

A Geography of indoors for analyzing global ways of living using computer vision

Martina Mazzarello

mmazz@mit.edu

Massachusetts Institute of Technology <https://orcid.org/0000-0001-8777-3927>

Mikita Klimenka

Massachusetts Institute of Technology

Rohit Sanatani

Massachusetts Institute of Technology

Javad Eshtiyagh

Massachusetts Institute of Technology

Yanhua Yao

Tsinghua University <https://orcid.org/0009-0006-2321-8426>

Paolo Santi

Massachusetts Institute of Technology <https://orcid.org/0000-0002-8942-8702>

Fabio Duarte

Massachusetts Institute of Technology <https://orcid.org/0000-0003-0909-5379>

Richard Florida

University of Toronto

Carlo Ratti

Massachusetts Institute of Technology

Article

Keywords: Imagery, Computer Vision, Indoor, Globalization, Domestic Environment

Posted Date: August 5th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4676415/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

A Geography of indoors for analyzing global ways of living using computer vision

Martina Mazzarello^{1†}, Mikita Klimenka^{1†}, Rohit Sanatani^{1†},
Javad Eshtiyagh¹, Yanhua Yao², Paolo Santi^{1,4}, Fabio Duarte^{1*},
Richard Florida³, Carlo Ratti¹

¹Senseable City Lab, Massachusetts Institute of Technology, 77
Massachusetts Avenue, Cambridge, 02139, MA, United States.

²Academy of Art and Design, Tsinghua University, Haidian District,
Beijing, 100089, China.

³Rotman School of Management, University of Toronto, 105 St George
St, Toronto, M5S 3E6, Canada.

⁴Istituto di Informatica e Telematica del CNR, Pisa, Italy.

*Corresponding author(s). E-mail(s): mmazz@mit.edu; fduarte@mit.edu;

Contributing authors: klimenko@mit.edu; sanatani@mit.edu;

javade@mit.edu; yyh19@tsinghua.org.cn; psanti@mit.edu;

florida@rotman.utoronto.ca; ratti@mit.edu;

[†]These authors contributed equally to this work.

Abstract

Globalization is often seen as having a homogenizing effect, reducing pronounced local cultural differences. The kinds of housing people live in are one of the most pronounced expressions of local culture. Our research utilizes a unique dataset of more than 640,000 Airbnb images and employs a visual AI framework to examine the differences and similarities in people’s living spaces across 80 cities. We show that geographic proximity and socio-economic connectivity—measured by air traffic data—are significantly correlated with the visual features of indoor spaces ($R = 0.19$ to 0.32), with cities closer to each other or having high passenger traffic between them also exhibiting high visual similarity. Our results challenge the prevailing notion that cultural homogenization may be overdrawn. Despite the homogenizing influence of globalization, distinctive local styles persist in domestic living arrangements. This research highlights the resilience of local identities and cultural distinctions in the face of global pressures.

1 Introduction

The question of whether and how global pressures lessen or eliminate local cultural distinctiveness has interested scholars over the past several decades. It is widely believed that globalization has led to a marked reduction in local cultural distinctiveness. For example, Roland Robertson’s concept of “glocalization” discusses the interplay between global and local influences, suggesting that local cultures adapt and coexist with global trends rather than being entirely subsumed [1]. Similarly, George Ritzer’s “McDonaldisation” thesis argues that the principles of fast-food restaurants are coming to dominate more sectors of society globally, leading to a homogenized culture [2]. Moreover, Thomas Friedman’s “The World is Flat” posits that globalization has created a level playing field, reducing the uniqueness of local cultures [3]. Despite these arguments, living spaces—the physical layout, the configuration, and the presence of objects—are critical indicators of cultural identity, arguably more significant than the types of food we eat, how we dress, or the cars we drive. We spend about 90% of our time indoors [4]. Thus, the way we organize our living spaces, including how to arrange furnishings and household objects, or the materials and colors we use in building our homes, play a crucial role in shaping human activities [5, 6]. They are cultural signifiers, revealing insights into the lives and values of those who inhabit these spaces [7, 8, 9]. Historically, these cultural signifiers used to be specific of certain times and geographies: an apartment in Manhattan Upper East Side in the 1960s, an English townhouse in the early 20th century, the minimalist Japanese home modulated on its tatamis.

To test propositions regarding the persistence of local distinctiveness in the face of globalization, we gather and study a unique dataset of more than 640,000 Airbnb images across 80 global cities, ranging from large metropolitan areas to small urban settings. Our research employs a visual AI framework to examine the differences and similarities in living spaces globally. We aim to underscore the enduring presence of local features within indoor environments, highlighting that, despite the onslaught of global trends, distinct local visual elements and features prominently persist. Globalization has changed the dynamics of cultural expression.

For decades now, scholars [10, 1] have discussed the dialectics between global forces and local practices, with mutual influences. Recent work has been taking advantage of cutting-edge computer vision techniques to analyze the plethora of visual data produced in cities worldwide, shedding light on multiple aspects of urban dynamics. It includes finding hidden styles in city dwellings by analyzing geolocated Airbnb rental images in ten major cities [11], unveiling the potential of image recognition techniques for visual information mining in hospitality and tourism [12, 13], or connecting cities geographically distant through the visual similarities of pictures posted on social media [14]. However, the exploration of indoor images, particularly with extensive scale and

depth, remains largely unexplored, arguably due to the lack of data accessible to researchers in such intimate spaces.

The omnipresence of social media and global markets is reshaping and often homogenizing our cultural references, from music to food to home furniture. This is also true for the hospitality industry, with global hotel chains dominating and setting the standards in the industry. Even personal transportation, be it Uber or Didi or 99taxi, they are basically the same—after all, they already rely on the same technology: cars. However, in 2008, another player entered the hospitality industry, promising to transform global travel behavior [15]: Airbnb, the online platform on which homeowners put their houses for short-term rental. Different from the global homogenization of hotel chains, Airbnb offers more than a shelter: it is a chance to share the domestic environment of an unknown person. From a scientific viewpoint, it offers a unique opportunity to investigate how living spaces are designed, utilized, and perceived across different cultures.

Having become a global platform, despite offering real homes for short-term rental, in several contexts, the platform has also attracted a large number of investors preparing properties solely as Airbnb rentals which diminishes the characteristic of being a "real" living space. Another challenge is that these properties are tailored for a "global" taste—whatever this "global" is, it implies similar house configuration, furniture, and decoration would be found anywhere, ready for global tourists [16], [17].

However, these challenges can also be seen as research opportunities. First, Airbnb's global footprint provides a large source of visual data on domestic spaces. Second, it is the perfect case to investigate cultural homogeneity and heterogeneity globally by examining indoor living spaces. This study employs a visual AI methodology to dissect and understand indoor living spaces through design features, appearance, and the prevalence of objects across different regions. We investigate the extent to which indoor spaces retain their local cultural identities amidst the homogenizing forces of globalization. Our analysis encompasses both external factors, such as geographic proximity and the volume of flights (indicative of socio-economic connectivity), and in-house factors, including the diversity in room aesthetics, the assortment of items within these spaces, and the presence of distinctive objects that contribute to visual classification.

The central inquiry of our research is articulated as follows: Do Airbnb listings exhibit a global homogeneity, or do they retain distinctive local characteristics? Should patterns of similarity emerge, what factors predominantly contribute to these patterns? Are they influenced more significantly by geographic proximity, the strength of social connections, or the typology of the city (whether it is a large metropolis or a smaller urban center)?

2 Results

Our investigation utilized a dual-framework analysis to decipher the visual diversity observed in Airbnb listings globally. This analysis differentiated between external and in-house determinants shaping the interiors' aesthetic uniqueness. Leveraging deep learning for image analysis and object recognition algorithms allowed us to evaluate

640,000 images across a strategic selection of 80 cities. For each of the eight geographic regions of North America, South America, Europe, Africa, West Asia, South Asia, East Asia, and Oceania, by selecting the five most populous cities in which Airbnb is active. Then for each of these large cities, we select a local city nearby with a relatively small population that has a sufficiently large number of Airbnb listings for training and testing purposes.



Fig. 1: The map displays 80 cities. Large cities are represented by large circles, while small cities by small ones. North America: marked in Blue. Large cities are Mexico City, New York City, Los Angeles, and Toronto and small cities are Toluca, Albany, Bakersfield, Quebec, and Madison. South America: Marked in Green. Large cities are Sao Paulo, Buenos Aires, Rio de Janeiro, Lima, and Bogota and small cities are Piracicaba, Cordoba, Vitoria, Trujillo, and Ibague. Europe: Marked in Red. Large circles are London, Berlin, Madrid, Milan, and Paris, and small circles are Bath, Heidelberg, Malaga, Siena, and Bordeaux. Africa: Marked in Plum. Large cities are Kinshasa, Cairo, Abidjan, Nairobi, and Cape Town, and small cities are Gaborone, Luxor, Yamoussoukro, Nakuru, and Paarl. West Asia: Marked in Brown. Large cities are Istanbul, Riyadh, Ankara, Dubai, Lahore and small cities are Eskisehir, Manama, Yerevan, Sharjah, and Islamabad. South Asia: Marked in Olive Green. Large circles are Delhi, Mumbai, Dhaka, Chennai, and Kolkata, and small cities are Amritsar, Varanasi, Sylhet, Lalitpur, and Kandy. East Asia: Marked in Orange. Large cities are Tokyo, Seoul, Osaka, Hong Kong, and Bangkok, and small cities are Miyakami, Suncheon, Okayama, Hualien, and Pattaya. Oceania: marked in Violet. Large cities are Sydney, Melbourne, Auckland, Jakarta, and Adelaide, and small cities are Sunshine Coast, Canberra, Wellington, Denpasar, and Margaret River.

The main finding of our study is that there is still a pronounced retention of local cultural attributes in living spaces, defying the common belief in globalization’s homogenizing influence. Firstly, we establish that geographic proximity is a critical factor. We find geographically closer cities tend to exhibit more visual similarities, with a significant correlation ($R = 0.23$) between geographic distance and visual difference.

This phenomenon is not confined to local and relatively isolated cities but extends to cosmopolitan centers, which also exhibit correlations ($R = 0.26$) between distance and visual difference, and thus indicate similar retention of local attributes.

In addition to geographic proximity, we explored the role of socio-economic connectivity—measured through the volume of direct flights between cities (238 large city pairs)—as a determinant of visual characteristics in living spaces. Our findings suggest a marked impact of socio-economic connectivity, with negative correlations ($R = -0.18$) between airline traffic and visual difference. However, the correlation between visual similarities and geographic proximity emerged as stronger, suggesting that physical distance exerts a more decisive influence on the preservation of local aesthetic distinctions in the face of global standardization.

We will present, moreover, how our analysis delves into city-specific object taxonomies, revealing global differences and identifying distinct indoor elements that define the most visually divergent cities. These patterns of heterogeneity, present both in isolated locales and global metropolises, reinforce the importance of physical location in maintaining the unique visual identity of living spaces amidst widespread standardization. The study identifies elements that contribute to the indoor character of cities, either simply due to their presence, or due to their visual quality, thus revealing the different ways in which such objects shape visual identity.

The *Methods* section provides an in-depth motivation for the external and in-house factors and how we calculated them, along with our approach towards the identification of important indoor elements.

2.1 Local Persistence in the Face of Globalization: Image Analysis Results

The embedding of images in latent space enables us to explore how images of interiors from different cities share stylistic or architectural similarities, even when the model might not accurately identify the city from which each image originates. Specifically, Figure 2 illustrates the latent space representations of bathrooms and bedrooms from a dataset encompassing 40 cities, with each city represented by a unique color and cities within the same region sharing similar color tones (shown in Figure 1). These plots reveal clear regional patterns, demonstrating that the neural network can recognize and group images by regional traits, despite having no prior information about the geographical proximity of the cities, neither during its testing phase nor in its training dataset. This finding underscores that the visual grouping of cities reflects genuine regional design identities, and that unique local interior characteristics persist, notwithstanding the broader trend towards global homogenization of interior design styles.

2.2 The influence of location and socio-economic connectivity on the visual profile of interiors

The analysis of the latent space representation indicates that physical distance significantly influences the visual similarity between living spaces. This phenomenon presents one of the first empirical demonstrations of the “first law of geography” in the

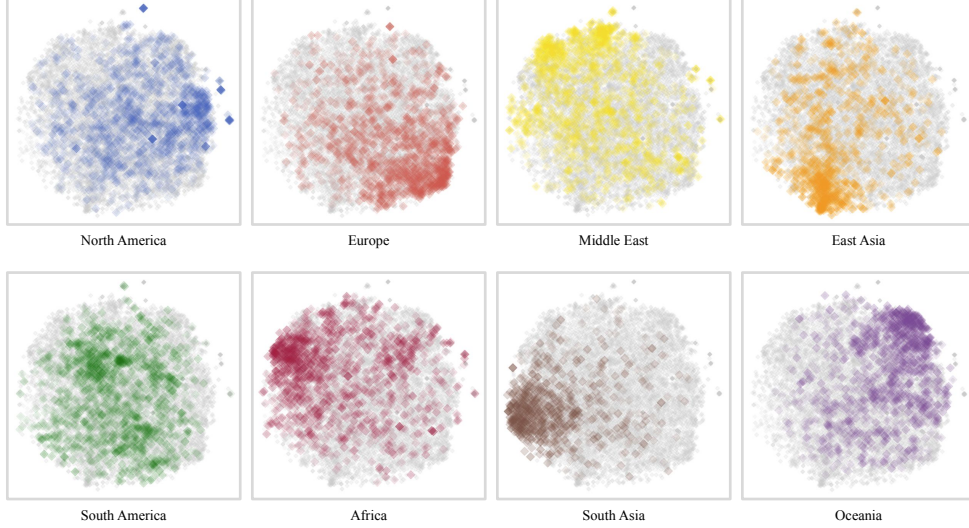


Fig. 2: Latent space (t-SNE) representations of visual identities of kitchens display clear regional affinities. For example, kitchens in South Asia tend to concentrate all in the left part of the latent space whereas kitchens in South America are less concentrated over the whole space. We also observe gradual regional differences. For example, South Asia is closer to East Asia than North America. Similar patterns are observed in the case of bathrooms, living rooms, and bedrooms.

context of domestic interior spaces’ visual qualities—everything is related to everything else, but near things are more related than distant things [18]. Furthermore, the connectivity between cities also plays a crucial role in shaping the visual similarities and distinctions among these spaces across various urban environments. To further examine these influences, we will present an analysis that highlights how these factors differently affect room types and the visual profiles of interiors in both large metropolitan areas and smaller cities.

2.2.1 Locational analysis - visual similarity and geographic proximity

In our study, we analyze the relationship between the physical location of living spaces and their visual characteristics by examining two specific types of distances. The first, geographic great-circle distance, refers to the shortest distance between any two points on the surface of the Earth, measured along the surface of the sphere [19]. This is how we quantify the physical separation between city centers. The second, Euclidean feature-space distance, involves measuring the straight-line distance between points in a multi-dimensional space [20] where each dimension represents a specific visual feature of the interior spaces. This method helps us to quantify the visual similarity or difference between these spaces. Through comprehensive analysis comparing every possible pair of city centroids, our findings reveal consistent and statistically significant

correlations between these two measures of distance. While the strength of this relationship, as indicated by correlation coefficients, varies among different room types, we observe correlation (R) values as high as 0.30 as shown in Figure 3. This provides strong evidence that the geographic location significantly influences the visual appearance of these spaces, suggesting that spaces closer together geographically tend to share more visual similarities than those farther apart.

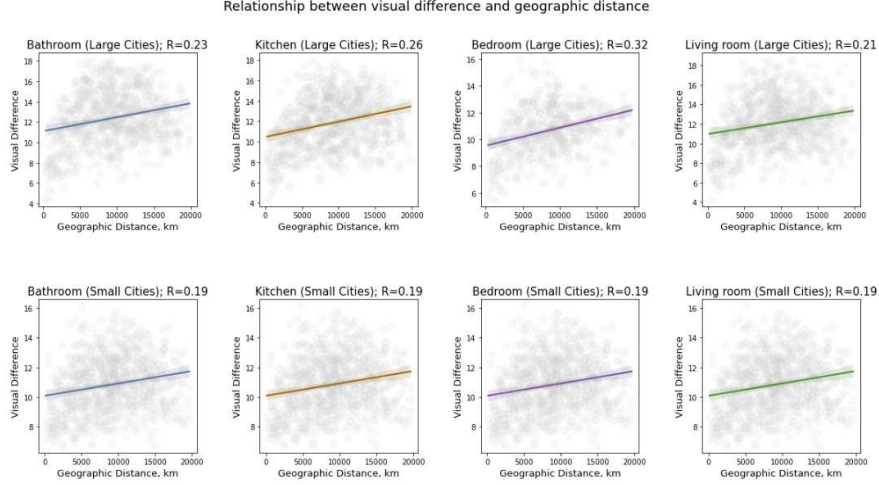


Fig. 3: Relationships between geographic distance and visual similarity for every large city (top) and small city (bottom) pair.

2.2.2 Socio-economic connectivity - visual similarity and social connectivity

While physical proximity has been shown to influence the visual similarity of living spaces, the correlation factor is moderate, hinting at the existence of additional factors that might influence visual similarity. We use the analysis of air travel—measured by the frequency and volume of direct flights between urban centers—as a proxy for the strength and depth of their socio-economic ties, used in this paper with the term socio-economic connectivity. This analysis focuses on a select subset of cities for which direct flight data is available. Our findings show a consistent and statistically significant negative correlation between the volume of direct air traffic for city pairs and the feature-space distances observed between them (Figure 4). In this context, a negative correlation indicates that as the number of direct flights between two cities increases, the visual dissimilarity (as quantified by feature-space distance) between their interior spaces decreases. Consequently, cities with higher degrees of connectivity tend to have more visually similar interior spaces compared to those with fewer direct connections.

Feature-space distance, introduced earlier in our study, measures the visual dissimilarity between spaces by quantifying the distance between their representations in

a multi-dimensional space, where each dimension corresponds to a visual characteristic of the interiors. This metric allows us to objectively compare the visual profiles of living spaces across different cities.

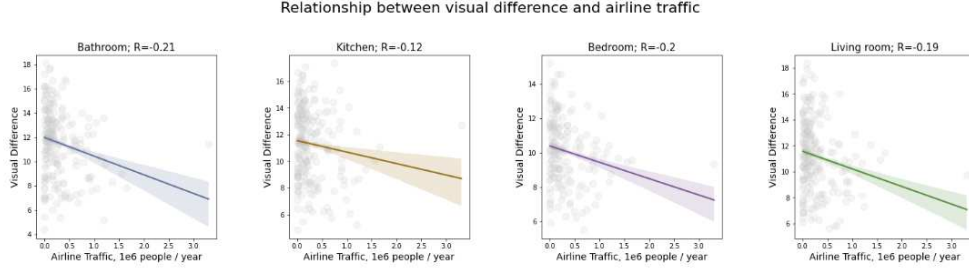


Fig. 4: Relationships between direct passenger traffic and visual similarity for large city pairs.

While such correlations do not necessarily imply any direct causal effects of physical connectivity, they do however empirically reveal the relationships between connectivity and indoor visual quality within an increasingly globalized world.

2.2.3 The Persistence of the Local

While the findings discussed above reveal important macro-trends pertaining to indoor visual qualities across the world, an added focus of this work was also to uncover qualitative and explanatory aspects behind such quantitative findings. The analysis of Gradient-weighted Class Activation Mapping (GradCAM) [21] revealed insights into indoor objects which helped our supervised model identify cities from their images. These become anchors of visual identity, and allow us to map out unique taxonomies for different cities.

Figure 5 summarizes the relative importance of distinct objects within living rooms of cities. Essentially, we identify two primary types of elements that contribute to a room’s visual identity across different room types. In the first type, it is the presence of that element itself that makes the room uniquely identifiable as belonging to a specific city. Examples of these include flowers in bathrooms in Chicago, pictures in living rooms in Auckland, cushions in Living rooms and bedrooms in Auckland and Albany, or foodstuff (identified as creature/fauna) in kitchens in Yammoskuro. In cases like Los Angeles, these can also include outdoor vegetation like palm trees seen through the windows of interior spaces. The second type of elements is characterized not just by their presence but by their specific visual attributes, which include architectural features like walls, windows, doors, and flooring. What makes the interiors of cities, it is not the presence of walls in bedrooms in Chennai or the presence of windows in Paris that make their spaces identifiable, but their unique appearance or “look.” Figure 5 shows selected examples, highlighting visual differences between common elements across cities.

differences in interior spaces. It’s important to note that the visual differences we identify through our neural network approach involve both the appearance of objects and their presence within a scene. To isolate the purely visual difference —disentangling it from the presence of specific objects —we calculated a residual variable between the visual similarity and object difference. This variable captures the variance in visual appearance not explained by the mere presence of an object. To evaluate the significance of geographic proximity versus socio-economic connectivity in influencing these aspects of indoor design, we employed the Akaike Information Criterion (AIC)[22]. The AIC helps in quantifying the strength of a statistical relationship, and the difference between AIC values is used to confirm whether one relation is significantly stronger than another one. Our analysis, based on AIC comparisons, reveals that geographic proximity has a significantly stronger correlation with visual differences across interiors. Conversely, socio-economic connectivity shows a significantly more pronounced relationship with the diversity of objects present. Figure 6 illustrates the relationships between object difference, visual difference, residual (pure visual) difference and geographic and socioeconomic differences.

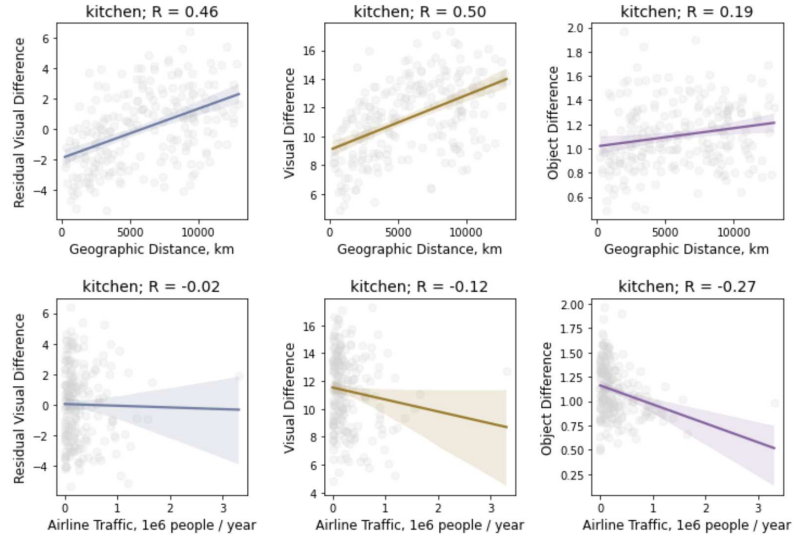


Fig. 6: Relationships between object difference, visual difference, residual (pure visual) difference and geographic and socioeconomic differences. For socioeconomic distance, the slope gets steeper transitioning from pure visual (no correlation) to visual (some correlation) to object composition (medium correlation). For geographic distance, this transition is the opposite.

3 Discussion

In contrast to a large body of prior studies which find that globalization erodes or even annihilates local distinctiveness, the results of our study on living spaces provide evidence of significant persistence of local differences despite globalization. We do this by exploiting a database of more than 640,000 Airbnb images across 80 cities across the world. We use a visual AI approach to underscore the enduring presence of local features within indoor environments, highlighting that, despite the onslaught of global trends, distinct local visual elements and features prominently persist. Our models adeptly classify regions based on visual data alone with an accuracy range of 0.43 - 0.47, reinforcing the concept that indoor elements serve as potent visual indicators of geographic proximity.

Further, our examination reveals a correlation, even though weaker, between air traffic data and the visual similarity of interiors, suggesting that socio-economic connectivity—exemplified by high passenger traffic—contributes significantly to the aesthetic convergence of domestic spaces. This illustrates how the circulation of people and ideas via air travel influences the visual quality of living environments. The vast dataset at our disposal opens up possibilities for more granular investigations into how specific cultural imprints shape domestic spaces. The proliferation of visual data from home streaming, social media, and short-term rentals presents a fertile ground for exploring cultural patterns and human behavior.

However, this study is not without limitations, which in turn pave the way for future research directions. One area of focus is the challenge in detecting less distinct personal items through visual AI, such as small decorative objects, personal artifacts, or unique cultural items, which may evade current detection capabilities. Moreover, distinguishing between the visual composition and the appearance of objects—where ‘composition’ refers to the arrangement and prevalence of elements within a space, and ‘appearance’ pertains to the aesthetic and stylistic attributes of these elements—poses its own set of challenges. This distinction is crucial for understanding how the visual identity of spaces is crafted not just by the types of objects present, but by their stylistic presentation and arrangement. Furthermore, while our classifier was trained on room type without explicit geographic labels, local knowledge may be implicitly captured in the visual data. This subtlety can complicate the recognition of specific features, such as the traditional layout of a Japanese tatami room, which may not be directly labeled but is inherently distinguishable through its unique visual and compositional characteristics. These observations highlight the interplay between global and local factors in shaping indoor aesthetics and point toward the necessity for enhanced algorithms for object detection and better interpretation of latent image features. Future research should aim to refine these techniques and consider incorporating additional data sources, such as cultural artifacts and architectural styles, to deepen our understanding of these dynamics.

In sum, the implications of our findings extend across various fields, offering a novel computer vision framework for examining indoor living spaces. This tool holds promise for anthropologists, sociologists, and cultural researchers, providing a new lens through which to examine the nexus of individual identity, cultural expression, and global influences.

4 Methods

4.1 External factors

The external factors include visual difference, represented by the physical distance between cities, and socio-economic connectivity, represented by the volume of air traffic between cities. By training a neural network on images from the dataset across the world, we obtained a model that captured essential visual distinctions of cities and was capable of identifying a city type from a given image. Instead of using the trained network in a classification task, we instead study how the network describes images visually relative to each other. To do so, we ran model prediction inference on each test image and extracted a penultimate layer from the neural network, which corresponded to the placement of a certain image in a space of latent features. By using dimensionality reduction, we were able to encode the visual representation of each image in a 2-D plot and observe the visual identity of images across cities relative to each other, similar to [23] (Figure 7).

4.1.1 Visual Difference

A robust comparison of diverse images requires developing a common set of parameters that the visual data points can be evaluated upon. Such parameters can include distribution of colors and textures [11], count of specific objects [24], visual clutter [25], or any other visual or indoor layout conditions [26] that can help us draw meaningful conclusions about the difference between images. However, the introduction of highly specific comparison parameters such as color or object count imposes bias on the study as it limits the comparison to a specific parameter. For example, comparing apartments in different continents by wall color is one of the many experiments we may design which asserts the assumption of walls as a defining cultural feature of indoor spaces. We instead seek to first observe universal differences across indoor spaces and further explore the factors driving that difference. In this case, deep learning models are suitable for encoding such comparison frameworks as they are capable of learning highly complex relations between visual data points across billions of dimensions [27].

While deep convolutional neural networks are most often used for specific end-to-end tasks such as classification or object detection, they are also capable of acting as dimensionality reduction mechanisms and are capable of representing a complex RGB input as a datapoint in an n-dimensional latent space that the network learns over training [28]. While we cannot extract the meaning of the latent space dimensions, such a reduction technique allows us to track how visually similar or dissimilar particular two images are, given how far they are positioned from each other in the latent space representation [29]. Neural networks pre-trained on the ImageNet dataset [30] are often used as out-of-the-box visual embedding frameworks. However, since ImageNet is trained to distinguish across a wide array of objects, its latent space embeddings are significantly geared towards object types instead of general appearance. As a result, ImageNet-pretrained models did not allow us to analyze beyond the type of commonplace objects present in the scenes.

In order to learn the latent features relevant exclusively to our dataset, we trained our own ResNet-50 network [31] to distinguish between 40 image classes, each representing a separate city in the dataset. Instead of using the output classification heads of the neural network, we instead extract its penultimate layer (see Figure 7) to obtain the latent space embedding associated with a specific input image [32], which represents the position of that visual data point in a latent space. These latent space vectors contain 1024 dimensions as designed by the architecture of ResNet-50. We further reduced the embedding from 1024 to 2 dimensions using t-SNE, a common dimensionality reduction algorithm [33]. We conducted experiments with reducing dimensions to 3, 512 and with no t-SNE reduction and did not find a significant difference in our results. For convenience of visualization, we proceeded with the reduction to 2 dimensions.

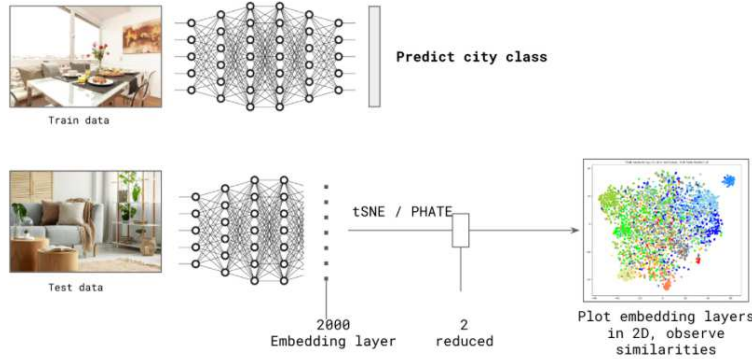


Fig. 7: The unsupervised clustering approach

We conducted training and dimensionality reduction separately for each of the room types (Figure 8). The coloring of the cities by distance shows that while the network was not aware of the regional clustering of input classes, similar cities tend to cluster closer together.

4.1.2 Object Difference

While a dimensionality reduction based on training neural network classifiers from image input allows us to record the degree of difference between individual cities, such ‘black-box’ models do not explain what the contributing factors to those differences are. Using a segmentation network pre-trained on ADE20K dataset [34, 35], we are able to parse each image into semantic regions corresponding to each of the 150 classes in the ADE20K object dataset. If there are some pixels detected belonging to a specific object class, we add their count per given image. We subsequently sum the occurrences of all objects across images per room type and aggregate the sums in a dataframe for each city. We subsequently normalize the values per row to account for potential differences in data points between cities. As a result, we obtain a table that displays the strength of occurrence of each specific object in a given city, which allows us to



Fig. 8: Dimensionality reduction output from a pre-trained network, dots colored by city

obtain 150-number embeddings of object occurrences for each city (See Figure 9a). We may subsequently find differences between the object occurrence vectors by taking the absolute value of their difference.

4.1.3 GradCAM

GradCAM [21] is an insightful approach to investigate how a neural network makes its classification decision. Equipped with this attention visualization tool, we can trace what specific regions in the picture the network focuses on. Using a segmentation network pre-trained on the ADE20K dataset [36], we are able to separate image regions belonging to different types of furniture. By overlaying attention maps over segmentation masks, we were able to calculate the relative degree of attention per each of the 150 object classes segmented from images. By dividing each attention sum by the respective area of the detected object and further normalizing the results, we are able to compare by how much the total attention towards a specific object or a category outweighs its spatial weight per image. This allows us to arrange object categories by their relative importance in a weight distribution (See Figure 9b). By running such analyses across all images belonging to a specific city, we are able to see what specific objects categories are the most characteristic for a particular city in the dataset. Figure 5 displays an array of such comparisons stacked and arranged by objects of average most importance. The gradient from red to yellow visualizes the ‘decay’ of relative importance of specific object categories to the network. Upon general observation, the gradient trends are similar across the cities, with a few common objects achieving high importance. While the same objects generally earn most of the attention, the specific ranking of the objects is also disrupted due to local differences. For example, curtains in Turkey or kitchen furniture in Boston ranks significantly higher than its relative average rank, signifying that this object category is of specific importance for

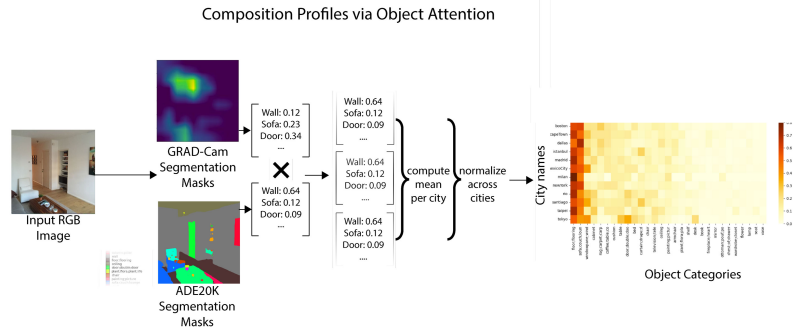
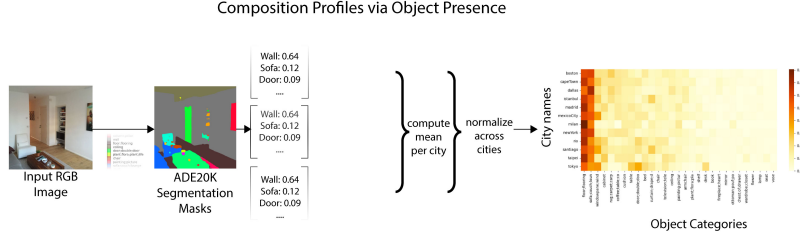


Fig. 9: Two workflows for determining object weight distribution for city classification

the network as it assigns a city class to a given image. This allows us to quantitatively compare cities to each other by examining the similarity of their object decay curves. While such decay curves can be compared to each other in their unique representation, they may also reflect unique differences in characteristic objects between cities.

4.2 In-house factors

For a systematic inquiry into the influences driving visual similarity and difference amongst cities, two in-house factors are analyzed – geographic proximity, and socio-economic connectedness. These are analyzed for every possible city pair, considering the set of large and small cities separately. For both sets of $n=40$ cities, the analysis is carried out on $n(n-1)/2 = 780$ pairs. For an inquiry into the effect of geographic proximity, the correlation between geographic great-circle distance [19] and Euclidean feature-space distance [20] is examined, based on every pair of city centroids. This provides an indication of the degree to which location influences visual similarity within living spaces. For assessing the impact of socio-economic connectedness, we rely on air traffic volume data between cities, as recorded by the International Civil Aviation Organisation (ICAO) [37]. We use flight counts for direct flights between city

pairs and analyze correlations with feature-space distance as above. It is worth noting that not all cities in our dataset have direct flights between them. Our analysis is thus based only on a subset of large cities (238 city pairs).

Acknowledgments. M.M., M.K., R.S., J.E., F.D., P.S. and C.R. acknowledge all members of MIT Senseable City Lab consortium.

Author contributions

M.M., N.K., R.S., F.D., P.S., J.E., Y.Y., and C.R. contributed to the conceptualization of the study. All authors participated in the discussions and commented the manuscript. The writing was carried out by M.M., N.K., R.S., F.D., P.S., J.E., and Y.Y. Data collection and analysis were performed by M.M., N.K., R.S., J.E., and Y.Y. All authors were involved in the analysis process.

Competing interests

The authors declare no competing interests.

Data and Code Availability

The data and code to replicate this research can be obtained from the authors upon reasonable request.

Appendix A Supplementary Material

A.1 Data structure

The dataset for this study encompasses Airbnb listings from 80 cities worldwide. For each of the eight geographic regions of North America, South America, Europe, Africa, West Asia, South Asia, East Asia, and Oceania, we begin by selecting the five most populous cities in which Airbnb is active. Then for each of these large cities, we select a local city nearby with a relatively small population that has a sufficiently large number of Airbnb listings for training and testing purposes. This sample provides us with a list of cities representing different geographic regions, cultures, and climates. Table A.1 presents a list of selected cities and the number of listings in the sample. The downloaded dataset contains posted images of an average of 33,212 listings for large cities and 16,177 listings for small cities. After processing the images and balancing the dataset, this results in a total of around 400,000 images for the selected four rooms. We then use the described machine learning model to identify pictures of kitchens, living rooms, bedrooms, and bathrooms.

Regions	Large Cities	Small Cities	#Listings (large, small)
North America	Mexico City, New York City, Los Angeles, Toronto, Chicago	Toluca, Albany, Bakersfield, Quebec, Madison	(238692, 127698)
South America	Sao Paulo, Buenos Aires, Rio de Janeiro, Lima, Bogota	Piracicaba, Cordoba, Vitoria, Trujillo, Ibague	(173660, 39776)
Europe	London, Berlin, Madrid, Milan, Paris	Bath, Heidelberg, Malaga, Siena, Bordeaux	(128428, 116561)
Africa	Kinshasa, Cairo, Abidjan, Nairobi, Cape Town	Gaborone, Luxor, Yamoussoukro, Nakuru, Paarl	(165324, 71735)
Middle East	Istanbul, Riyadh, Ankara, Dubai, Lahore	Eskisehir, Manama, Yerevan, Sharjah, Islamabad	(195063, 74133)
South Asia	Delhi, Mumbai, Dhaka, Chennai, Kolkata	Amritsar, Varanasi, Sylhet, Lalitpur, Kandy	(115590, 59045)
East Asia	Tokyo, Seoul, Osaka, Hong Kong, Bangkok	Miyakami, Suncheon, Okayama, Hualien, Pattaya	(152907, 62884)
Oceania	Sydney, Melbourne, Auckland, Jakarta, Adelaide	Sunshine Coast, Canberra, Wellington, Denpasar, Margaret River	(158830, 95275)

Table A1: List of selected cities and number of downloaded images per city.

Socioeconomic connectivity is proxied by the number of flights between the pairs of cities. This information is obtained from the dataset ICAO Data+ M2: Traffic By Flight Stage. The International Civil Aviation Organisation (ICAO) collects air transport statistics data from all of its 193 member states. These include passenger and flight counts for and between airports, fleet and personnel data for carriers, financial data and the like. We use this dataset to query the number of direct flights between every city pair for large cities in our dataset. Since not all large cities have direct flights between them, we use a subset of 238 city pairs for our analysis. The number of flights between cities is used as a metric for capturing the strength of the socioeconomic connectivity between them.

1.2 Residual Analysis for 12.2.4

This section supports the analysis performed in Section 2.4 which aimed to explore the difference in the impact of geography and economic connectivity on visual profile and object composition in interiors. Specifically, we hypothesized that economic connectivity is more connected with the object composition of interiors, while geographic distance is more connected with the visual appearance of interiors.

One major issue hindering this comparison is the inability to properly decouple the appearance of objects from their typology. While the object composition dataset we collected indeed counts the occurrences of objects by type, the ‘visual differences’ between images come from a neural network’s latent space that considers the image as a whole. Therefore, the ‘visual differences’ in fact encompass both the appearance of objects and their composition (occurrence) in images.

In order to decouple the ‘true visual appearance’ factor from the composite one, we calculated a residual of the visual appearance with respect to the object [38]. Considering V as a model describing the visual distinctiveness between given city pairs and O as a model describing the difference in object composition, we obtained $RV = V(O) - \langle V \rangle (O)$ as the value of V that is not described by its relationship with O . We called the model RV residual visual factor as a measure of visual differences that does not include object morphology. In Figure Section 2.4 Figure 2 we visualized the Residual Visual, Visual and Object differences as plotted against geographic distance and socioeconomic connectivity. For the plots of residual visual, visual, and object difference plotted against geographic distance, the correlation gets weaker from residual visual to object difference. At the same time, the plots in the same order against socioeconomic connectivity show an increase in correlation towards object difference.

While we used the residual visual factor to illustrate the difference anecdotally, we exploited difference in Akaike Information Criterion (AIC) to confirm the significance in this difference [39]. AIC is a common method for best model selection widely applied in a variety of domains, from environmental studies [40] to sociology [41]. We computed AIC for socioeconomic connectivity with respect to visual difference and object composition and obtained a difference for each room type. We repeated the same experiment for geographic distance. The objective of computing AIC difference

was to explore whether adding another external variable (visual difference to object composition) could make the model significantly more informative. The result in AIC differences are stored in Table X. Considering that the difference above 2-5 AIC values is considered significant [42], we see that for socioeconomic connectivity, object composition is significantly more informative in 3 out of 4 rooms, and visual appearance is more informative for geographic distance for all 4 room types.

$\delta AIC = AIC(\text{SocioEconomicProximity} \text{VisualDifference}) - AIC(\text{SocioEconomicProximity} \text{ObjectDifference})$	
Room type	δAIC
Bathroom	-3.1
Bedroom	8.2
Living room	18.2
Kitchen	13.4

Table A2: A table of differences between the AIC value for a regression model of Socioeconomic Proximity vs visual difference and the AIC value for a regression model of Socioeconomic Proximity vs object difference, exploring how better the object distance is for predicting socioeconomic distance as opposed to visual difference, with a significant difference being larger than 2. In 3 out of 4 room types, object difference is a significantly better explanation than visual difference

$\delta AIC = AIC(\text{GeographicProximity} \text{VisualDifference}) - AIC(\text{GeographicProximity} \text{ObjectDifference})$	
Room type	δAIC
Bathroom	-36.3
Bedroom	-12.3
Living room	-16.2
Kitchen	-53.1

Table A3: A table of differences between the AIC value for a regression model of Geographic Proximity vs visual difference and the AIC value for a regression model of Geographic Proximity vs object difference, exploring how better the object distance is for predicting socioeconomic distance as opposed to visual difference, with a significant difference being larger than 2. In 4 out of 4 room types, object difference is a significantly worse explanation than visual difference

References

- [1] Roland Robertson et al. “Glocalization: Time-space and homogeneity-heterogeneity”. In: *Global modernities* 2.1 (1995), pp. 25–44.
- [2] George Ritzer. *The McDonaldization of society* 6. Vol. 6. Pine Forge Press, 2011.
- [3] Thomas L Friedman. *The world is flat: A brief history of the twenty-first century*. Macmillan, 2005.

- [4] Neil E Klepeis et al. “The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants”. In: *Journal of exposure science & environmental epidemiology* 11.3 (2001), pp. 231–252.
- [5] Robin Hanson. “Combinatorial information market design”. In: *Information Systems Frontiers* 5 (2003), pp. 107–119.
- [6] Daniel Miller. *Home possessions: Material culture behind closed doors*. Routledge, 2021.
- [7] Mike Featherstone. “Global culture: An introduction”. In: *Theory, Culture & Society* 7.2-3 (1990), pp. 1–14.
- [8] John Tomlinson. “Cultural globalization”. In: *The Blackwell companion to globalization* (2007), pp. 352–366.
- [9] Dorothy E Leidner. “Globalization, culture, and information: Towards global knowledge transparency”. In: *The Journal of Strategic Information Systems* 19.2 (2010), pp. 69–77.
- [10] Anthony Giddens. “The consequences of modernity. 1990”. In: (2007).
- [11] Sohrab Rahimi, Xi Liu, and Clio Andris. “Hidden style in the city: an analysis of geolocated airbnb rental images in ten major cities”. In: *Proceedings of the 2nd ACM SIGSPATIAL workshop on smart cities and urban analytics*. 2016, pp. 1–7.
- [12] Xi Liu et al. “Inside 50,000 living rooms: an assessment of global residential ornamentation using transfer learning”. In: vol. 8. 4.
- [13] Xianwei Liu et al. “Applying image recognition techniques to visual information mining in hospitality and tourism”. In: *International Journal of Contemporary Hospitality Management* 35.6 (2023), pp. 2005–2016.
- [14] Fan Zhang et al. “Discovering place-informative scenes and objects using social media photos”. In: *Royal Society open science* 6.3 (2019), p. 181375.
- [15] Nicole Gurran. “Global home-sharing, local communities and the Airbnb debate: a planning research agenda”. In: *Planning theory & practice* 19.2 (2018), pp. 298–304.
- [16] R Bachar et al. “The De-Domestication of Home-The Story of Airbnb”. In: (2020).
- [17] Marian Makkar and Sheau-Fen Yap. “Managing hearts and minds: romanticizing Airbnb experiences”. In: *Current Issues in Tourism* 25.21 (2022), pp. 3461–3480.
- [18] Waldo R Tobler. “A computer movie simulating urban growth in the Detroit region”. In: *Economic geography* 46.sup1 (1970), pp. 234–240.
- [19] R Bullock. “Great circle distances and bearings between two locations”. In: *MDT, June* 5 (2007), pp. 1–3.
- [20] Kimberly L Elmore and Michael B Richman. “Euclidean distance as a similarity metric for principal component analysis”. In: *Monthly weather review* 129.3 (2001), pp. 540–549.
- [21] Ramprasaath R Selvaraju et al. “Grad-CAM: Why did you say that?” In: *arXiv preprint arXiv:1611.07450* (2016).
- [22] Hirotugu Akaike. “Factor analysis and AIC”. In: *Psychometrika* 52 (1987), pp. 317–332.

- [23] Mathilde Caron et al. “Deep clustering for unsupervised learning of visual features”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 132–149.
- [24] Henk Staats and Ritwik Swain. “Cars, trees, and house prices: Evaluation of the residential environment as a function of numbers of cars and trees in the street”. In: *Urban Forestry & Urban Greening* 47 (2020), p. 126554.
- [25] Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. “Measuring visual clutter”. In: *Journal of vision* 7.2 (2007), pp. 17–17.
- [26] Mohammad Javad Parseh et al. “Semantic-aware visual scene representation”. In: *International Journal of Multimedia Information Retrieval* 11.4 (2022), pp. 619–638.
- [27] Yoshua Bengio, Aaron Courville, and Pascal Vincent. “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.
- [28] Wei Wang et al. “Generalized autoencoder: A neural network framework for dimensionality reduction”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014, pp. 490–497.
- [29] Yang Liu et al. “Latent space cartography: Visual analysis of vector space embeddings”. In: *Computer graphics forum*. Vol. 38. 3. Wiley Online Library. 2019, pp. 67–78.
- [30] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [31] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [32] Matthew D Zeiler and Rob Fergus. “Visualizing and understanding convolutional networks”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I* 13. Springer. 2014, pp. 818–833.
- [33] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).
- [34] Bolei Zhou et al. “Scene parsing through ade20k dataset”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 633–641.
- [35] Bolei Zhou et al. “Semantic Understanding of Scenes through the ADE20K Dataset”. In: *International Journal on Computer Vision* 127 (2019), pp. 302–321.
- [36] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. “Vision transformers for dense prediction”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 12179–12188.
- [37] David Mackenzie. *ICAO: A history of the international civil aviation organization*. University of Toronto Press, 2010.
- [38] JRJ Belloto and TD Sokolovski. “Residual analysis in regression”. In: *American Journal of Pharmaceutical Education* 49.3 (1985), pp. 295–303.

- [39] Eric-Jan Wagenmakers and Simon Farrell. “AIC model selection using Akaike weights”. In: *Psychonomic bulletin & review* 11 (2004), pp. 192–196.
- [40] Anthony T Cahill. “Significance of AIC differences for precipitation intensity distributions”. In: *Advances in water resources* 26.4 (2003), pp. 457–464.
- [41] Raymond Sin-Kwok Wong. “Model selection strategies and the use of association models to detect group differences”. In: *Sociological Methods & Research* 22.4 (1994), pp. 460–491.
- [42] Kenneth P Burnham, David R Anderson, and Kathryn P Huyvaert. “AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons”. In: *Behavioral ecology and sociobiology* 65 (2011), pp. 23–35.