idiap.ch

1 Introduction

The online availability of large-scale urban imagery coming from social media or Google Street View, in combination with crowdsourcing-based image labeling and machine learning for visual recognition, are offering the possibility to build systems that can reason about a variety of urban phenomena [49, 6, 16]. In particular, understanding how people perceive and experience the urban environments we inhabit or visit - in aesthetics, affective, and social terms - is relevant for ubiquitous computing given the multiple connections between urban perception and personal and community well-being, e.g., the value of spending time in environments perceived as restorative [17, 44, 29].

The state-of-the-art on machine recognition of urban perception (i.e., identifying if a place is perceived as safe, beautiful, or interesting) has essentially followed two stages: data labeling of perceived attributes by online crowdworkers who look at urban scenes (either volunteers [49, 43] or remunerated workers [50]); followed by supervised learning methods that use such labels as ground-truth and attempt to generalize to unseen data [36, 38, 16]. An increasing body of evidence is showing that blindly treating perceived attributes in this way can be problematic, and that research needs to account for the variety of contexts that affect human perception [51, 32, 31]. Two fundamental issues in urban perception are: (1) how different kinds of observers might differ in their urban perceptions given their prior exposure to a place (i.e., locals vs. visitors) and thus produce different aesthetics or affective labels about the same scenes; and (2) what is the effect of these differences once they are implemented in machine inference systems. This is especially true when applying algorithms to world regions that are not well represented in digital terms, as is the case with many countries in the Global South [53, 14]. This also has important implications for urban analytic systems trained from crowdsourced data, given an (implicit or explicit) assumption in much of the current literature, namely that a model trained on a specific set of cities and human observers might generalize to other cities and to aesthetic or affective impressions by other people [36, 38, 16].

In the context of Global South cities, and using three cities in Mexico as a concrete case study, we address two research questions:

RQ1: Do local and non-local observers agree on the perception of urban dimensions in such cities? If not, what are the dimensions for which differences in perception are the largest, and what can explain such differences?

RQ2: Are there differences in the performance of machine inference systems trained to recognize urban scenes as locals perceive them, compared to how non-locals do? Are generic deep learning features equally effective for the two cases?

In this chapter, we address the above questions using the following approach. First, on a dataset of 1,200 images collected from three cities in central Mexico, which correspond to different examples of urban density and economic activity, we collected a set of crowdsourced labels of urban impressions volunteered by young local inhabitants along six dimensions, namely *dangerous*, *dirty*, *interesting*, *pleasant*, *polluted*, and *pretty*. Local observers viewed and annotated images in an

online setting comparable to the one used in our previous work [50], where we gathered impressions from Amazon Mechanical Turk (MTurk) workers.

Second, to characterize the content of the urban scenes under study, we extracted visual cues using both manual coding of high-level attributes derived from environmental psychology literature, and three versions of machine features extracted from convolutional neural networks (CNNs) pre-trained on large-scale scene data. This diversity of visual cues allows us to compare across representations of urban scenes: while manual cues correspond to semantic descriptors that one could expect to find in less privileged areas of cities worldwide (e.g. neglected vegetation), the CNN-derived features correspond to either scene types (e.g. alley) or object types (e.g. sky) present in urban images.

Third, we conducted a comparative analysis of urban perception by local and non-local online observers, using the impressions provided by the two groups of observers. Based on this analysis of crowdsourced labels, we find that non-local observers reach higher inter-observer agreement compared to local ones for most dimensions; and that these two groups disagree on what they report to perceive with respect to dimensions like danger and interest; namely, locals tend to perceive urban scenes as more dangerous than non-locals; while non-locals tend to see urban scenes as more interesting than locals. This result confirms, using 10 times more image data, a preliminary finding reported in our previous work [51]. Furthermore, we use additional sources of information, including open text provided by MTurk workers and in-situ discussions with local observers, to suggest plausible explanations for this finding.

Finally, we conducted a systematic evaluation of machine inference of urban perception variables in a regression setting. Using eight models corresponding to two sources of human labels (local and non-local) and four sources of visual cues, we find that (1) inference systems trained on impression labels provided by non-local observers result in higher performance (measured by the standard R^2 coefficient of determination) for three dimensions in the case of manual visual cues, and all six dimensions in the case of CNN features; (2) CNN features outperform manual cues for all the six urban perception dimensions for regressors trained on non-local labels, and (3) in contrast, manual visual cues outperform CNN features for all six urban dimensions when regressors are trained on labels by local observers. We discuss possible reasons for these trends, and also discuss about the implications of our findings for computing systems that use crowdsourced generation of subjective labels to analyze urban environments.

The chapter is organized as follows. Section 2 reviews related work and frames our work within the existing literature. Section 3 summarizes our methodology. Section 4 describes the image data and the protocol for collection of crowdsourced urban perception. Section 5 describes the methods used for manual and automatic extraction of visual cues from urban images. Section 6 presents the comparative analysis of urban perception by local and non-local observers. Section 7 presents and discusses the results of inferring urban perception. Section 8 discusses the implication of our findings. Section 9 provides final remarks.

2 Related Work

In this section, we review work related to perception of urban attributes using crowdsourcing, and machine recognition of urban perception from visual data.

2.1 Perception of Urban Attributes using Crowdsourcing

In environmental psychology, urban planning, and architecture, the visual assessment of landscapes (including urban ones) has used a suite of well-established methodologies [24, 13], spanning the in-situ observation of environments [48], the use of on-street pedestrian surveys [39], and the assessment of both real and synthetic urban imagery [29]. Online, geo-referenced imaging resources like Google Street View (GSV) have been more recently used as part of visual assessment tools of urban environments [7]. Crowdsourcing research has also proposed to use GSV in combination with online crowdsourcing to collect labels of urban perception, making use of the large-scale nature of GSV and the potential availability of online workers who can observe images and provide their impressions [49, 43]. In the work by Salesses et al. [49], a set of 4,000 GSV images from four cities, two in the US (New York City and Boston) and two in Austria (Salzburg and Linz) were labeled with respect to three dimensions: *class* (later renamed as *wealth*), *safety*, and *uniqueness*. This dataset (dubbed Place Pulse 1.0) was labeled by online volunteers using a pairwise procedure, where pairs of images were relatively ranked with respect to each of these attributes. In the work by Quercia et al. [43], following similar goals and techniques, a dataset of 500 GSV images from London were labeled for three attributes: *beauty*, *happiness*, and *quietness*. In the work by Dubey et al. [16], a large dataset of pair-wise rankings for 100K GSV images from 56 large cities (Place Pulse 2.0) was sparsely labeled for six attributes: *beautiful, boring, depressing, lively, safe,* and *wealthy*. While the attributes used in these works have been adapted from existing environmental psychology literature, no systematic methodological justification for their choice was provided.

Most of the research described above has focused on cities in the US and Western Europe. This opens an opportunity to study cities in the Global South with similar approaches, as three quarters of the 100 largest populated urban areas worldwide are in the Global South [2]. However, key factors need to be reconsidered. In many cities, online imaging resources like GSV might not have wide coverage. The use of phones as infrastructure to collect urban images becomes an attractive alternative in these cases [46]. Furthermore, citizens from countries in the Global South do not often have the same access to online platforms to contribute annotations (e.g. Amazon's Mechanical Turk is not available to workers in most countries). This can pose significant limits to the collection of crowdsourced urban impressions from local inhabitants. Some work in these two directions (collecting crowdsourced impressions about Global South cities, and collecting urban impression labels specifically from locals) has been proposed in our previous work [50, 51] for the case of Mexico and by

Candeia et al. [10] for Brazil. Although not necessarily focused on urban perception as studied here, there has been considerable work on citizen participation for mapping and community-related purposes in Latin America [37, 8, 1, 11], Asia [55], and Africa [3]. The starting point for the work described in this chapter is the dataset from our previous work [50]. This consists of 1,200 urban images collected in three cities in Central Mexico of diverse characteristics, for which we have additionally collected online impressions from local volunteers for six attributes, and which allows us to systematically compare the views of local inhabitants and AMT crowdworkers (*dangerous*, *dirty*, *interesting*, *pleasant*, *polluted*, and *pretty*.) As shown in [50], these variables provide some coverage of the circumplex model of affect for environments [48].

2.2 Situated Crowdsourcing and Local Knowledge

Another related topic is situated crowdsourcing. This form of crowdsourcing exploits the availability of users to provide on-demand information via input devices like public displays embedded in the physical space [20]. This view of crowdsourcing has seen interest in the last few years [33, 21, 19, 20] in the context of citizen participation and healthcare. Situated crowdsourcing requires available local contributors, who can respond to the needs of the task requester. Specifically to our research, some work has been done to understand the differences between local and non-local contributors with respect to performance in tasks that may need local knowledge. In the work by Goncalves et al. [19], it was shown that situated crowdsourcing involving local contributors (university students interacting with large displays deployed on campus) was comparable to online crowdsourcing involving MTurk workers, in terms of task accuracy and task uptake rate, for tasks that did not require local knowledge (e.g., counting malaria infected blood cells in medical images). In other related work [20], crowdsourced tasks that required local knowledge of the city were compared to more generic tasks. Some of the tasks involved photo taking. This work found that certain tasks requiring local knowledge were more attractive to the participants, as they gave them an opportunity to share such knowledge. In contrast to the above work, in which task performance can be objectively estimated given the fact-finding nature of the task (e.g. counting cells), we collect online judgments of urban impressions from locals and non-locals. In principle, many of these attributes have no unique ground-truth, so we can expect differences between the two types of observers, given their exposure to specific places.

2.3 Machine Recognition of Urban Perception Attributes

Approaches for automatic recognition of urban perception attributes have been proposed, using the datasets described in section 2.1. The visual cues used in previous

work have ranged from standard low-level descriptors, including color and a variety of texture features like Histogram of Oriented Gradients (HOG) [12] and Scale-Invariant Feature Transform (SIFT) [30], to more recent CNN-derived features. In the work by Naik et al. [36] and Ordonez et al. [38], recognition of the Place Pulse 1.0 urban perception attributes (*safety*, *uniqueness*, *wealth*) was done for New York City and Boston. In the work by Porzi et al. [42], a study on automatic pairwise ranking of urban images was conducted for the *safety* attribute from the four cities of Place Pulse 1.0 using a CNN. This approach did not provide an automatically generated rating for a place, but rather compared pairs of places, which departs from the standard way of reasoning about place attributes in environmental psychology, in which places are independently rated [48]. In the work by Arietta et al. [6], a method that also used GSV images was proposed to automatically infer urban attributes (e.g. housing prices) from visual cues (HOG+color). In the work by Gebru et al. [18], image data from GSV processed with CNNs to extract vehicle semantic descriptors (brand, model, and year) was used to find connections between neighborhoods in 200 US cities and income indicators (an attribute connected to the *wealth* dimension studied in previous work). Finally, in the work by Dubey et al. [16], the Place Pulse 2.0 dataset was used to compare pairs of city scenes using a CNN for the six available urban perception dimensions. This approach thus follows the same idea of [42] of inferring relative rankings among pairs of places, reporting a pairwise accuracy of 73.5%. This is one of the few works reporting automatic inference of urban perception attributes on data from Global South cities, although the paper itself does not discuss the specific performance on such cities. A second exception is our previous work [52], which presented an approach based on visual feature extraction using a pre-trained CNN, followed by a second regression module to infer urban perception attributes collected from AMT workers as presented in [50].

We have used our previous work [52] as starting point, and substantially extend the studied visual representations, which include both high-level descriptors of urban scene components manually generated by young observers, as well as three automated CNN-derived image representations. Furthermore, we study the effects of training CNNs with urban perception labels generated by local and non-local observers, and quantify the differences in performance obtained for each observer group. As we discuss later in the paper, this has important implications for AI systems for urban analytics trained from crowdsourced data, given the assumption in much of the current literature [36, 38, 16], that a model trained on a specific set of cities and human observers will generalize to other cities and to impressions by other people.

3 Methodology

Our methodology involves the following stages: selection of urban perception labels; collection of impressions; extraction of visual cues; and inference from visual cues. Each stage is summarized as follows.

Selection of urban perception labels (Section 4): Six attributes were chosen following the methodology in [46] and [51]: *dangerous*, *dirty*, *interesting*, *pleasant*, *polluted* and *pretty*. These attributes describe the scene and environment in which the image dataset is based. Let us recall that the state of Guanajuato, as a tourist destination in Mexico, can elicit different responses from observers, from negative ones (e.g., concerns about streets with tag graffiti) to positive ones (e.g., enjoyment of the colonial architecture of some buildings, see Figure 1 for examples). In the rest of the chapter, we use the terms *urban perception label* and *label* interchangeably.

Collection of impressions (Section 4): To collect impressions of urban visual attributes, we use two approaches: (1) a crowdsourcing task where the annotators were local with respect to the cities presented in the image dataset; and (2) a crowdsourcing task in which the raters were part of a foreign-born population that did not know anything specific about the images they assessed.

Extraction of visual cues (Section 5): Two different methods to extract visual cues are used: CNNs and manual coding. First, CNNs are an example of what can be automatically extracted with current computer vision methods. Concretely, we use 3 different pre-trained CNNs (DilatedNet Semantic Segmentation: 150 features; GoogLeNet places205: 205 features; and GoogLeNet places365: 365 features) to extract visual cues at the object- or scene-level, based on the final layer of class probabilities using the Caffe [23] framework. Before extraction, images were re-sized to 256x256 pixels and pre-processed by mean image subtraction. In contrast, manual coding provided by local observers, following a procedure adapted from the Block Environmental Inventory (BEI) [40], is a way to explore specific high-level visual cues that are not included in the set of features extracted by the pre-trained CNNs. Throughout the chapter, we use the terms *visual cues* and *features* interchangeably.

Inference from visual cues (Section 6): The automatic inference methods produce continuous values for each inferred urban perception attribute. Specifically, we use Random Forests to implement a regression task in which the dependent variables are the urban perception labels and the independent variables are the visual features. To measure the performance of the regression task, we use the coefficient of determination (R^2). We use cross-validation with k=10 folds and report the mean over these 10 runs.

4 Data: images and impressions

4.1 Image dataset

We conduct our study on the image corpus generated in the Urban Data Challenge (UDC) explained in [50], where 7,000 images were taken from a first-person perspective by young student volunteers with their smartphones in the state of Guanajuato (located in central Mexico with a population of about 6 million inhabitants, most of them urban (70%).) Students attended a public high-school in Guanajuato, the *Colegio de Estudios Científicos y Tecnológicos del Estado de Guanajuato* (CE-

CYTE). This high school provides education on science, technology, and humanities to low-income youth living in Guanajuato City and surrounding areas. Students were altruistically motivated and eager to contribute to improve the understanding of their city. A partnership that included school authorities, teachers, parents, and a local research team supported the experiment.

From the collected images, 1,200 were selected in groups of 400 images for each of the three cities that were chosen to collect the images: (1) Guanajuato City (170,000 inhabitants), which is a UNESCO world heritage site and whose economic activity relies mostly on tourism – a fact that makes local inhabitants care about the image of their city; (2) Leon City (1.6 million inhabitants, the seventh most populated metropolitan area in Mexico), which is an industrial and business center with factories specialized on leather and footwear products; and (3) Silao City (147,000 inhabitants), which is an industrial city, with industrial parks and automotive component companies due to the presence a major US car assembly plant. The images were collected in 2015 during daytime.

The volunteers participating in the image data collection tried to capture in pictures the characteristics of each of the three cities, documenting different neighborhoods and iconic places. Volunteers ventured in many of the neighborhoods of the city, except those in the suburbs of two of the cities (known to be unsafe.) Please refer to [50] and [47] for a detailed explanation of the urban image collection process. Examples of the images taken can be seen in Figure 1.

Fig. 1: Samples from the image corpus. The shown images were selected randomly; each row represents a city. Top: Guanajuato; Middle: Silao; Bottom: Leon. For privacy reasons, images have a reduced resolution.

4.2 Impressions by local observers

In our experimental protocol, 120 additional student volunteers from CECYTE provided the local impressions (43 women and 77 men). At the time of the study, 80% were 16 years old, 15% were 17 years old, and 5% were 18 years old. To gather annotations, we followed a protocol similar to the one used in our previous work [51]. After conducting recruiting activities for a one-month period, 120 volunteers were chosen from a school population of 1,100 students. Each volunteer was required to have a signed parental approval and travel insurance to take part in the image labeling experiment. The experiment was conducted over a period of a month, in which groups of about 40 students visited computing facilities at IPICYT to perform the task. During the visit, students were given a meal and a guided tour to the computing facilities, after which they performed the labeling experiment. Data collected from volunteers during the experiment was anonymized, and personal information (age and gender) was only gathered for general statistics but not linked to each person's annotations.

The gathering of local impressions was conducted through a custom-built website, comparable in basic functionalities to the one used by MTurk workers (see next subsection). Six urban perception attributes were labeled during the experiment (*dangerous*, *dirty*, *interesting*, *pleasant*, *polluted* and *pretty*). In comparison to previous work, Salesses et al. [49] labeled three attributes (*safety*, *uniqueness*, and *wealth*), two of which are also covered by our work with different names and in some cases inverted scale (*dangerous* for *safety*). Quercia et al. [43] labeled three attributes (*beauty*, *happiness*, and *quietness*), one of which is also covered by our work with a different name (*pretty* for *beauty*). Dubey et al. [16] labeled six attributes (*beautiful*, *boring*, *depressing*, *lively*, *safe*, and *wealthy*), three of which are covered by our work with different names and sometimes inverted scale (*dangerous* for *safe*; *interesting* for *boring*; *pretty* for *beautiful*). In summary, three of our attributes have not been studied in these previous works (*dirty*, *pleasant*, *polluted*), while the other three have been studied. Besides the labels, students also labeled 10 high-level semantic descriptors following the procedure described in section 5.2.

Each volunteer worked independently with a high resolution monitor to annotate 50 images. A team of three supervisors facilitated the process. On average, each volunteer spent 1.5 hours to complete all annotations. Five annotations from independent observers were gathered per image, per label, and per semantic descriptor. A seven-point Likert scale was used to assess the urban perception labels (from 1: Strongly disagree to 7: Strongly agree). The online experiment followed the Code of Ethics and Conduct of the British Psychological Society.

In the online experiment, students were not given any information about the urban place being displayed. However, we acknowledge that some of the images contain identifying characteristics of the captured places, such as signs in Spanish as well as other details (e.g. flags) that might give away the location to some degree. It is known that people can pinpoint where a photo was taken [41]. To quantify this issue, we manually annotated the 1,200 image corpus for the following attributes: presence of people passing by, who could have dressing style characteristics more

commonly expected in certain neighborhoods of a big city or in a small town [34]; and presence of signs in Spanish, which could help observers identify specific local businesses. Table 1 shows the number of images of our dataset that contain these attributes. While the presence of passersby is frequent (44% of the images), signs are substantially less common (4.8%.)

Feature	Cities			Total	Percentage (N=1200 images)
	Silao	Leon	Guanajuato		
People	168	291	70	529	44.08
Signs in Spanish	41	16	1	58	4.83

Table 1: Number and percentage of images containing potentially identifying place characteristics.

4.3 Impressions by non-local observers

Crowdsourcing of non-local impressions was conducted through Amazon's Mechanical Turk, in a process originally discussed in [50]. 146 US-based Master workers, with a minimum of 95% approval rate, were chosen to complete the corresponding HITs (Human Intelligence Tasks). Every HIT consisted of observing one image and providing impressions for each of the six urban perception labels. All annotations were done on a seven-point Likert scale (from 1: Strongly disagree to 7: Strongly agree), just as local observers did. The workers were not told about the source or location of the images, or about the cities in the study, to reduce potential bias. A number of 10 annotations were gathered per image and per label, making a total of 144,000 individual judgments.

Regarding demographics, 77 of the workers answered a post-task survey. For gender, 58% of respondents were women and 42% were men. For ethnicity, 80% of respondents were White/Caucasian, 12% Asian, 3% Hispanic/Latino, and 3% Black/African American. For age group, the distribution was as follows: 3% of respondents were 18-24 years-old, 32% were 25-34 years-old, 43% were 35-50 years-old, and 22% were 50+ years-old. Regarding residence, 18% lived in a big city, 18% in a small-to-mid-sized town, 45% in the suburbs, and 18% were rural. Finally, 23% of respondents reported visits to developing countries, and 44% of these respondents (i.e., 10% of the total workers) had visited Mexico. This last response was collected via free text (i.e., no written clues about the country where the images were taken were provided by the experiment.) In sum, the combination of low self-reported Hispanic/Latino ethnicity (3%) and the low number of people reporting previous visits to Mexico (10%) suggests that the MTurk workers who participated in our task qualify overall as non-local observers. More details about the online crowdsourcing task can be found in [50].

5 Visual cue extraction

In this section, we describe the visual cues, both machine-generated features and manually labeled semantic descriptors used in the study.

5.1 Extraction of visual cues via CNNs

Using CNN architectures, we extract the following features:

DilatedNet Semantic Segmentation: 150 features. This consists of 150 visual features resulting from applying deep learning techniques, using a pre-trained semantic segmentation network called DilatedNet [56]. The features include both indoor and outdoor generic elements such as wall, building, sky, floor, tree, etc. The images are described with the actual proportion of each of the features.

GoogLeNet places205: 205 features. This is based on the extraction of 205 visual cues, using the final layer with class probabilities of a pre-trained CNN based on the GoogLeNet architecture trained on the places205 database [58]. This popular database, in its first version, allows to describe an image with place-related labels related to our image dataset, like alley, basilica, corridor, church, residential neighborhood, etc.

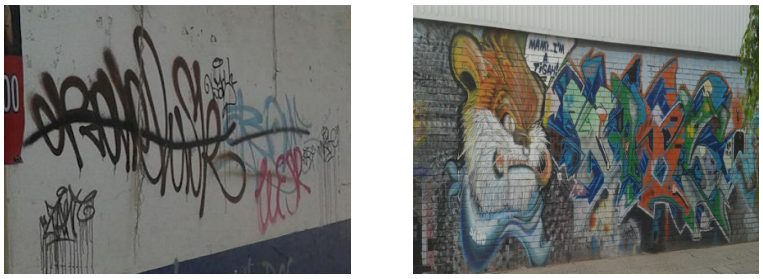
GoogLeNet places365: 365 features. The places365 database is an update of the places205 database and consists of around 1.8 million of training pictures. In this case, it uses the same CNN architecture (GoogLeNet), but trained on the new database [57]. This contains 365 visual features that were extracted by using the final layer with class probabilities. The features include almost the 205 from the places205 database and approximately 160 more, including categories like bazaar, downtown, flea market, house, industrial area, junkyard, park, promenade, etc. We decided to use both versions of the places databases to examine concrete benefits of using a more expressive vocabulary of scene labels during the automatic inference experiments.

5.2 Extraction of visual cues via manual coding and local annotation

While CNNs offer good performance in terms of extracting object-level and scene-level elements from the pictures, other high-level visual cues can evoke certain atmospheres. For example, how people can perceive danger is discussed in [9] through the concept of perceived personal danger, i.e., the general fear of people to become a victim. More specifically, people walking at night might look at physical elements that can become an obstacle from escaping in case of danger (blocking elements or enclosed spaces), and also look at places with poor outdoor lighting. Other authors studied the relationship between tranquility and danger by conducting correlation analyses for natural and urban settings and three elements, namely nature,

openness, and degree of care [22]. These three elements (nature, openness and care) are very general, yet one could relate them with semantic cues, e.g. *nature* can be related to the presence of visual cues such as plants, trees, or grass. *Openness* can be related to the presence of parks, plazas, or wide streets. Finally, *care* could be negatively related to visual cues such as litter on the street.

To manually extract a set of visual cues relevant to our dataset, we implemented a two-stage process. In the first stage, we defined a set of high-level visual cues. For this, we did a manual analysis based on a sample of images for two of the six labels, namely *dirty* and *dangerous*. We decided to focus on these two labels as they could correspond to visual concepts that are not sufficiently represented in existing databases used to train CNN models [53]. This procedure was based on a random sample of images for which attention was paid to understand specific visual cues, for example, understanding differences between tag graffiti (Figure 2a) from artistic graffiti (Figure 2b). This task was implemented by the first author, who was born and raised in Mexico, and therefore understands the context of the pictures. We then clustered the features defined from the above procedure using the *Block Environmental Inventory* (BEI) defined in [40], which uses three types of physical cues for *incivilities, vandalism and dilapidated houses*; *signs of territorial functioning*; and *defensible space features*. Finally, we found 10 clusters of high-level visual cues that are part of the BEI and used them as part of the image annotation procedure described in section 4.2. In alphabetical order, the semantic descriptors are: *deteriorated roads*; *lack of maintenance*; *lack of outdoor lighting*; *lack of security elements*; *littering*; *neglected vegetation*; *poor urban planning*; *unkempt houses/buildings*; *vacant lots*; and *vandalism*.



(a) Tag graffiti (b) Graffiti

Fig. 2: Examples of tag graffiti and artistic graffiti.

In the second stage, we gather the annotations of these 10 semantic descriptors in a binary way for all images (1 if the semantic descriptor appears in the image, and 0 otherwise) as part of the annotation process by student volunteers described in Section 4.2. Finally, for each picture and for each semantic descriptor, we aggregate the annotations to compute the mean over all annotators. The result is a number between 0 and 1 for each of the semantic descriptors, that can be seen as an empirical

probability of the descriptor appearing in the picture. As mentioned earlier in this section, the process above was implemented only for two of the six attributes we analyze. However, as we show later, these manual cues have also explanatory power for some of the other labels.

6 Comparing impressions between local and non-local observers (RQ1)

In this section, we first present an analysis of the reliability of the collected impressions, followed by the presentation of their descriptive statistics; a comparison of the impressions for local and non-local observers; and a qualitative analysis of the collected impressions, which complements the quantitative analysis.

6.1 Annotation quality: inter-rater reliability

To measure the quality of the labels in terms of inter-rater reliability, we computed intraclass correlation (ICC) on each of the labels across the image corpus [27]. Since each of the images is rated by a set of k annotators randomly selected from a larger population of K annotators [51], we chose the *One-Way Random-Effects Model* [27], known as *ICC(1,k)*.

The inter-rater agreement of the non-local observers was detailed in [50]. However, since we only consider 5 annotations per image for the local observers and to be fair in the comparison, we computed *ICC(1,k)* by randomly sampling 5 (out of 10) annotations, doing this 10 times and getting the mean and standard deviation over these 10 values. Results of the computed *ICC(1,k)* are reported in Table 2. Note that throughout this section, we will also use the results obtained in [51], in which another set of 99 images from Guanajuato City was annotated. With respect to other previous work on crowdsourced urban perception [49, 43, 16], note that the way in which the labels were obtained is different. The three works above used pair-wise comparisons, i.e., observers saw pairs of images and gave a relative ranking for the pair with respect to each attribute. In contrast, we collect individual ratings, where each image is assessed independently. This procedure allows to estimate measures of inter-observer agreement like ICC, unlike [49, 43, 16], which do not report such measures of agreement.

According to Table 2, we see that for k=*5* annotators there is higher agreement among non-locals compared to locals. ICC values indicate moderate reliability (values between 0.5 and 0.75) for all labels except for *polluted* [27]. Note that we use the term reliability in the usual sense in psychology (i.e., a measure of inter-observer agreement). For locals, three labels indicate moderate reliability: *interesting*, *pleasant* and *pretty*, while *polluted* is the label with the lowest agreement among raters. Label *pleasant* (resp. *pretty*) achieved the highest agreement among non-locals (resp.

Label	Non-locals			Locals	
	k=5 Mean±SD	*k=10* [52]	*k=10* [51]	*k=5*	*k=10* [51]
Dangerous	0.62±0.01	0.76	0.83	0.34	0.63
Dirty	0.65±0.01	0.78	0.85	0.36	0.68
Interesting	0.54±0.01	0.70	0.63	0.52	0.70
Pleasant	0.67±0.01	0.79	-	0.56	-
Polluted	0.46±0.02	0.64	-	0.28	-
Pretty	0.61±0.01	0.73	0.83	0.58	0.80

Table 2: *ICC(1,k)* scores, including standard deviation and mean values for the non-local case . Please note that the "-" indicates that the label was not included in the corresponding study.

locals). Comparing the two groups for *k=10*, we see a similar tendency as shown in [51]: non-local raters tend to agree more for most of the labels compared to local raters. Note that by definition the ICC values are higher for larger k.

6.2 Descriptive Statistics

We follow the procedure described in [51] to aggregate the values of the scores: the annotations rely on an ordinal scale (that also describes a ranking), and knowing that one of the statistics to get the central tendency of an ordinal scale is the median [54], we compute it on each of the 5 scores per image. Given these median scores, we compute the mean and the standard deviation per label using the image corpus.

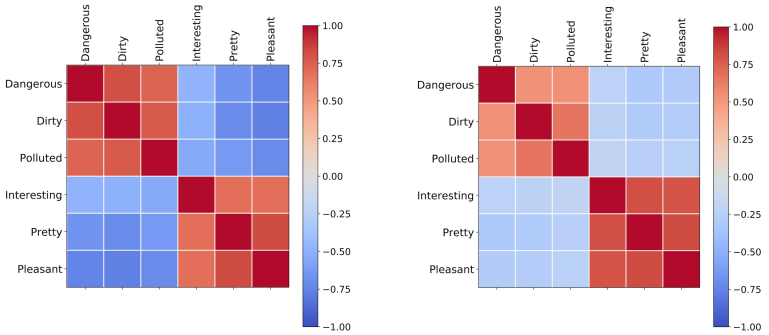
Label	Non-locals		Locals	
	Mean±SD [52]	Mean±SD [51]	Mean±SD	Mean±SD [51]
Dangerous	2.98±1.00	3.19±1.20	3.78±1.45	4.43±0.91
Dirty	3.16±1.10	3.25±1.26	3.53±1.53	4.33±1.24
Interesting	3.84±0.90	4.14±1.10	3.00±1.54	3.55±1.23
Pleasant	3.82±1.00	-	3.08±1.58	-
Polluted	2.89±0.90	-	3.39±1.38	-
Pretty	3.11±1.00	3.25±1.36	3.04±1.65	3.47±1.38

Table 3: Means and standard deviations of the annotation scores for each label and group. Please note that the "-" indicates that the label was not included in the study.

Table 3 shows the descriptive statistics of the labels. All mean scores for both locals and non-locals, on the 1,200 image corpus, show a trend towards disagreement, as values are below 4 on the seven-point Likert scale, which corresponds to *somewhat disagree*). Examining previous work, when comparing the two groups based on the 99 image corpus in [51], we see that local observers overall tend to perceive

the images as more *dangerous*, *dirtier* and *prettier*, while the non-local observers tend to perceive the images as more *interesting*. We find the same pattern in our data, except for the label *pretty*. Regarding *polluted* and *pleasant*, we see that locals perceive images as more *polluted*, while non-locals perceive them as more *pleasant*. A detailed comparison is presented in Section 6.3.

We also perform a Spearman's correlation analysis of the local and non-local median scores (Figure 3a), and we get similar results as with the non-local people's study in [52] (Figure 3b): There are two groups of labels which are positively correlated among them and negatively correlated with the other group. These are *interesting*, *pretty*, and *pleasant* (referred to as positive labels in the rest of the analysis); and *dangerous*, *dirty*, and *polluted* (referred to as negative labels.) We see that correlation absolute values are generally higher for the non-local annotations than for the local ones.



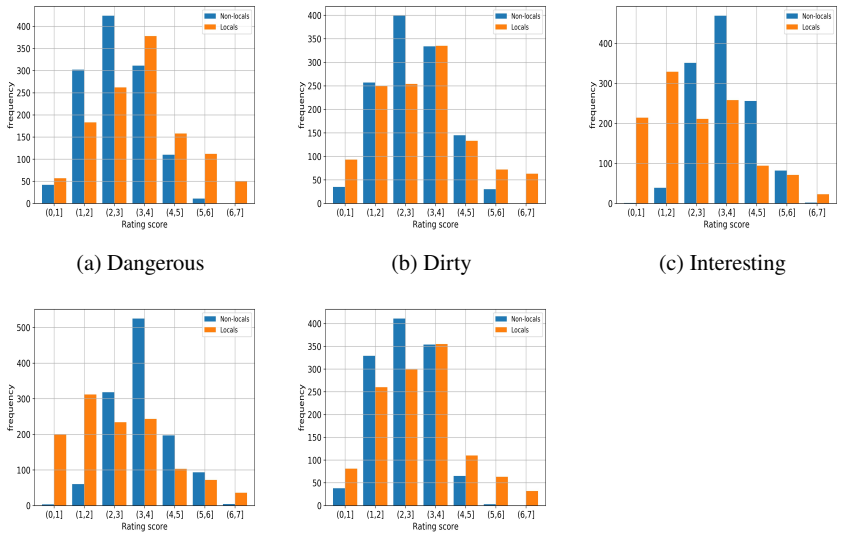
(a) Non-locals (b) Locals

Fig. 3: Correlation matrix of the six urban perception labels for (a) non-locals, and (b) locals (N=1200, $p<0.05$ for all entries of the matrix).

6.3 Comparing Impressions between Groups

6.3.1 Pair-wise analysis

We now compare the impressions between the two groups of observers. We want to understand if the mean difference of the labels between the two groups is statistically significant. We perform the Tukey's Honest Significant Difference (HSD) test, and present it in Table 4. Furthermore, to compare our findings, we also include the results of our previous work, which investigated such differences on a smaller, 99-image dataset [51]. We complement this by plotting the distribution of the perception scores between the two groups of raters in Figure 4.



(a) Dangerous (b) Dirty (c) Interesting

(d) Pleasant (e) Polluted

Fig. 4: Plots comparing the distributions of perception scores between non-locals (blue) and locals (orange). Only those labels whose mean difference in the Tukey's HSD test were statistically significant at $p<0.05$ are included.

Based on the statistics summarized in Table 4, we observe that:

1. Images were perceived as more *dangerous*, *dirtier*, and more *polluted* by the local observers compared to non-locals (local-to-non-local differences: +0.79, +0.37, and +0.50 in Table 4, respectively). This tendency was also seen in [51] for *dangerous* and *dirty*. When looking at the individual median scores per image, we found that for locals, 73% of the images were rated to be more *dangerous*, 61% were rated to be *dirtier* and 75% were rated to be more *polluted*. For these three labels, one can see in Figure 4 that they have a very similar distribution for locals along the Likert scale.

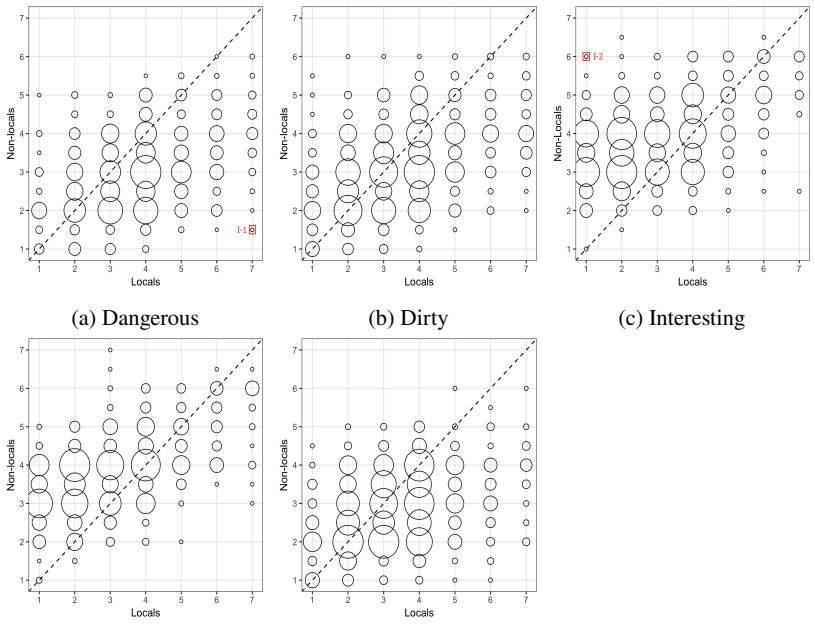
2. Images were perceived as more *interesting* and *pleasant* by the non-locals (local-to-non-local differences: -0.85, and -0.74 in Table 4, respectively). A similar result was also obtained in [51] for the *interesting* attribute. When looking at the individual median scores per image (Figure 4), we found that for non-locals 64% of the pictures were rated as more *interesting*, and 58% were rated as more *pleasant*.

3. Finally, we found that the range of perceptions for the label *pretty* is not statistically different between local and non-local people. (local-to-non-local difference: -0.07, in Table 4). A similar result was also obtained in [51]

Overall, it is relevant that these results match several of our previous findings [51], while using 12 times more data.

6.3.2 Scatter plot analysis

The previous analysis is complemented by a scatter plot analysis. The scatter plots of the 5 statistically significantly different labels (based on the Tukey's HSD test) are shown in Figure 5. Based on them, we confirm that for *interesting* and *pleasant*, many points are above the 45° line, which means that non-locals had a tendency to give higher scores to many images for these labels. In contrast, for *dangerous* and *polluted*, many points are below the 45° line, meaning that locals had a tendency to give higher scores to many images for these labels. In the next section, we will discuss two pictures with opposite scores for the two groups for two labels: *dangerous* and *interesting*. These pictures are identified with the tags *I-1* and *I-2* within the scatter plots shown in Figure 5.



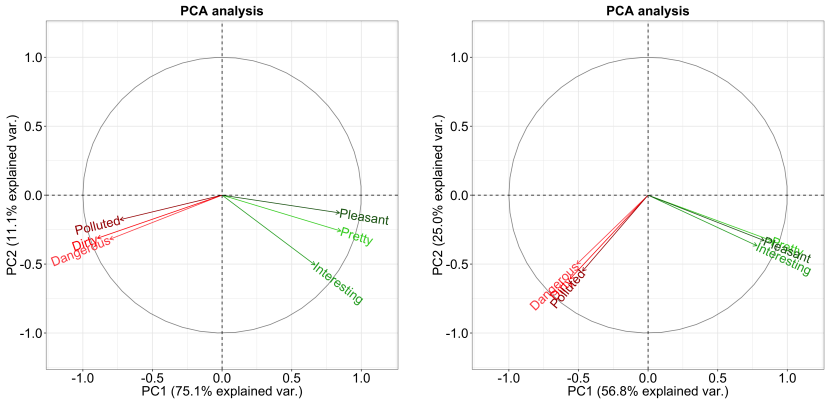
(a) Dangerous (b) Dirty (c) Interesting

(d) Pleasant (e) Polluted

Fig. 5: Scatter plots showing the pair-wise annotator scores by locals and non-locals, for the 5 labels which statistically significant mean difference in Tukey's HSD test. Each circle corresponds to an image, with the size of the circle proportional to the number of observations. A 45° line is also shown in all the plots. Two dots are highlighted in the plots as I-1 and I-2, corresponding to Figure 9.

6.3.3 Correlation analysis

Finally, we perform a PCA analysis on the aggregated median scores for both locals and non-locals. Figure 6 shows the projection of the labels on the two main principal components. We note that the results, while not exactly the same, are quite similar: the first two principal components explain over 80% of the variance, and the loading weights point to the same quadrants. We also see that *dangerous*, *dirty*, and *polluted* point to the negative side of the first principal component, while *pleasant*, *interesting*, and *pretty* point to the positive side. The first principal component seems to correspond to a valence-like dimension in the valence/arousal circumplex model of affect for environments proposed by [48]. If the first component is seen as valence, then the *dirty* and *dangerous* attributes are projected as negative, while *pretty* and *pleasant* are projected as positive (see Figure 6). We corroborate this trend when we compute the correlation between the aggregated scores of both groups (Figure 7): the same subset of labels (positive and negative) are positively correlated within the same subset, but negatively correlated with the opposite subset.



(a) Non-locals (b) Locals

Fig. 6: PCA analysis on the aggregated median scores. The six variables are projected on the first two principal components for (a) non-local impressions; and (b) local impressions. Positive (resp. negative) variables are plotted in shades of green (resp. red).

6.4 Qualitative analysis of impressions

During the data collection process, all non-local MTurk observers were asked to optionally provide comments about the images they labeled, as a means to document

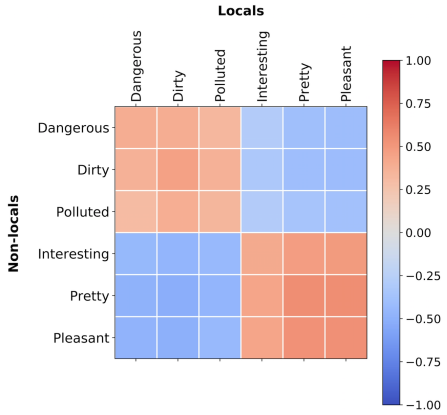


Fig. 7: Correlation matrix of the annotations of non-locals and locals (N=1200, $p<0.05$ for all entries of the matrix).

(a) (b) (c)

Fig. 8: Images that elicited diverging perceptions from local and non-local annotators. (a) The scene in image was perceived as dangerous by locals, less so by non-locals. (b) This scene was perceived as interesting by locals, and differently by non-locals. (c) This scene elicited comments by locals about being interesting, while it elicited other opinions by non-locals, e.g. about perceived safety.

additional impressions of the depicted scenes. A subset of local observers was also asked to provide their comments on a small number of images. The qualitative analysis of these comments complements the statistical analysis of quality and pairwise differences of the annotations presented in Section 6.1. We provide five examples of images (Figures 8 and 9) that evoked different impressions between locals and non-locals.

A first example is shown in Figure 8a. For the local annotators, the depicted site is perceived as dangerous, while for the non-locals it evokes other reactions. Locals seem to use specific visual cues to establish that the urban site depicted in the image is dangerous (stairs poor condition, lack of illumination, the presence of a water spill on the stairs, and exposed pipes.) More specifically, they argued that the stairs and walls are quite deteriorated and the presence of liquids make them slippery and

(a) Dangerous: I-1

(b) Interesting: I-2

Fig. 9: Images with large differences in perception between local and non-local annotators. (a) The scores for *dangerous* are 7 for locals and 1.5 for non-locals. (b) The scores for *interesting* are 1 for locals and 6 for non-locals. Reduced image resolution and pixelation have been used in (a) for privacy reasons.

difficult to use for children and the elderly: *"[Stairs are] dangerous because there is water and the steps are not well built, moreover, the space to pass is narrow"*, *"The steps on the alley look slippery due to [water] spills"* and *"The steps are short, people could run into the water pipes and fall down also due to the poor condition of the steps"*. Some annotators also express concern about vandalism and the lack of light in the space at night. *"[The alley] is dangerous because there are no light sources and very few people walk by"*, *"I think that this place is dangerous because it is a narrow alley and there is no light"* and *"Given the area and location, it looks like people gather to paint graffiti on the walls at night"*. MTurk workers, on the other hand, focused their comments on emotions that do not necessarily reflect a sense of danger: *"It makes me feel sad, oppressed and depressed, like the world is closing in on me in a non-physical sense"* or *"It makes me feel claustrophobic"*. Other comments are more linked to danger: *"Makes me feel as if I have turned down the wrong path into a back alley, or stairwell, that I should not have"* or *"It makes me feel cramped and like I might trip and fall"*.

A second example is shown in Figure 8b. Local annotators perceive the site as interesting, in contrast with non-locals, who focus on visual cues that relate to other attributes. Local annotators say: *"[The place is interesting] because the street is wide and you have a view of the city"*, *"It looks interesting to me because of the structure and the steep ramp"*, *"It is interesting because of the view, the sunlight and it looks peaceful"*, *"This place looks interesting given the distribution of houses and the stone pavement"*, *"I like how it looks like and the trees make it look better"* or *"The place looks well maintained and pretty"*. Non-local annotators expressed a different view: They noted that: *"[the place] makes me feel like I should just keeping moving through if I found myself there"*, *"This area seems slightly isolated so it makes me feel slightly unsafe and weary"* or *"Looks like poor area where there is poverty"*.

A third example is shown in Figure 8c. For some locals, the place is uninteresting: *"[The place] is boring, because it is only one alley with many steps"*, *"Boring, because it is only one alley with many steps"*, *"There are only stairs, I would not like living there, it will not be easy to walk or run"* or *"I like that the alley has many*

steps; by climbing up couple of steps to enter a house and I like the way to reach other houses through the alley". For others, *"the place is interesting because of the long and high stairs"*, and *"because the plants on the alley stand out"*. Non-locals focused on attributes leading to a different view: *"One structure in the background in what appears to be perfect condition. Then, along the stairs a building that does not appear to be in good repair"* or *"Lack of bars and graffiti types of things lead me to believe the criminal element in this neighborhood may be less than in some of the others, so the level of danger might be less"*. One non-local annotator observed that the place is *"Obviously from another time in history."*

As a final example, Figure 9 shows two images that were selected as they had significant score differences across groups, for specific labels. Figure 9a is labeled as *I-1* in Figure 5(a); the scores for the *dangerous* label are 7 for locals and 1.5 for non-locals. This seems to be an outlier case. From our own inspection, it is not clear what kind of features are used by locals to perceived the place as dangerous, except perhaps the sidewalk that seems to blend into the street. We speculate that in a local context, background and prior experiences may play a key role. In contrast, and the basis of the visual cues present in the images, non-locals provide a low score that does not reflect a sense of danger. For Figure 9b, labeled as *I-2* in Figure 5(b), the scores for the *interesting* label are 1 for locals and 6 for non-locals, respectively. In this case, the local score seems reasonable, given the visual cues present in the image. However, for non-locals, the site seems to be interesting. We speculate that annotators might have not previously seen a site like this, which may compel them to find out more about it.

This anecdotal evidence complements the quantitative analysis presented earlier in this section, and seems to confirm that the background and previous experience of observers play a crucial role when forming urban perceptions.

7 Inference from visual cues for local and non-local impressions (RQ2)

The automatic inference methods produce continuous values for each inferred urban perception attribute. For the regression task, we first train Random Forest models using the annotation of each perception label as the dependent variable, and the datasets with visual cues based on CNNs (Subsection 5.1) and the one with semantic descriptors (Subsection 5.2) as the independent variables. We evaluate four models for each group of raters (locals and non-locals): (1) semantic segmentation features; (2) places205 features; (3) places365 features; and (4) semantic descriptors. Finally, we compare the obtained results with our previous work described in [52], which uses a fully connected layer of a GoogLeNet CNN pre-trained on places205 database (CNN-FC) and non-local scores. This model is denoted by M0 in the rest of the section. The eight regression models are summarized in Table 5.

Model 1 (M1-1 & M1-2): This corresponds to the 150 automatically extracted semantic segmentation features. The results are shown in Table 6. The R^2 values

range between 0.10 to 0.30. Furthermore, the results for the systems trained on non-local scores are better than those for the systems trained on local scores, in particular for *dangerous*, *dirty*, and *polluted*. We believe that this is partly due to the lower ICCs obtained for these attributes, as discussed in section 6. Furthermore, the results with this model are below the results obtained with the CNN-FC in [52]. This can be partly explained by the fact that the number of features of the semantic segmentation is lower than the one for CNN-FC.

Model 2 (M2-1 & M2-2): This model corresponds to the automatically extracted places205 features. The results explain reasonably well the positively phrased labels (*pretty*, *interesting* and *pleasant*) as their R^2 values are above 0.25. We see the same pattern as the *model 1*, namely that the results obtained for non-local impressions are better than those for locals. On the other hand, the remaining three variables (*dangerous*, *dirty*, and *polluted*) cannot be inferred when using the local scores.

Model 3 (M3-1 & M3-2): This model corresponds to the automatically extracted places365 features. In this case, the model with the scores generated by non-locals (M3-1) produced better R^2 values when compared to CNN-FC for the *dangerous* and *polluted* labels. As for the remaining labels, results were comparable to those of *model M2-1*. For the model trained on labels by locals, there is not much improvement when compared to *model M2-2*. The regression models trained on *dangerous*, *dirty*, and *polluted* local labels still get low R^2 values.

Model 4 (M4-1 & M4-2): This corresponds to the 10 manually annotated semantic descriptors. The results show that better R^2 is obtained using the annotations by non-locals. We speculate that since the non-local annotators lack contextual information about the cities where the images were taken, they rate mainly based on visual cues they perceive in the pictures as seen in Section 6. Although these results are lower than those obtained with both *M0* and *M3*, they are interesting as only 10 features were used in the regression task. For the case of the local scores, the R^2 of two of the negatively phrased labels (*dirty* and *polluted*) improved considerably. This suggests that when analyzing scenes from a local view, it is possible to achieve better inference performance if high-level visual elements, rather than object-level ones, are considered.

We remark that the results obtained with the local annotator scores are not as good as those produced by the non-locals. This could be explained by the finding that there is more disagreement among local annotators, particularly when scoring the negative labels (Section 6). Notice that this pattern is also visible in the correlation matrices. (Figures 3b and 3a).

In summary, we found that (1) inference systems trained on impression labels provided by non-locals result in higher performance numbers (measured by the standard R^2 coefficient of determination) for all six dimensions in the case of CNN features, and for three dimensions in the case of manual visual cues; (2) positively phrased attributes are inferred with higher R^2 than negatively phrased ones; and (3) for local labels, the three negative variables cannot be recognized at acceptable levels using CNN features, although the performance improved for the case of manual, high-level descriptors.

Label	Group pair	Image corpus: 1,200 images		Image corpus: 99 images [51]	
		Mean difference	p-value	Mean difference	p-value
Dangerous	Lo-NLo	+0.79	**<0.001**	+1.24	**<0.001**
Dirty	Lo-NLo	+0.37	**<0.001**	+1.08	**<0.001**
Interesting	Lo-NLo	-0.85	**<0.001**	-0.59	**0.005**
Pleasant	Lo-NLo	-0.74	**<0.001**	-	-
Polluted	Lo-NLo	+0.50	**<0.001**	-	-
Pretty	Lo-NLo	-0.07	0.21	+0.22	0.24

Table 4: Tukey's HSD statistics. Lo and NLo stands for locals and non-locals. Values in bold are statistically significant with the indicated p-value. Note that the "-" indicates that the label was not included in the study used for comparison [51].

Model	Labels (locals or non-locals)	Visual cues (CNN or manual)
M1-1	Non-local	CNN DilatedNet semantic segmentation
M1-2	Local	CNN DilatedNet semantic segmentation
M2-1	Non-local	CNN GoogLeNet places205
M2-2	Local	CNN GoogLeNet places205
M3-1	Non-local	CNN GoogLeNet places365
M3-2	Local	CNN GoogLeNet places365
M4-1	Non-local	Manual semantic descriptors
M4-2	Local	Manual semantic descriptors

Table 5: Definition of the eight regression models used for inference, according to the choice of labels and visual representation.

Label	M0	M1-1	M2-1	M3-1	M4-1	M1-2	M2-2	M3-2	M4-2
	R^2	R^2	R^2	R^2	R^2	R^2	R^2	R^2	R^2
Dangerous	0.38	0.28	0.35	**0.39**	0.32	0.08	0.13	0.16	0.18
Dirty	**0.37**	0.24	0.35	0.35	0.33	0.07	0.12	0.16	0.24
Interesting	**0.45**	0.28	0.41	0.41	0.19	0.14	0.26	0.24	0.28
Pleasant	**0.45**	0.31	0.42	0.42	0.37	0.18	0.26	0.26	0.32
Polluted	0.30	0.21	0.28	**0.32**	0.21	0.05	0.07	0.08	0.21
Pretty	**0.46**	0.29	0.43	0.42	0.33	0.15	0.27	0.25	0.33

Table 6: Regression results. The values in bold represent the best result for a given label. All models use Random Forests. Models are defined in Table 5. M0 corresponds to the results obtained by using non-local scores and a fully connected layer of a GoogLeNet CNN pre-trained on places205 [52].

8 Discussion

In this section, we discuss some of the main findings of our work, and their implications.

Reliability of local annotations. First, in contrast with our previous findings of [51], which relied on a different cohort of local annotators, we found that the annotations generated by locals are less consistent (i.e., lower ICC values). Since the experimental procedure to gather the annotations was the same, we investigated which factors might have contributed to this lower inter-observer agreement. One possible explanation is that the population of locals is more heterogeneous. In our previous work [51], the majority of participating volunteers (selected from a population of 600 students of CECYTE Guanajuato) lived in the downtown area. In contrast, the cohort participating in this study were from a new campus of the same institution, hosting 1,100 students, a number of whom live in the suburbs. As life in the suburbs has differences with life in the downtown area, it is reasonable to think that the way urban spaces are perceived may differ to some degree. This hypothesis suggests the need to conduct a finer study that examines possible differences in perception across different sub-populations inhabiting a city.

Similarly, we need to consider gender differences (e.g. women vs. men). An example is the perception of danger. On one hand, there is an increasing sense on insecurity in many regions in Mexico [35]; on the other hand, gender violence is a prevalent phenomenon in the country [5]. As a result, women tend to be cautious when walking outside. According to the members of our local research team, city alleys in Guanajuato that lack security elements are seldom used by women walking alone. We speculate that the perception of a dangerous place might differ between local men and women. As an alternative explanation, it is possible that some of the local participants had actually visited the locations they rated, and thus might have first-hand knowledge of the actual level of danger in those locations. These hypotheses would have to be tested as part of future work, since our experimental protocol considered anonymized data, i.e., personal information (age and gender) was only obtained for general statistics but was not linked to each person's annotations.

Comparing local and non-local annotator populations. With respect to demographic comparisons with MTurk workers, recent work showed using a large-scale survey that the large majority of MTurk workers are from the US (75%), followed by India (16%), Canada (1.1%), Great Britain (0.7%), Philippines (0.35%), and Germany (0.27%). As for age, 20% of MTurk workers are born after 1990, 60% are born after 1980, and 80% are born after 1970. As for gender, 51% are women workers and 49% are men [15]. This raises the question of how comparable our two populations of annotators are, and whether such differences might lead to biases. We acknowledge such possibility, yet remark that the unavailability of MTurk in Mexico limits direct comparisons of people working on the same platform. It is also worth considering the possibility that some of the US MTurk workers had Mexican heritage and thus are likely familiar with the urban scenes depicted in our corpus. While we cannot estimate this number accurately as we did not collect such data, we can provide partial evidence based on those MTurk workers who responded to our

demographic survey (N= 77). As discussed in Section 4.3, only 3% of the respondents self-reported their ethnicity as Hispanic/Latino. Furthermore, only 10% of the respondents reported having visited Mexico. Another partial answer can be inferred from [15], which reports that about 0.16% of MTurk workers are from Mexico. A 2018 estimate of the upper bound of the total number of MTurk workers is around 200,000 people, with about 2,450 workers available at a given time [15]. These numbers suggest a low probability that MTurk annotators are of Mexican origin.

Implications of using non-local or local annotations for machine learning. The generation of subjective urban labels exclusively from online crowdsourcing platforms like MTurk has the potential risk of inducing biases. This could in turn lead to machine learning systems that incorporate country- or population-specific perceptions, or that lack diversity. As previously mentioned, recent work found that 75% of MTurk workers are from the US [15]. Furthermore, the majority of people in Mexico do not have access to work on platforms like MTurk due to a variety of factors, including platform restrictions, tax regulations, and lack of access to credit card services. Some of the challenges faced by crowdworkers have been discussed in [26, 25]. For these reasons, conducting future urban crowdsourcing experiments in Global South countries calls for the design of digital platforms adapted to the local conditions, the diversification of urban perception labels that reflect local views, and the use of culture-specific processes to effectively engage workers.

Practical uses of our work. Understanding the urban perception of local inhabitants in Global South cities clearly goes beyond scientific inquiry, which brings up the question on how to provide city officials with tools that leverage upon our findings. We can discuss two directions of the work presented here.

According to the United Nations Human Settlements Programme [4], 85% of the Mexican population will live in cities by 2030. For this reason and as a first direction, it is important to develop methods to understand how these increasing numbers of local inhabitants perceive and experience their environment, as part of participatory processes in collaboration with authorities to develop new public policy [28]. Since the beginning of our research, our team has been in close contact with city authorities in Guanajuato and Leon [47], who were interested in understanding how youth perceive their city and what urban issues they considered more relevant. Over time, our team has shared tools with government officials from Guanajuato City and Leon City, Guanajuato's Youth Institute, the Institute of Legislative Investigations, the Leon's Teenagers House , and the Institute of Planning of Guanajuato State [45]. This shows that the approach is valuable for city offices interested in youth.

In a second direction, contrasting differences of perceptions between locals and visitors has potential practical value for cities. According to the United Nations and the World Trade Organization, Mexico is consistently ranked as one of the ten most visited countries in the world. Understanding how visitors (both real and potential) perceive cities in Mexico, and what urban features are most important for them, could be studied through comprehensive urban perception studies to inform tourism strategies. Our team has also shared insights with government officials from Guanajuato's Tourism Observatory and Guanajuato's Tourism Ministry, with the

goal of developing methods to explore how international tourists perceive urban spaces in touristic cities across Guanajuato state.

9 Conclusions

This chapter presented a study on urban perception by humans and machines, using images as input and six urban perception variables as inference targets. Our work used three cities in Mexico as a case study. We now summarize the answers to the two research questions we addressed.

Our first RQ inquired whether local and non-local observers agreed on the perception of urban dimensions in such cities. We found that non-locals reached higher agreement compared to locals for most dimensions; and that the impression scores of the two groups presented statistical differences for some of the dimensions. More specifically, locals had a tendency to score urban scenes as more dangerous than non-locals; in contrast, non-locals tended to score scenes as more interesting and pleasant than locals. We have discussed possible explanations for these findings. Future work involving a mixed-method approach (collecting additional online observations and interviews with observers) could refine some of the analysis presented here.

Our second RQ asked whether visual machine learning systems would result in comparable performance, when trained to infer subjective attributes of urban scenes with local or non-locals labels. Based on eight models trained on the two types of labels and four types of visual cues (three of them coming from standard CNN systems, and one obtained by manual coding), we found that systems trained with non-local labels produced higher performance. This result could likely follow from the higher inter-observer agreement obtained for the non-local labels. At the same time, this result highlights the importance of understanding the potential impact of systems deployed to recognize perceptual aspects in Global South cities, when these are learned from perceived labels generated by different groups of people, including local inhabitants and external observers.

The results of our work highlight the need for further studies about the influence of demographic and cultural diversity of crowdworkers on subjective label production, and about the implications of this on automation, through empirical analyses of systems that use crowdsourced subjective labels for machine learning in urban environments.

Acknowledgments. L. Medina Rios acknowledges the support of Mexico's Consejo Nacional de Ciencia y Tecnología (CONACYT). S. Ruiz-Correa also acknowledges the support of CONACYT through project 247802. D. Gatica-Perez acknowledges the support of the Swiss National Science Foundation (SNSF) through the Dusk2Dawn Sinergia project. All authors thank the CECYTE Guanajuato community for their enthusiastic participation, and Yassir Benkhedda for technical support. We also thank the book editors for their valuable suggestions to improve the chapter.

References

1. Caminos de la Villa, Argentina. https://www.caminosdelavilla.org. Accessed March 2021
2. List of largest cities. https://en.wikipedia.org/wiki/List_of_largest_cities. Accessed March 2021
3. Map Kibera, Kenya. https://mapkibera.org. Accessed March 2021
4. UN Habitat. https://unhabitat.org. Accessed March 2021
5. UN Women, the long road to justice, prosecuting femicide in Mexico. https://www.unwomen.org/en/news/stories/2017/11/feature-prosecuting-femicide-in-mexico. Accessed March 2021
6. Arietta, S.M., Efros, A.A., Ramamoorthi, R., Agrawala, M.: City forensics: Using visual elements to predict non-visual city attributes. IEEE Transactions on Visualization and Computer Graphics **20**(12), 2624–2633 (2014)
7. Bader, M.D., Mooney, S.J., Lee, Y.J., Sheehan, D., Neckerman, K.M., Rundle, A.G., Teitler, J.O.: Development and deployment of the computer assisted neighborhood visual assessment system (canvas) to measure health-related neighborhood conditions. Health & Place **31**, 163 – 172 (2015)
8. Balestrini, M., Bird, J., Marshall, P., Zaro, A., Rogers, Y.: Understanding sustained community engagement: A case study in heritage preservation in rural argentina. In: Proceedings ACM Conference on Human Factors in Computing Systems, pp. 2675–2684 (2014)
9. Blobaum, A., Marcel, H.: Perceived danger in urban public space: The impacts of physical features and personal factors. Environment and Behavior pp. 465–486 (2005)
10. Candeia, D., Figueiredo, F., Andrade, N., Quercia, D.: Multiple images of the city: Unveiling group-specific urban perceptions through a crowdsourcing game. In: Proceedings ACM Conference on Hypertext and Social Media, HT '17, pp. 135–144 (2017)
11. Connors, W.: Google, Microsoft expose Brazil's favelas. Wall Street Journal (2014)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, p. 886–893. USA (2005)
13. Daniel, T.C.: Whither scenic beauty? visual landscape quality assessment in the 21st century. Landscape and Urban Planning **54**(1), 267 – 281 (2001)
14. DeVries, T., Misra, I., Wang, C., van der Maaten, L.: Does object recognition work for everyone? In: Proceedings of CVPR Workshop on Computer Vision for Global Challenges (2019)
15. Djellel Difallah, E.F., Ipeirotis, P.: Demographics and dynamics of mechanical turk workers. In: Proceedings ACM International Conference on Web Search and Data Mining (2018)
16. Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A.: Deep learning the city : Quantifying urban perception at a global scale. In: Proc. European Conference on Computer Vision, pp. 196–212 (2016)
17. Florida, R., Rentfrow, P.J.: Place and well-being. In: K.M. Sheldon, T.B. Kashdan, M.F. Steger (eds.) Designing Positive Psychology: Taking Stock and Moving Forward. Oxford University Press (2011)
18. Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E.L., Fei-Fei, L.: Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. Proceedings of the National Academy of Sciences **114**(50), 13108–13113 (2017)
19. Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H., Kostakos, V.: Crowdsourcing on the spot: Altruistic use of public displays, feasibility, performance, and behaviours. In: Proceedings ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 753–762 (2013)
20. Goncalves, J., Hosio, S., Van Berkel, N., Ahmed, F., Kostakos, V.: Crowdpickup: Crowdsourcing task pickup in the wild. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(3), 51:1–51:22 (2017)
21. Heimerl, K., Gawalt, B., Chen, K., Parikh, T., Hartmann, B.: Communitysourcing: Engaging local crowds to perform expert work via physical kiosks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12, pp. 1539–1548 (2012)

22. Herzog, T.R., Chernick, K.K.: Tranquility and danger in urban and natural settings. Journal of Environmental Psychology **20**(1), 29 – 39 (2000)
23. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
24. Kaplan, R., Kaplan, S., Brown, T.: Environmental preference: A comparison of four domains of predictors. Environment and Behavior **21**(5), 509–530 (1989)
25. Kingsley, S.C., Gray, M.L., Suri, S.: Accounting for market frictions and power asymmetries in online labor markets. Policy & Internet **7**(4), 383–400 (2015)
26. Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., Horton, J.: The future of crowd work. In: Proceedings ACM Conference on Computer Supported Cooperative Work, pp. 1301–1318 (2013)
27. Koo, T.K., Li, M.Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine **15**, 155–163 (2016)
28. Le Dantec, C.A., Asad, M., Misra, A., Watkins, K.E.: Planning with crowdsourced data: Rhetoric and representation in transportation planning. In: Proceedings ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 1717–1727 (2015)
29. Lindal, P.J., Hartig, T.: Architectural variation, building height, and the restorative quality of urban residential streetscapes. Journal of Environmental Psychology **33**, 26 – 36 (2013)
30. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
31. Ma, X., Hancock, J.T., Lim Mingjie, K., Naaman, M.: Self-disclosure and perceived trustworthiness of airbnb host profiles. In: Proc. ACM Conference on Computer Supported Cooperative Work and Social Computing (2017)
32. Ma, X., Neeraj, T., Naaman, M.: A computational approach to perceived trustworthiness of airbnb host profiles. In: Proc. AAAI Int. Conference on Web and Social Media (2017)
33. Marshall, P., Cain, R., Payne, S.: Situated crowdsourcing: a pragmatic approach to encouraging participation in healthcare design. In: 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), pp. 555–558 (2011)
34. Matzen, K., Bala, K., Snavely, N.: StreetStyle: Exploring world-wide clothing styles from millions of photos. arXiv preprint arXiv:1706.01869 (2017)
35. Monroy-Hernández, A., boyd, d., Kiciman, E., De Choudhury, M., Counts, S.: The new war correspondents: The rise of civic media curation in urban warfare. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, CSCW '13, pp. 1443–1452 (2013)
36. Naik, N., Philipoom, J., Raskar, R., Hidalgo, C.: Streetscore – predicting the perceived safety of one million streetscapes pp. 793–799 (2014)
37. Offenhuber, D., Lee, D.: Putting the informal on the map: Tools for participatory waste management. In: Proceedings of the 12th Participatory Design Conference: Exploratory Papers, Workshop Descriptions, Industry Cases - Volume 2, PDC '12, pp. 13–16 (2012)
38. Ordonez, V., Berg, T.L.: Learning high-level judgments of urban perception. In: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (eds.) ECCV 2014, pp. 494–510. Springer International Publishing (2014)
39. Painter, K.: The influence of street lighting improvements on crime, fear and pedestrian street use, after dark. Landscape and Urban Planning **35**(2), 193 – 201 (1996)
40. Perkins, D., Meeks, J., Ralph, T.: The physical environment of street blocks and resident preceptions of crime and disorder: Implications for theory and measurement. Journal of Environmental Psychology **12**, 21–34 (1992)
41. Piasco, N., Sidibea, D., Demonceaux, C., Gouet-Brunet, V.: A survey on visual-based localization: On the benefit of heterogeneous data **74**, 74–109 (2018)
42. Porzi, L., Rota Bulò, S., Lepri, B., Ricci, E.: Predicting and understanding urban perception with convolutional neural networks. In: Proceedings of the 23rd ACM International Conference on Multimedia, MM '15, pp. 139–148 (2015)
43. Quercia, D., O'hare, N., Cramer, H.: Aesthetic capital: What makes london look beautiful, quiet, and happy? (2014)

44. Rentfrow, P.J.: The open city. In: Handbook of Creative Cities. Cheltenham, UK (2011)
45. Ruiz-Correa, S., Hernandez-Huerfano, E., Alvarez-Rivera, L., Islas-Lopez, V., Ramirez-Sanchez, V., Gonzalez-Abundes, M., Hernandez-Castaneda, M., Carrillo-Sanchez, E., Hasimoto-Beltran, R., Plata Ortega, I.: Urbis: A mobile crowdsourcing platform for sustainable social and urban research in Mexico. In: Sustainable Development Research in Mexico (SDR'17). Springer World Sustainability Series (2018)
46. Ruiz-Correa, S., Santani, D., Daniel, G.P.: Young and the city: Crowdsourcing urban awareness in a developing country. In: Proceedings Int. Conf. on Internet of Things in Urban Space (2014)
47. Ruiz-Correa, S., Santani, D., Ramirez-Salazar, B., Ruiz-Correa, I., Rendon-Huerta, F.A., Olmos-Carrillo, C., Sandoval-Mexicano, B.C., Arcos-Garcia, A.H., Hasimoto-Beltran, R., Gatica-Perez, D.: SenseCityVity: Mobile crowdsourcing, urban awareness, and collective action in Mexico. IEEE Pervasive Computing **16**(2), 44–53 (2017)
48. Russell, J., Pratt, G.: A description of the affective quality attributed to environments. Journal of Personality and Social Psychology **38**, 311–322 (1980)
49. Salesses, P., Schechtner, K., Hidalgo, C.A.: The collaborative image of the city: Mapping the inequality of urban perception. PLOS ONE **8**(7), 1–12 (2013)
50. Santani, D., Ruiz-Correa, S., Daniel, G.P.: Looking at cities in Mexico with crowds. Proceedings ACM Symposium on Computing for Development pp. 127–135 (2015)
51. Santani, D., Ruiz-Correa, S., Daniel, G.P.: Insiders and outsiders: Comparing urban impressions between population groups. In: Proc. ACM International Conference on Multimedia Retrieval. ACM (2017)
52. Santani, D., Ruiz-Correa, S., Daniel, G.P.: Looking south: Learning urban perception in developing cities. ACM Transactions on Social Computing **1**(3) (2018)
53. Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., Sculley, D.: No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In: Proceedings of NIPS Workshop on Machine Learning for the Developing World (2017)
54. Stevens, S.S.: On the theory of scales of measurement. Science **103**(2684), 677–680 (1946)
55. Sturgis, S.: Kids in India are sparking urban planning changes by mapping slums, Bloomberg CityLab (2015)
56. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: Proceedings Int. Conf. on Learning Representations (ICLR) (2016)
57. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
58. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 27, pp. 487–495 (2014)