Check Out This Place: Inferring Ambiance From Airbnb Photos

Laurent Son Nguyen, Salvador Ruiz-Correa, *Member, IEEE*, Marianne Schmid Mast, and Daniel Gatica-Perez, *Member, IEEE*

***Abstract*—Airbnb is changing the landscape of the hospitality industry, and to this day, little is known about the inferences that guests make about Airbnb listings. Our work constitutes a first attempt at understanding how potential Airbnb guests form first impressions from images, one of the main modalities featured on the platform. We contribute to the multimedia community by proposing the novel task of automatically predicting human impressions of ambiance from pictures of listings on Airbnb. We collected Airbnb images, focusing on the countries Switzerland and Mexico as case studies, and used crowdsourcing mechanisms to gather annotations on physical and ambiance attributes, finding that agreement among raters was high for most of the attributes. Our cluster analysis showed that both physical and psychological attributes could be grouped into three clusters. We then extracted state-of-the-art features from the images to automatically infer the annotated variables in a regression task. Results show the feasibility of predicting ambiance impressions of homes on Airbnb, with up to 42% of the variance explained by our model, and best results were obtained using activation layers of deep convolutional neural networks trained on the Places dataset, a collection of scene-centric images.**

***Index Terms*—Ambiance prediction, first impressions, home environments, Airbnb, social media, image processing.**

I. Introduction

AMBIANCE, defined as "*the character and atmosphere of a place*" [8], encompasses what people think and feel about a physical environment. The problem of automatically characterizing place ambiance has gained increased interest in multimedia research, with works mainly focusing on outdoor [33], [35] and indoor [38], [43] public spaces. In the context of online peer-accommodation platforms like Airbnb, ambiance plays an important role in the process of deciding on what place to book for a stay. Being able to characterize place ambiance of Airbnb listings could enable applications ranging from ambiance-based search engines (e.g., a *romantic* studio for a couple's week-end trip, or a *practical*, *spacious*, and *comfortable* house for a week-long family vacation) to automated recommendation systems for hosts to improve the presentation of their listing (e.g., interior design, decoration). To our knowledge, no work has addressed the problem of ambiance prediction in a setting related to private indoor places, homes, or peer-accommodation platforms. In this work, we address the new multimedia task of automatically predicting human impressions of ambiance from pictures of the online peer-accommodation platform Airbnb.

Airbnb is changing the landscape of the hospitality industry. As part of the *sharing* or *on-demand economy*, the peer accommodation platform offers guests an inexpensive and credible alternative to the traditional hotel industry, by providing benefits associated with staying in a local residence [17]. For hosts, Airbnb is an opportunity to monetize their extra space for short-term rentals, by advertising the place online through text descriptions, ratings, comments, and photos. Despite the increasing prevalence of peer accommodation platforms, relatively little academic literature has investigated such systems. This work constitutes a first attempt at understanding how potential Airbnb guests form first impressions from images, one of the main modalities featured on the platform.

Closely related to our work, psychology researchers were the first investigating the formation of first impressions in home environments, focusing on the desired ambiance [16], perceived ambiance [14], and the personality of the owner [14] of home environments. As a main limitation, most of the existing research in this field has been conducted using surveys [16] or visiting homes in person [14], which makes it difficult to scale with respect to the number of places and their geographic distribution. Because Airbnb spaces are meant to be rented by guests, their presentation can be polished to look beautiful and catchy; thus, their appearance can differ from real personal living spaces. However, Airbnb listings are reflective of homes *as presented in social media*, and the platform is what approaches the most to home environments online and at scale, thus making Airbnb an opportunity to study indoor private places at a scale never achieved before. In this sense, our work also contributes to the

Manuscript received August 2, 2016; revised March 18, 2017 and September 12, 2017; accepted October 14, 2017. Date of publication November 2, 2017; date of current version May 15, 2018. This work was supported in part by the EPFL-UNIL CROSS program through the Mi Casa es su Casa project, and in part by the Swiss National Science Foundation through the UBImpressed Sinergia project. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xilin Chen. *(Corresponding author: Laurent Son Nguyen.)*

L. S. Nguyen is with the Social Computing Group, Idiap Research Institute, Martigny 1920, Switzerland (e-mail: lnguyen@idiap.ch).

S. Ruiz-Correa is with the Centro Nacional de Supercómputo, Instituto Potosino de Investigacion Cientifica y Tecnologica, San Luis Potosí 78216, Mexico (e-mail: src@cmls.pw).

M. Schmid Mast is with the Faculté des Hautes Etudes Commerciales (HEC Lausanne), Universite de Lausanne, Lausanne CH-1015, Switzerland (e-mail: Marianne.SchmidMast@unil.ch).

D. Gatica-Perez is with the Ecole Polytechnique Federale de Lausanne (EPFL), Idiap Research Institute, Martigny CH-1920, Switzerland (e-mail: gatica@idiap.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2769444

psychology literature by proposing a scalable methodology, as well as by providing insights about the structure of judgments of personal indoor environments.

Our work addresses the following research questions:

RQ1 Can the ambiance and the physical attributes of on-demand home environments listed on Airbnb be consistently assessed by external observers? If so, what are the dimensions with highest agreement/reliability?

RQ2 What is the underlying structure of ambiance and physical attributes? In other words, can ambiance and physical attributes be grouped into clusters of low-dimensional space?

RQ3 What types of Airbnb images best convey the ambiance of a home environment? Can the selection of images to be shown to observers be automated?

RQ4 Can high-level judgments of ambiance and physical attributes be automatically inferred from images?

To address these research questions, we collected images from Airbnb, focusing on two countries where Airbnb is popular (Switzerland and Mexico) as case studies. We then gathered a first round of annotations on 200 Airbnb listings by trusted research assistants on 49 physical and ambiance attributes. To understand the structure of these annotations, we performed a cluster analysis and found that the variables could be grouped into three clusters. We then used Amazon Mechanical Turk (MTurk) to gather annotations on 1200 listings, in which workers had to watch three images and answer questions on physical and ambiance attributes of the place; inter-rater agreement was high for most of the dimensions, showing that MTurkers provide reliable impressions of ambiance from automatically selected images of Airbnb listings. We then extracted state-of-the-art features from the images to automatically infer the high-level variables in a regression task. Results showed the feasibility of inferring ambiance impressions of homes on Airbnb, with up to 42% of the variance explained by our model. In terms of image representations, best results were obtained using activation layers of deep convolutional neural networks trained on the Places dataset, a collection of scene-centric images.

Our paper has the following contributions:

1) We propose a new task in multimedia research, namely the inference of ambiance in on-demand home environments using Airbnb images. To this end, we collected images from Airbnb listings from Mexico and Switzerland. Our dataset contains 350 K images from 22 K+ Airbnb listings in these two countries.
2) To address **RQ1**, research assistants in our team annotated 200 listings on 49 dimensions, including physical, ambiance, and overall attributes, and results showed that 42 out the 49 dimensions had moderate-to-high inter-rater agreement, suggesting that most attributes could be consistently assessed by external raters.
3) To address **RQ2**, we performed a cluster analysis, and results showed that the dimensions could be grouped into three main clusters: *positive*, *negative*, and *decorated/unique*.
4) To further address **RQ1** in an online crowdsourcing setting, we used Amazon Mechanical Turk to collect annotations of 1200 Airbnb listings on 11 physical and ambiance dimensions derived from the cluster analysis. Agreement among MTurkers was moderate-to-high for most attributes, showing that these dimensions could be consistently annotated.
5) To address **RQ3**, we propose an automatic image selection method to represent the Airbnb listings, which did not decrease the validity of the annotations compared to selecting images manually, indicating the feasibility of such an approach.
6) To address **RQ4**, we applied a pipeline to automatically infer annotations of Airbnb places based on a large variety of image features, including color histograms, histogram of gradient (HOG), and features based on pre-trained convolutional neural networks (CNNs). Results showed that most dimensions could be predicted up to $R^2 = 0.42$, and that CNNs trained on the Places dataset overall yielded the best prediction accuracies.

II. Related Work

Given the multi-faceted nature of this work, the related work spans multimedia, social media analysis, and psychology.

A. Related Work in Psychology

Environmental psychologists were among the first to study how humans select, affect, and get influenced by home environments. Spaces in which we live and work tend to shape us: housing quality and form, interior design, and colors were found to be related with social constructs and outcomes as diverse as satisfaction, problems with children, or warmth [13]. Gosling *et al.* have identified three main factors to explain how people affect their personal living spaces: (1) identity claims, i.e., deliberate symbolic statements about how occupants would like to be regarded; (2) thought and feeling regulators, i.e., features or objects able to positively affect the occupant's feeling and thoughts (e.g., family pictures, wall colors); and (3) behavioral residue, i.e., physical traces left in the environment by behavioral acts [13]. This results in physical alterations of the space, and environmental cues can be used by observers to make inferences about the owner's personal characteristics as diverse as gender [39], political preferences [7], and even personality [14]. These works were based on in situ visits of personal spaces, where observers gave ratings and coded physical features of the space; this approach is arguably the most ecologically valid one, but makes it difficult to conduct large-scale studies. Our study contributes to the growing body of related work leveraging social media and crowdsourcing to overcome this limitation [15], [16], [42] (see Section II-B).

Recent studies in environmental psychology [7], [14] were based on the Personal Living Space Cue Inventory (PLSCI) [12], an instrument designed to systematically capture features of personal spaces. PLSCI consists of a questionnaire containing 42 bipolar attributes related to physical properties (e.g., cluttered *vs.* uncluttered) and ambiance (e.g., comfortable *vs.* uncomfortable) of the space, as well as a check-list of 700 individual items; it is to this day the most complete instrument to document personal spaces. However, little is known about the inherent structure of the attributes coded in PLSCI; our work

contributes to this issue by performing a cluster analysis of the annotated skills in the context of Airbnb places.

B. Related Work in Social Media

Social multimedia platforms that allow users to share photos, comments and reviews have gained widespread adoption, and have been used by researchers as data sources. Bakhshi *et al.* [4] analyzed the relationship between the presence of faces in pictures and social engagement on 1M Instagram images, and found that photos containing a face received more *likes* and *comments*. Hays *et al.* [18] constructed a geo-localized dataset of 6M+ images from Flickr to estimate the geographic information from the image content.

Graham *et al.* [15] used Foursquare to investigate the reliability of human impressions of place ambiance based on the profile of Foursquare users who had visited the place, and in situ visits of the place; results showed that observers were able to form consistent impressions in both settings. This study, however, only examined 49 places in one city and involved the personal visit of the places. Santani *et al.* [42], [43] addressed these issues by investigating the use of crowdsourcing mechanisms to understand social ambiance in 300 popular Foursquare indoor places from six cities based on manually-selected user-contributed images, and found that crowdworkers were consistent in judging public spaces from pictures. Our work is closely related to these studies in terms of the social impressions under analysis. Our work further contributes to the existing literature by extending the analysis of ambiance on home environments (as opposed to public spaces) shared on Airbnb, and by automatically inferring ambiance impressions from images.

C. Related Work on Airbnb

The body of research on Airbnb spans the fields of management, tourism, architecture, user experience design, and multimedia. Concepts as diverse as discrimination [10], review bias [48], hospitality exchanges [26], price and neighborhood prediction [46], or socio-economic characteristics of areas [36] have been investigated on Airbnb. Zervas *et al.* [48] examined the rating mechanisms of Airbnb on 12K+ listings and found that over 50% of the properties had 5-star ratings (best possible rating), and 94% had over 4.5-star ratings, highlighting the limitations of Airbnb rating mechanisms. Lampinen *et al.* [26] investigated the motivations of hosts to monetize network hospitality in several qualitative studies, and found that the financial exchange facilitates social exchanges between hosts and guests. Rahimi *et al.* [37] recently investigated the relationships between the level of decoration and geographical differences using Airbnb images from 10 cities, and did not observe major differences across cities (suggesting a globalization effect), whereas noticeable differences were observed across neighborhoods of a same city.

In multimedia analysis, Quattrone *et al.* [36] used a computational approach to determine the socio-economic conditions that benefit from Airbnb, using all London listings spanning a period from 2012 to 2015; results showed that the density of Airbnb listings was correlated with the inverse distance to the city center, and that the Airbnb properties were associated with attractive areas with young, tech-savvy residents. Lee *et al.* [28] quantitatively characterized consumption behaviors in Airbnb, using 4K+ listings; results showed that host responsiveness, wish-list counts, number of reviews, and price were correlated with the number of rentals, while the overall rating of the listing was not, which can be explained by the high skewness of ratings found in [48]. To our knowledge, the only work involving Airbnb images in multimedia analysis is the one by Tang *et al.* [46], where the problem of neighborhood and price prediction using image and textual features was addressed, using San Francisco data obtained from the Inside Airbnb project [2].

Our work contributes to the literature on Airbnb by analyzing images for the automatic inference of first impressions, one of the most relevant modalities of the platform in the context of zero-acquaintance situations. Specifically, we analyze what crowdsourced impressions can be reliably annotated from images, and use a computational framework to predict physical and ambiance attributes.

D. Related Work in Multimedia Analysis

The last decade has witnessed significant progress in the field of automatic scene recognition. Although most of the works have focused on objective image classification tasks like object detection or scene type recognition (e.g., [27], [49]), some studies have addressed the problem of inferring high-level human impressions about images. Such tasks include the prediction of aesthetic qualities [29], [30], [47] and memorability [19] of images, or visual style of objects [20].

More closely related to our work are studies [33], [35] addressing the problem of predicting high-level judgments from the Place Pulse dataset, a corpus of 4K+ images of urban places obtained from Google Street View, including crowdsourced annotations of safety, social class, and uniqueness [41]. Ordonez *et al.* [33] used GIST, SIFT + Fischer Vectors, and deep convolutional features from a pre-trained model to predict safety, uniqueness, and wealth in both classification and regression tasks. Porzi *et al.* [35] formulated the problem of urban perception prediction as a ranking task. They trained a convolutional neural network on the annotated pair-wise comparisons of pictures of urban scenes inherent to the Place Pulse dataset, and compared their approach to state-of-the-art features (GIST, HOG, and CNN-based features). Although our approach is similar in terms of the extracted image features, our work significantly differs from [33], [35] in both the type of places (home environments *vs.* outdoor urban places) and the constructs under analysis (a number of previously unexplored dimensions of ambiance *vs.* safety, uniqueness, wealth).

Related to impressions of indoor environments, Redi *et al.* [38] addressed the problem of automatically inferring place ambiance based on profile pictures of patrons. To this end, they automatically extracted image features from profile pictures, including aesthetics, colors, emotions, demographics (age and gender), and self-presentation. Results showed the feasibility of such an approach, even if the validity of the results is limited by the small number of data points ($N = 49$). Santani *et al.* [43] also addressed the problem of automatic ambiance prediction of popular public spaces on Foursquare. They extracted visual

TABLE I
NUMBER OF COLLECTED AIRBNB LISTINGS AND IMAGES; MEAN (AVGIM) AND MEDIAN (MEDIM) NUMBER OF IMAGES PER LISTING FOR MEXICO AND SWITZERLAND

	#listings	#images	avgIm	medIm
Mexico	14206	246518	17.35	14
Switzerland	8462	103210	12.19	10
Total	22668	349728	15.43	12

features from the pictures of the public places, including color distributions, GIST, HOG, and pre-trained CNN-based features, and results showed the validity of this approach. Our work shares with [38] and [43] some annotated ambiance attributes, but differs in terms of the place type (home environments *vs.* public indoor places).

III. DATA COLLECTION AND SELECTION

We collected Airbnb data from Mexico and Switzerland as case studies. The choice of these countries was motivated by the fact that they both represent distinct parts of the world (developing and developed world, respectively), are varied in terms of tourist/non-touristic regions, and have significant Airbnb penetration. We queried `www.airbnb.com` for both countries; IDs and corresponding latitude/longitude coordinates were returned by the platform. Listings outside the country borders were discarded using latitude/longitude coordinates combined with a point-in-polygon method. Images were then downloaded for each unique listing ID. Data was collected in August 2015. Table I displays basic statistics of the collected data. In total, we collected 350K images from 22K+ individual Airbnb listings.

In order to gain basic understanding of the content of the downloaded images, we used a computational method to automatically label images from 1500 randomly sampled listings (750 for Mexico and 750 for Switzerland), for a total of 21998 images. We used the GoogLeNet convolutional neural network trained on the Places dataset [49], a corpus of 2.4M+ images from 205 scene categories. The use of this specific model is justified by the significant overlap of scene types between our dataset and the one used for training. For each image, we applied a forward pass of the neural network using the Caffe framework [22], and obtained the class probabilities of each 205 place category. Fig. 1 displays the distribution of the most probable place label; we observe that the 6 most frequent classes (bedroom, shower, living room, parlor, and kitchenette) are directly related to indoor home environments, while other frequent classes are related to outdoor environments (swimming pool, patio, veranda, chalet, courtyard), which suggests that Airbnb is mainly composed of pictures of the physical environment. In this respect, Airbnb significantly differs from other social media platforms such as Foursquare, where a significant proportion of the images is related to food, drinks, or people [42]. Furthermore, on the 21998 images analyzed, the median resolution was 1440×1080 pixels. These observations constitute a strong justification to use Airbnb for home environment research: the data is available in large quantities, most images are depicting views of the indoor physical environment, and the resolution is high enough to capture detailed information about the place.

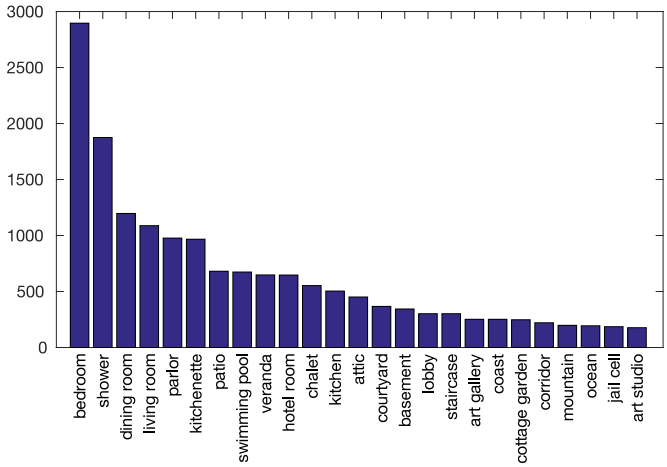

Fig. 1. Distribution of the most probable image label from 1500 randomly sampled Airbnb listings (750 for Mexico, 750 for Switzerland, $N_{images} = 21998$), using GoogLeNet Places CNN.

In this work, we use images from 1200 randomly sampled Airbnb listings (600 for both Mexico and Switzerland). This number constitutes a significant increase with respect to previous work on home environments. While we were limited by the cost of annotating the listings, this number is sufficient to obtain statistically significant and representative results. Furthermore, our methodology is fully scalable, and the price of collecting annotations is directly proportional to the number of desired listings.

IV. IMPRESSIONS OF AIRBNB PLACES

Three of our research questions in Section I relate to the types of impressions that observers make about Airbnb places: (**RQ1**) Can the ambiance and the physical attributes of Airbnb listings be consistently assessed by external observers? If so, what are the dimensions with highest agreement? (**RQ2**) What is the underlying structure of ambiance and physical attributes? (**RQ3**) What types of Airbnb images best convey the ambiance of a home environment? Can the selection of images to be shown to observers be automated? Similarly to [42], we used three images to represent each Airbnb listing to address these questions.

We used the following three-step approach. (1) We tested the "ideal" case on 200 listings, where the images to be shown were manually selected to represent the listing in the best possible way. To remove the annotation noise due to the potential presence of spammers, we used trusted research assistants for the annotation of a total of 49 physical, psychological, and overall attributes based on the PLSCI [12]. This step is presented in Section IV-A. (2) We analyzed the structure of the first round of annotations in a clustering experiment in the principal component analysis (PCA) space. This step is detailed in Section IV-B. (3) We conducted a large-scale annotation campaign on 1200 listings using Amazon Mechanical Turk (MTurk) and automatically selected images to represent the listing. This step is discussed in Section IV-C.

A. Trusted Annotations on Manually Selected Images

In order to assess the level of agreement that one can expect in an ideal case, we collected annotations on 200 randomly-sampled listings (100 for both Switzerland and Mexico). To ensure that the Airbnb listings could be represented in the best possible manner, we manually selected the three images to be displayed to observers. Images were selected such that they clearly depicted the indoor environment of the listing: whenever possible, the views were varied and pictures of different rooms were used. Manual selection was performed by one of the co-authors. To remove annotator noise due to the potential presence of spammers, we asked a pool of five trusted research assistants to give their impressions on the 200 Airbnb listings. In addition to the research assistants, 8 other trusted observers annotated an average of 25 listings. In total, each listing was annotated by five observers.

We compiled a list of adjectives to qualify the physical and ambiance properties of home environments based on the Personal Living Space Cue Inventory (PLSCI) [12], a questionnaire including 42 bipolar physical and ambiance attributes used to characterize home environments. We discarded attributes which could only be rated in situ (e.g., odor, temperature, noise level) and added to the list adjectives from related work [15], [42] and words we judged to be complementary to qualify indoor home environments. Our final list comprised 16 physical attributes (*clean*, *colorful*, *dark*), 28 ambiance attributes (*artsy*, *comfortable*, *upscale*), and 5 overall attributes (*I like*, *others like*). The full list of annotated attributes is displayed in Table II. For each adjective, we provided a definition of the word. All attributes were annotated on a 7-point Likert scale. Likert scales were recently found to yield reliable results for image aesthetics [44].

To assess the reliability of each annotated variable in the absence of ground truth, we used one of the Intraclass Correlation Coefficients (ICC), used in similar settings [15], [42], [44]. $ICC(1,k)$ assesses the degree of agreement in rating the targets (i.e.,, Airbnb listings) when the ratings are aggregated across the judges, and each target is assumed to be rated by a different set of judges. $ICC(1,k)$ is dependent on the number of annotations and the variance of the data; furthermore, $ICC(1,k)$ values can be problematic to interpret as no standard threshold exists to segment between e.g., moderate and high agreement. To address this, we compared our results with the literature investigating the agreement of judges on related social dimensions. We used a threshold of $ICC(1,k) = 0.50$ as a cut-off between low and high inter-rater agreement.

Table II displays the inter-rater agreement for all annotated variables. Most physical attributes had medium-to-high agreement, with 14 out of the 16 variables with $ICC(1,k)$ values over the 0.50 threshold. Physical attributes related to lighting (*dark*, *well-lit*), decoration (*colorful*, *decorated*), and size (*large*, *spacious*) had the highest agreement, with $ICC(1,k) > 0.70$. *Full* had the lowest agreement ($ICC(1,k) = 0.37$), which can be explained if the concept of fullness was either ambiguous or irrelevant in the context of Airbnb properties. The *clean* attribute had low agreement ($ICC(1,k) = 0.45$): this can be explained by the low variance combined with the strong negative skewness. The mean value for *clean* was high (5.85 for a maximum of 7), whereas the mean for *cluttered* was low (2.28), which indicates that most annotated Airbnb listings were tidied up before taking the pictures. This finding highlights a major difference between this study and the work of [14], which explicitly instructed occupants not to tidy up their spaces before the visit of observers. In this sense, photos of Airbnb listings have a decreased ecological validity due to self-presentation bias: hosts have a clear motivation to present their space in the best possible manner as it will directly affect their number of rentals. Additionally, in certain places, Airbnb provides photographers to take professional pictures of the place.

TABLE II
INTER-RATER AGREEMENT AND DESCRIPTIVE STATISTICS OF ANNOTATIONS MADE BY RESEARCH ASSISTANTS ON MANUALLY SELECTED IMAGES ($N_{listings} = 200$, $N_{raters} = 5$)

	ICC(1,k)	mean	std	skew
Physical attributes				
Clean	0.45	5.85	0.73	-0.97
Cluttered	0.57	2.28	0.87	1.28
Colorful	0.85	4.09	1.52	-0.01
In good condition	0.53	5.44	0.76	-0.45
Cramped	0.63	2.87	1.09	0.46
Dark	0.86	3.33	1.53	0.41
Decorated	0.77	4.25	1.24	-0.38
Full	0.37	3.27	0.90	0.55
Large	0.77	3.99	1.14	0.01
Neat	0.53	5.09	0.84	-0.48
New	0.73	3.48	1.17	0.62
Organized	0.57	4.81	0.77	-0.48
Practical	0.56	4.68	0.88	-0.50
Good use of space	0.60	4.67	0.86	-0.20
Spacious	0.72	4.32	1.10	-0.01
Well-lit	0.79	4.44	1.39	-0.16
Ambiance attributes				
Artsy	0.64	3.32	1.23	0.15
Bohemian	0.66	3.71	1.31	-0.00
Charming	0.76	3.61	1.30	0.14
Cheesy	0.50	3.28	1.06	0.43
Comfortable	0.72	4.09	1.11	-0.11
Conservative	0.50	3.71	1.07	-0.09
Contemporary	0.75	3.34	1.29	0.50
Cozy	0.68	3.92	1.15	-0.04
Dull	0.80	3.53	1.52	0.35
Eclectic	0.44	3.55	0.97	0.03
Kitsch	0.66	3.53	1.34	0.42
Luxurious	0.76	2.74	1.21	0.85
Modern	0.81	3.29	1.39	0.52
Old-fashioned	0.71	3.67	1.31	-0.05
Off-the-beaten-path	0.64	3.19	1.24	0.34
Pleasant	0.61	4.10	1.19	0.01
Pretentious	0.57	3.12	1.12	0.32
Relaxed	0.13	3.94	0.87	0.27
Romantic	0.55	2.95	1.14	0.56
Simple	0.77	4.36	1.39	-0.23
Sophisticated	0.69	3.27	1.22	0.23
Offbeat	0.59	3.04	1.15	0.35
Stylish	0.66	3.59	1.29	0.22
Traditional	0.35	3.67	0.90	-0.11
Trendy	0.70	3.02	1.25	0.71
Unique	0.72	3.18	1.35	0.30
Upscale	0.61	2.90	1.14	0.60
Overall attributes				
Short stay ("I would like to stay here for a short stay")	0.59	4.15	1.16	0.07
Long stay ("I would like to stay here for a long stay")	0.62	3.45	1.25	0.47
Estimated price	0.62	3.30	0.88	0.32
I like ("I like this place")	0.66	4.13	1.15	0.07
Others like ("I think others like this place")	0.69	4.39	1.02	-0.08

23 out of the 28 ambiance attributes had $ICC(1,k)$ over the 0.50 threshold. Ambiance attributes related to simplicity (*simple*, *dull*), modernity (*modern*, *contemporary*, *old-fashioned*), as well as *unique*, *charming*, and *luxurious* had the highest agreement ($ICC(1,k) > 0.70$). *Relaxed* and *traditional* had the lowest agreement ($ICC(1,k) = 0.13$ and 0.35, respectively), which suggests that they were difficult to rate from Airbnb images. All overall attributes had $ICC(1,k)$ over the 0.50 threshold; the highest values were obtained for *I like this place* and *I think others like this place*, which suggests that our research assistants agreed on what Airbnb places were attractive.

TABLE III
COMPARISON OF INTER-RATER AGREEMENT ($ICC(1,k)$) BETWEEN OUR STUDY AND THE ONES OBTAINED IN [42], FOR $k = 5$

	ICC(1,k) current study	**ICC(1,k)** from [42]
Artsy	0.64	0.61
Bohemian	0.66	0.45
Conservative	0.50	0.61
Dull/Dingy	0.80	0.59
Old-fashioned	0.71	0.57
Off-the-beaten-path	0.64	0.42
Romantic	0.55	0.70
Sophisticated	0.69	0.76
Trendy	0.70	0.53
Upscale	0.61	0.76

See Section IV-A for details.

We compared the inter-rater agreement found in this experiment with related works focusing on ambiance impressions [15], [42]. Because of the dependence of $ICC(1,k)$ on the number of ratings, we processed the data shared by the authors of [42] to do a fair comparison with our work. We sampled 5 out of 10 the MTurk annotations 200 times, and averaged the obtained $ICC(1,k)$. Table III shows the comparison between our results and the ones obtained in [42] for the overlapping attributes. We observe that our results are in general comparable with [42]: six out of the 10 ambiance attributes had higher agreement in our study. It should however be noted that the settings differed (home environments from Airbnb *vs.* popular indoor places on Foursquare, research assistants *vs.* MTurkers).

Overall, the results obtained in this first round of annotations indicate that observers can form reliable impressions on most physical, ambiance, and overall attributes from three manually selected pictures of Airbnb listings. This partly answers our first research question (**RQ1**): using trusted annotators, reliable impressions of physical, ambiance, and overall attributes can be made from manually selected images to represent Airbnb listings.

B. Clustering of Impressions

To understand the structure of impressions from Airbnb places, we conducted cluster analysis. Attributes with $ICC(1,k)$ lower than 0.50 were discarded (leaving 42 variables), and we standardized each variable such that it had zero-mean and unity variance. We then conducted principal component analysis (PCA) on the annotated data; the projection of the physical, ambiance, and overall attributes onto the two first components is shown in Fig. 2. We observe that the first principal component, accounting for 51.9% of the variance, had positive attributes on the right-hand side (*well-lit*, *upscale*, *romantic*, *charming*), while negative were found on the left-hand side (*dark*, *dull*, *cramped*). This observation is evocative of the fact that Airbnb listings were mainly rated on a good *vs.* bad basis, which is an effect known in psychology as the halo effect [32], where global evaluations tend to induce altered evaluations of other attributes. The second principal component, accounting for 15.7% of the variance, seemed to enclose attributes related to decoration (*decorated*, *colorful*), uniqueness (*unique*, *bohemian*, *offbeat*), and cheesiness (*cheesy*, *kitsch*) on the positive side.

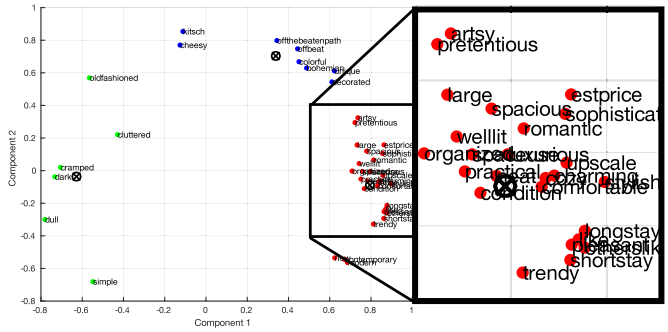

Fig. 2. Annotated attributes projected onto the first two components of the PCA space. Colors (red, green, and blue) denote the clusters to which the attributes belong, and the black crosses surrounded by circles are the cluster centers. Please view this with a PDF reader. The black rectangular border represents a closeup view of a cluster. For the complete list of attributes belonging to each cluster, please refer to Fig. 3.

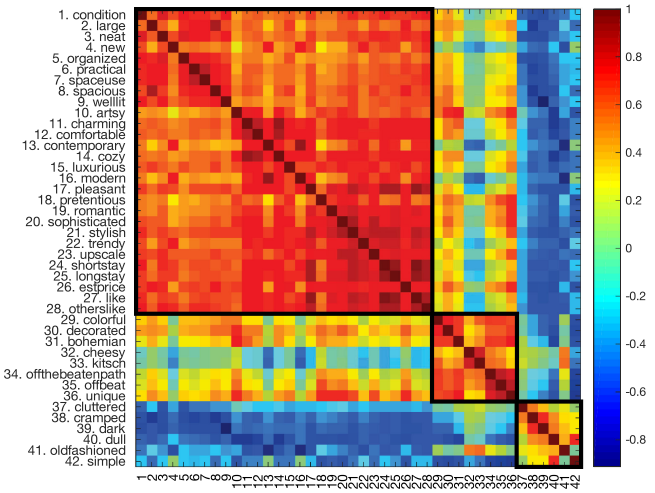

Fig. 3. Correlation matrix between the annotated variables listed in Table II. Black rectangular borders indicate the three distinct clusters found.

We then performed K-means clustering in the principal component space. We used the method proposed in [31], which used the Euclidean distance as a distance measure, and where the PCA coordinates were scaled by the square-root of the eigenvalues corresponding to each component prior to applying K-means clustering. Experiments for varying values of K were conducted, and $K = 3$ was a subjectively optimal choice; the main limitation of this approach is the manual selection of parameter K (number of clusters). Fig. 3 displays the correlations between the annotated variables, where attributes belonging to each cluster are grouped together. We first observe a dense cluster containing 28 out of the 42 annotated variables, including attributes that could be qualified as positive (*I like*, *comfortable*, *romantic*, *stylish*, etc.). The second cluster encompassed attributes related to decoration (*colorful*, *decorated*) and uniqueness (*unique*, *offbeat*), while the third one included negative attributes (*cluttered*, *dark*, *cramped*). This clustering procedure was also applied to physical and ambiance attributes taken separately, and similar results were obtained (not included here for space reasons). Also, as a verification step, factor analysis was also applied, and similar results were found. This answers

TABLE IV
LIST OF 'ACCEPTED' AND 'BORDERLINE' CLASSES USED FOR AUTOMATIC IMAGE SELECTION

Accepted classes	Borderline classes
attic, bedroom, dining room, living room, parlor	basement, closet, corridor, dorm room, dinette/home, game room, home office, hotel room, kitchen, kitchenette, lobby, office, patio, shower, staircase, veranda

our second research question (**RQ2**), annotated attributes can be grouped into three clusters: *positive*, *negative*, and *decorated/unique*. Please also note that the cluster analysis was also performed on physical and ambiance attributes separately, and similar results were obtained and are not reported due to space constraints.

In [38], clustering of ambiance dimensions from Foursquare public places was performed using a semi-automated method based on K-means, resulting in 18 clusters. The large difference between [38] and our work in terms of the number of clusters found (18 *vs.* 3) highlights the trade-off between high-granularity/numerous variables *vs.* low-granularity/few variables to characterize place ambiance that researchers have to make, in a somewhat arbitrary manner.

In any case, the results found in our work indicate that although a high granularity of annotations can be desirable, they can be partly redundant in reality, due to their strong inter-correlations. Also, the annotation cost is largely proportional to the number of desired variables; we therefore believe that understanding the structure of physical and ambiance attributes is a step forward to the understanding of home environments, as it enables the analysis of a larger number of places for the same cost. We also believe in the importance of aiming for a shorter version of the PLSCI [12], in a similar fashion to Ten-Item Personality Inventory (TIPI) to quantify personality traits of individuals [11].

C. MTurk Annotations on Automatically Selected Images

In addition to the 200 listings used in the two previous subsections where images were manually selected, we used 1000 randomly sampled Airbnb listings (500 for both Mexico and Switzerland), where images to be used to represent the places were automatically selected. To this end, we sampled 1500 listings (750 for both Mexico and Switzerland), where at least 6 images were present. We then applied a forward pass of the GoogLeNet CNN trained on the Places dataset [49] on the images to obtain the class probabilities. Then, for each place we randomly selected three images for which the highest class probabilities corresponded to indoor home environments (i.e.,, *accepted classes*, see Table IV). If fewer than three images belonged to the *accepted classes*, we added randomly sampled images from the *borderline* classes (see Table IV). In case the union of images belonging to the *accepted* and *borderline* classes was lower than three, the listing was discarded. The choice of *accepted* and *borderline* was motivated by the necessity to represent the indoor environment of Airbnb listings in the best possible way. We manually inspected the selected images and discarded listings with duplicate pictures and places where recognizable faces were present. From the 1500 original listings, 93 (34 for Switzerland, 59 for Mexico) were discarded because the number of accepted/borderline images was lower than 3. We further discarded 69 listings after manual inspection (36 for Switzerland, 33 for Mexico). As a last step, we randomly picked 1000 listings from the remaining places.

TABLE V
INTER-RATER AGREEMENT OF ANNOTATIONS COLLECTED ON MTURK ($N_{listings} = 1200$, $N_{raters} = 5$) AND COMPARISON BETWEEN AUTOMATIC AND MANUAL IMAGE SELECTION TO REPRESENT THE AIRBNB LISTINGS, USING $ICC(1,k)$

	Total N=1200	Automatic N=1000	Manual N=200
Physical attributes			
Large, spacious	0.74	0.74	0.70
Dark, badly-lit	0.71	0.70	0.76
Colorful, decorated	0.73	0.71	0.79
Cramped, confined	0.68	0.69	0.61
Bright, well-lit	0.71	0.70	0.73
Ambiance attributes			
Comfortable, cozy	0.54	0.53	0.60
Dull, simple	0.67	0.66	0.68
Cheesy, kitsch	0.29	0.29	0.27
Sophisticated, stylish	0.63	0.63	0.65
Off-the-beaten-path, unique	0.52	0.50	0.57
Overall attribute			
Overall like	0.65	0.64	0.66

Recent research in both psychology and computer science has demonstrated the feasibility of using crowdsourcsing to conduct behavioral studies in a fast and fully scalable manner [24]. We used MTurk to collect impressions on 11 dimensions from 1200 Airbnb properties. The choice of attributes to be annotated was based on the results of the cluster analysis (Section IV-B). For both physical and ambiance attributes, we selected adjective pairs which we judged to be most representative of the clusters. For instance, for physical attributes, *large, spacious* and *bright, well-lit* represented the *positive* cluster while *colorful, decorated* represented the cluster of the same name. In addition to the physical and ambiance attributes, we used the *overall liking* variable to capture the holistic impression about the listings.

The list of annotated variables is displayed in Table V. All variables were rated on 7-point Likert variable. Five annotations were collected per listing by MTurkers located in the U.S. who had over 95% approval rate. MTurkers were required to click on the three images to view them in full resolution before being able to start completing the questionnaire. Definitions of the words were available to raters in order to avoid confusion.

Table V shows the inter-rater agreement for the MTurk annotations. We first observe that all physical and ambiance attributes except *cheesy, kitsch* had high agreement, with $ICC(1,k)$ over the 0.50 threshold. This confirms the validity of ambiance impressions from Airbnb listings represented by three pictures of the indoor environment, which further answers our first research question (**RQ1**): reliable impressions from Airbnb images can also be obtained from MTurk on most dimensions.

Overall, higher agreement was obtained on physical attributes ($ICC(1,k) \geq 0.67$ for all variables) than for ambiance, which indicates that the physical characteristics of a place are easier to rate. One hypothesis to explain this observation is that ambiance might require a higher level of abstraction in the formation of first impressions, and idiosyncrasies (e.g., in the form of stereotypes) might occur in the inference process, leading to

TABLE VI
SUMMARY OF THE EXTRACTED FEATURES, WHERE D DENOTES THE FEATURE DIMENSIONALITY

Feature set	Description	D
Non-CNN		
RGB	Color histogram in RGB space	384
HSV	Color histogram in HSV space	384
HOG	Histogram of oriented gradients	680
Trained on ImageNet		
AlexNet (FC6)	6th fully connected layer of AlexNet	4096
AlexNet (FC7)	7th fully connected layer of AlexNet	4096
GoogLeNet (FC)	Fully connected layer of GoogLeNet	1024
Trained on Places		
AlexNet (FC6)	6th fully connected layer of AlexNet	4096
AlexNet (FC7)	7th fully connected layer of AlexNet	4096
GoogLeNet (FC)	Fully connected layer of GoogLeNet	1024

Note that activation features were extracted from CNNs trained on both the ImageNet and Places datasets.

divergence across raters [14]. We believe that this process could have occurred when forming impressions about *cheesy, kitsch* ($ICC(1,k) = 0.29$): this attribute requires a sense of irony that might not be shared across MTurkers. When rated by our research assistants, these attributes had relatively high inter-rater agreement ($ICC(1,k) = 0.50$ and 0.66 for *cheesy* and *kitsch*, respectively), and this can be explained by the lower diversity of this set of raters (university students in their twenties) compared to MTurkers.

The last two columns of Table V show the differences of inter-rater agreement between the two image selection methods used to represent the Airbnb listings. We observe a very minor decrease in $ICC(1,k)$ for automatic selection compared to manual. This indicates that automatically selecting images to represent a place does not decrease the quality of the annotations, which answers our third research question (**RQ3**): automatic image selection can be used to represent Airbnb places as well as manual selection. This finding differs from the work on ambiance impressions of popular indoor places on Foursquare, where the authors found manual selection to better represent the place compared to a purely random selection [42]. Here, our approach is based on the automatic scene classification based on a CNN trained on the Places dataset [49], and the results indicate that using this approach for selecting images to represent places is valid.

V. IMAGE REPRESENTATION

One of the main objectives of this work is to automatically infer first impressions of physical and ambiance attributes from Airbnb places, using images as sole modality. Recent related works predicting high-level human judgments have used a wide variety of features to represent images, such as GIST, HOG, SIFT, color histograms, and activation layers of pre-trained deep convolutional neural networks (CNNs) [30], [33], [35]. In this work, we applied a similar methodology to represent Airbnb images: we extracted features related to color and gradient distributions, as well as activation layers of pre-trained CNNs. This section describes the method used to obtain the image representations, and Table VI summarizes the features used in this work.

A. Non-CNN Features

a) Color Histograms: To obtain a color representation of images, we extracted color histograms. The images were first divided into $n_x/4 \times n_y/4$ blocks, where n_x and n_y denote the original image width and height, respectively. For each block, we then computed the histograms of pixel intensities for each color channel, using 8 histogram bins; histograms were normalized. Color histograms were computed using both RGB and HSV representations. The dimensionality of the color histograms was $D = 384$ (16 image blocks $\times$ 8 histogram bins $\times$ 3 color channels).

b) Histograms of Oriented Gradients (HOG): To obtain a gradient representation of the Airbnb images, we extracted histograms of oriented gradients (HOG). To this end, we used the method presented in [5]. We used the default parameters: $n_{bins} = 8$, $L = 3$, with n_{bins} and L denoting the number of histogram bins and number of pyramid levels, respectively, the dimensionality of the feature representation is $D = 680$ ($1 + 4 + 16 + 64$ image blocks $\times$ 8 histogram bins).

B. Activation Layers of Deep CNNs

Deep convolutional neural networks have enabled significant improvements in image classification tasks [49]. Recent studies have shown that features extracted from upper layers of pre-trained CNN models can yield competitive results for generic tasks, even when the network was trained for an unrelated task [3], [9], which partially removes the need to train or adapt a CNN to a new dataset. This method has been successfully used in perceptual tasks including the prediction of aesthetic qualities of images [30] and high level urban perceptions [33], [35].

We investigate the use of activation features from pre-trained CNN models for the task of predicting first impressions of indoor home environments, which to our knowledge has not been previously studied. Specifically, we examine two CNN architectures trained on two specific datasets.

1) Models: We extracted activation features from two popular CNN architectures:

a) **AlexNet** [25] won the ILSRVC-2012 contest [40] and consists of five convolutional layers, three fully-connected layers, and a final softmax layer. In this work, we used the two last fully-connected layers (denoted *FC6* and *FC7*) as they were shown to perform competitively across datasets of different natures [3], [9].

b) **GoogLeNet** [45] constitutes the current state-of-the-art for image classification and is the winner of ILSRVC-2014 contest [40]. We extracted the fully-connected layer right before the softmax, denoted *FC*.

2) Datasets: Both the AlexNet and GoogLeNet architectures were trained on two specific datasets.

a) **ImageNet** [40] is one of the major benchmark datasets in image classification. It consists of 1.2M+ images, where each image belongs to one of 1000 categories. With image categories including animals, food, objects, or scenes, ImageNet constitutes a general-purpose dataset.

b) **Places** [49] is a scene-centric dataset of 7M+ labeled pictures of scenes, including 205 categories of indoor and outdoor scenes. To our knowledge, it constitutes the closest existing database to our specific setting of Airbnb home environments.

The activation layers from the four possible model/dataset pairs (AlexNet/GoogLeNet trained on ImageNet/Places) were extracted from pre-trained networks, using the Caffe framework [22]. We used the pre-trained models from the Caffe model zoo [1].

VI. Methodology

A. Inference Task

We defined the inference problem as a regression task, where the goal was to predict the MTurk annotations from the 1200 listings of Section IV-C. The inference task was conducted at the listing level, considering each Airbnb place as a separate data point, which differs from considering each image as a separate entity, as it is the case in other datasets such as ImageNet [40]. Prediction at the listing level constitutes a relevant task in the context of on-demand home environments, because a place is made of a collection of images.

B. Regression Method

We used Random Forest [6] as regression model, using the standard parameters ($n_{trees} = 500$, and $m_{try} = D/3$, where D denotes the feature dimensionality), with no dimensionality reduction. We used 10-fold cross-validation for training and testing, ensuring that the test set was completely separated from the training set. As the baseline regression model, we took the average annotated variable as the predicted value.

C. Evaluation Measures

To quantify the performance of the prediction models, we used Pearson's correlation coefficient (r), root-mean-square error ($rmse$), and coefficient of determination (R^2), as these are three widely used measures in both psychology and pattern recognition. R^2 is based on the ratio between the mean squared errors of the predicted values obtained using a regression model and the baseline-average model, and can be seen as the amount of variance explained by the tested model.

D. Images Used for Prediction

The fact that each listing includes a varying number of images raises the following question: What is the effect of the choice of images to represent an Airbnb listing on the prediction accuracy? To address this issue, we considered three different sets of images used for prediction.

1) **3-Image Subset** includes the three images that were used as stimuli to the MTurk annotators to provide their first impressions of ambiance and physical attributes. Because the stimulus is the same as the input used for prediction, we hypothesize that this image subset will yield the best performance; however, this image subset is unrealistic in the eventuality of a real-world deployment of the system, as these manually selected three images do not exist as such for unseen test data.
2) **Indoor Places Subset** includes all images of a place for which the most probable category belongs to the union of the *accepted* and *borderline* categories of Table IV, using the output of GoogLeNet trained on the Places dataset. This procedure ensures that the images used as input for ambiance prediction are similar in nature to the ones used as stimuli for annotation.
3) **All Images Subset** includes all images of a listing.

E. Aggregation Methods

To aggregate features stemming from multiple images belonging to an Airbnb listing, we evaluated two methods.

1) **Concatenation** is the baseline aggregation method and consists in concatenating the feature vectors of the images of a given place. Because the dimensionality of the feature representations of the listings must remain the same, the concatenation aggregation method was only used with the *3-Image Subset*. Note that in this case the order of concatenation is arbitrary.
2) **VLAD** stands for Vector of Locally Aggregated Descriptors [21] and is a feature encoding and pooling technique. VLAD first learns a codebook of k cluster centers c_i from a dictionary of descriptors, using k-means. Then, each local descriptor x is associated to its nearest cluster center; the VLAD descriptor accumulates, for each cluster c_i, the differences $x - c_i$ of the descriptors x assigned to c_i, and an l_2-normalization is applied to the final feature vector. Assuming the local descriptor to be d-dimensional, the obtained dimension of the new representation is $D = k \times d$. In our specific case, the local descriptor x was the feature vector representing one image, and the codebook was learned from the set of images belonging to the 1200 listings, using $k = \{1, 2, 3\}$ as the number of clusters. Please also note that Fisher vectors [34] were also used, but because the results were very similar to VLAD, they are not discussed in the paper.

VII. Experiments and Results

A. Image Representations

We first compared the performance obtained from the individual image representations presented in Section V. To this end, we used the *3-Image Subset* for the 1200 listings, and the *Concatenation* aggregation method. Table VII displays the inference results for physical and ambiance attributes for this experiment. Results demonstrate the possibility to infer first impressions from Airbnb pictures using a fully automated method, with over 40% of the variance explained by the model for three of the five physical attributes (*dark, badly-lit*; *colorful, decorated*; *bright, well-lit*), and over 30% for three of the six ambiance and overall attributes (*dull, simple*; *sophisticated, stylish*; *overall like*). This answers our fourth research question (**RQ4**): high-level judgments of ambiance and physical attributes can be automatically inferred from Airbnb images.

Our results indicate that physical attributes were easier to infer. This can be explained by the higher level of agreement among raters found in Section IV-C. Furthermore, we hypothesize that the process of forming ambiance impressions is more complex as idiosyncrasies might be part of the process, which might lead to divergence across raters [14]. Physical attributes

TABLE VII
INFERENCE RESULTS FOR VARYING IMAGE REPRESENTATIONS, USING THE *CONCATENATION* AGGREGATION METHOD ON THE *3 IMAGES SUBSET*

Ambiance attributes	**Comfortable, cozy**			**Dull, simple**			**Cheesy, kitsch**			**Sophisticated, stylish**			**Off-the-beaten-path, unique**			**Overall like**		
	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$
Baseline-Avg	0.00	0.00	0.92	0.00	0.00	1.26	0.00	0.00	0.86	0.00	0.00	1.23	0.00	0.00	1.12	0.00	0.00	1.11
Non-CNN																		
RGB	0.45	0.18	0.83	0.53	0.25	1.09	0.16	0.02	0.85*	0.38	0.14	1.14	0.44	0.18	1.01	0.46	0.20	0.99
HSV	0.45	0.19	0.83	0.54	0.26	1.08	0.22	0.05	0.84	0.38	0.13	1.15	0.46	0.20	1.00	0.46	0.19	0.99
HOG	0.29	0.08	0.88	0.32	0.09	1.20	0.03*	-0.02	0.87*	0.25	0.06	1.19	0.31	0.09	1.07	0.28	0.07	1.07
ImageNet																		
AlexNet (FC6)	0.46	0.17	0.84	0.53	0.23	1.11	0.24	0.05	0.84	0.44	0.15	1.13	0.52	0.21	0.99	0.48	0.19	1.00
AlexNet (FC7)	0.45	0.17	0.84	0.54	0.25	1.09	0.26	0.07	0.83	0.45	0.16	1.13	0.53	0.24	0.97	0.49	0.20	0.99
GoogLeNet (FC)	0.47	0.18	0.83	0.56	0.25	1.09	0.28	0.07	0.83	0.45	0.15	1.14	0.52	0.23	0.98	0.50	0.20	0.99
Places																		
AlexNet (FC6)	0.52	0.22	0.81	0.59	0.29	1.06	0.31	0.09	0.82	0.55	0.25	1.06	0.54	0.23	0.98	0.58	0.28	0.94
AlexNet (FC7)	**0.53**	**0.25**	**0.80**	0.62	0.34	1.03	0.33	0.10	0.81	0.55	0.28	1.05	**0.57**	**0.29**	**0.94**	0.57	0.29	0.93
GoogLeNet (FC)	0.51	0.24	0.80	**0.63**	**0.35**	**1.01**	0.32	0.10	0.82	**0.57**	**0.30**	**1.03**	0.56	0.28	0.95	**0.59**	**0.31**	**0.92**

Physical attributes	**Large, spacious**			**Dark, badly-lit**			**Colorful, decorated**			**Cramped, confined**			**Bright, well-lit**		
	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$
Baseline-Avg	0.00	0.00	1.20	0.00	0.00	1.19	0.00	0.00	1.19	0.00	0.00	1.20	0.00	0.00	1.14
Non-CNN															
RGB	0.36	0.12	1.13	**0.66**	**0.42**	**0.91**	0.50	0.23	1.04	0.34	0.11	1.14	**0.67**	**0.42**	**0.87**
HSV	0.37	0.12	1.13	0.64	0.38	0.94	0.56	0.28	1.00	0.33	0.10	1.14	0.65	0.39	0.89
HOG	0.30	0.08	1.15	0.31	0.09	1.14	0.25	0.06	1.15	0.27	0.07	1.16	0.32	0.10	1.08
ImageNet															
AlexNet (FC6)	0.48	0.19	1.08	0.61	0.30	1.00	0.62	0.30	0.99	0.43	0.15	1.11	0.61	0.29	0.96
AlexNet (FC7)	0.50	0.21	1.07	0.59	0.28	1.01	0.63	0.33	0.97	0.45	0.17	1.10	0.62	0.29	0.96
GoogLeNet (FC)	0.47	0.18	1.09	0.58	0.24	1.04	0.63	0.31	0.98	0.45	0.16	1.10	0.58	0.24	0.99
Places															
AlexNet (FC6)	0.57	0.27	1.03	0.65	0.34	0.97	0.64	0.33	0.97	0.54	0.24	1.05	0.65	0.35	0.92
AlexNet (FC7)	0.58	0.31	1.00	0.64	0.36	0.96	**0.69**	**0.41**	**0.91**	0.55	0.27	1.03	0.66	0.38	0.90
GoogLeNet (FC)	**0.62**	**0.34**	**0.98**	0.58	0.29	1.01	0.67	0.40	0.92	**0.57**	**0.30**	**1.01**	0.60	0.30	0.95

related to illumination (*dark, badly-lit*; *bright, well-lit*) and colors (*colorful, decorated*) obtained the best inference results, with $R^2 > 0.40$ ($r > 0.65$). This result is somewhat unsurprising as these attributes are closely linked to the distribution of pixel intensities of the images; the relatively high level of variance explained by the color histogram features supports this hypothesis. Attributes related to the perceived size of the place (*large, spacious*; *cramped, confined*) obtained lower yet still competitive results. For ambiance and overall attributes, all variables obtained $R^2 > 0.25$ ($r > 0.50$), except for *cheesy, kitsch*. The low performance ($R^2 = 0.10$) obtained for *cheesy, kitsch* can be explained by its low inter-rater agreement due to the difficulty in forming consistent impressions for this dimension. In contrast, our method allows to infer the holistic variable of *overall like* with $R^2 = 0.31$ ($r = 0.59$), which is a promising result.

In terms of individual feature sets, the activation features of CNNs trained on the Places dataset [49] consistently yielded the best results for all dimensions (see Table VII), except for illumination-related physical attributes (*dark, badly-lit*; *bright, well-lit*) where they still performed well. While activation features trained on ImageNet yielded competitive results, they were consistently less accurate than the ones trained on Places. This finding can be explained by the fact that Places includes a large number of images of home environments; in other words, the amount of visual content overlap between the dataset and our setting was high. Our results confirm the usefulness of the activation layer of pre-trained CNNs used as image representation for perceptual tasks [9], [33].

Our results also suggests that the network architecture did not play a crucial role for our specific task: minor differences were observed between AlexNet and GoogLeNet for a fixed training dataset. In terms of AlexNet fully-connected layers, results obtained for *FC7* were slightly but consistently higher than the ones for *FC6* for Places, while no difference could be observed for ImageNet.

The poor results obtained with HOG indicate that the texture-only representation of images is relatively uninformative of the ambiance and physical characteristics of indoor places, whereas simple features from color histograms were very informative of the illumination dimensions of the places (*dark, badly-lit*; *bright, well-lit*), while yielding quite competitive results for some other ambiance and physical attributes (*colorful, decorated*; *dull, simple*).

In comparison with related works using other types of variables and settings, the results obtained in this work are competitive. Ordonez *et al.* [33] reported Pearson correlation coefficients of 0.54, 0.67, and 0.72 for the inference of safety, uniqueness, and wealth, respectively, in outdoor public places. In terms of inference results, our work cannot be directly compared to [38] as the setting, scale, and evaluation measures largely differ.

B. Images Used for Prediction and Aggregation

To understand the effect of the images used for prediction on the accuracy of the inference, we evaluated the use of the image subsets of Section VI-D. Additionally, we tested the aggregation methods of Section VI-E. Please note that the *Concatenation* method can only be used for the *3-Image Subset*, because the number of images per place varies for both the *Indoor Places Subset* and the *All Images Subset*. For this experiment, we used the GoogLeNet activation features trained on Places, as they were found in Section VII-A to consistently yield top results; additionally, their dimensionality is four times smaller than AlexNet activation layers for which comparable results were obtained.

Table VIII presents the prediction results obtained for the different image subsets and aggregation methods. In terms of images used to represent the Airbnb listings, we observe an important decrease in performance between the *3-Image Subset* and the *Indoor Places Subset*, with $\Delta r \in [0.09, 0.19]$ and

TABLE VIII
INFERENCE RESULTS FOR VARYING IMAGE SUBSETS AND AGGREGATION METHODS, USING THE ACTIVATION LAYER OF GOOGLENET TRAINED ON PLACES

Ambiance attributes	**Comfortable, cozy**			**Dull, simple**			**Cheesy, kitsch**			**Sophisticated, stylish**			**Off-the-beaten-path, unique**			**Overall like**		
	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$
Baseline-Avg	0.00	0.00	0.92	0.00	0.00	1.26	0.00	0.00	0.86	0.00	0.00	1.23	0.00	0.00	1.12	0.00	0.00	1.11
3-Image Subset																		
Concatenation	0.51	0.24	0.80	0.63	0.35	1.01	0.32	0.10	0.82	0.57	0.30	1.03	0.56	0.28	0.95	0.59	0.31	0.92
VLAD (k=1)	0.53	0.28	0.78	0.65	0.41	0.97	0.34	0.12	0.81	0.59	0.34	1.00	0.58	0.33	0.92	0.59	0.34	0.90
VLAD (k=2)	0.51	0.25	0.80	0.64	0.39	0.99	0.34	0.11	0.81	0.58	0.32	1.01	0.57	0.30	0.94	0.58	0.32	0.91
VLAD (k=3)	0.51	0.25	0.80	0.63	0.36	1.01	0.33	0.11	0.81	0.57	0.30	1.03	0.55	0.28	0.95	0.57	0.31	0.92
Indoor Places Subset																		
VLAD (k=1)	0.42	0.17	0.84	0.53	0.27	1.08	0.22	0.05	0.84*	0.48	0.22	1.08	0.46	0.21	0.99	0.50	0.24	0.97
VLAD (k=2)	0.41	0.16	0.84	0.54	0.28	1.07	0.21	0.04	0.84*	0.48	0.22	1.09	0.46	0.21	1.00	0.49	0.23	0.97
VLAD (k=3)	0.39	0.15	0.85	0.53	0.26	1.08	0.23	0.05	0.84	0.47	0.20	1.10	0.47	0.22	0.99	0.46	0.21	0.98
All Images Subset																		
VLAD (k=1)	0.40	0.15	0.85	0.50	0.24	1.10	0.17	0.02	0.85*	0.45	0.20	1.10	0.46	0.21	0.99	0.49	0.23	0.97
VLAD (k=2)	0.40	0.15	0.85	0.51	0.25	1.09	0.18	0.03	0.85*	0.45	0.19	1.10	0.47	0.22	0.99	0.49	0.23	0.97
VLAD (k=3)	0.40	0.15	0.85	0.52	0.26	1.09	0.18	0.03	0.85*	0.47	0.21	1.10	0.47	0.22	0.99	0.49	0.22	0.98

Physical attributes	**Large, spacious**			**Dark, badly-lit**			**Colorful, decorated**			**Cramped, confined**			**Bright, well-lit**		
	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$	r	R^2	$rmse$
Baseline-Avg	0.00	0.00	1.20	0.00	0.00	1.19	0.00	0.00	1.19	0.00	0.00	1.20	0.00	0.00	1.14
3-Image Subset															
Concatenation	0.62	0.34	0.98	0.58	0.29	1.01	0.67	0.40	0.92	0.57	0.30	1.01	0.60	0.30	0.95
VLAD (k=1)	0.63	0.38	0.95	0.62	0.36	0.96	0.69	0.46	0.87	0.58	0.32	0.99	0.61	0.36	0.91
VLAD (k=2)	0.63	0.38	0.95	0.60	0.32	0.98	0.67	0.42	0.90	0.58	0.32	0.99	0.59	0.32	0.94
VLAD (k=3)	0.62	0.36	0.96	0.58	0.30	0.99	0.66	0.39	0.92	0.57	0.30	1.01	0.58	0.30	0.95
Indoor Places Subset															
VLAD (k=1)	0.52	0.26	1.03	0.43	0.18	1.08	0.54	0.28	1.01	0.48	0.22	1.06	0.44	0.18	1.03
VLAD (k=2)	0.53	0.27	1.03	0.44	0.17	1.08	0.56	0.29	1.00	0.49	0.23	1.06	0.44	0.18	1.03
VLAD (k=3)	0.51	0.25	1.05	0.41	0.16	1.10	0.55	0.28	1.01	0.47	0.21	1.07	0.43	0.17	1.04
All Images Subset															
VLAD (k=1)	0.51	0.25	1.05	0.41	0.16	1.09	0.51	0.25	1.03	0.48	0.22	1.07	0.41	0.16	1.04
VLAD (k=2)	0.50	0.24	1.05	0.41	0.16	1.09	0.51	0.25	1.03	0.47	0.21	1.07	0.41	0.16	1.04
VLAD (k=3)	0.51	0.25	1.04	0.41	0.16	1.09	0.54	0.27	1.01	0.50	0.23	1.06	0.44	0.18	1.03

$\Delta R^2 \in [0.10, 0.18]$. This can be explained by the fact that the *3-Image Subset* is composed of the three images used as stimuli to the MTurkers for rating the places, which constitutes an ideal (but unrealistic) case. Indeed, in the case of the deployment of an automated prediction system, where the ambiance of unseen Airbnb listings would be inferred, the *3-Image Subset* would not be available as such. The results obtained for *All Images Subset* are very similar to the *Indoor Places Subset*, which suggests that filtering out images unrelated to indoor environments prior to predicting the ambiance and physical attributes does not significantly improve the performance.

In terms of aggregation methods, *VLAD* outperformed the *Concatenation* method on the *3-Image Subset* for most attributes. This result can be explained by the fact that the order of the concatenation is arbitrary, which might introduce some bias. Overall, for the *3-Image Subset* the best prediction performances were obtained using *VLAD* with $k = 1$, which is equivalent to averaging the representations (average pooling) of the three images [21]. This result suggests that the aggregation step could be considered as *pooling*, and that other pooling methods (e.g., max-pooling, weighted pooling) could be used; this will be investigated as future work. An increase in the number of clusters resulted in a slight decrease in prediction accuracy, likely due to the relatively small number of images used in clustering. For the *Indoor Places* and *All Images* subsets, the number of clusters did not affect the prediction accuracy. Please also note that we conducted the same experiments with Fisher vectors [34], and results were very close to *VLAD*, therefore we did not present them here.

C. Visualization

As illustration, Fig. 4 displays scatter and histogram plots of annotated and predicted scores for the variables of *overall like* and *colorful, decorated*. The scatter plots show that predicted scores were biased towards the center: high scores were under-evaluated while low scores were over-evaluated; furthermore, upon inspection of the histogram plots, one can observe that the range of predicted scores is smaller than the annotated scores. This suggests that our method tends to have a limited prediction range and that the extreme values constitute the main sources of errors. This effect can be explained by the unbalance in the distribution of annotations: extreme instances were under-represented. This problem has recently been addressed by introducing sample weights during training [23], which constitutes a possible avenue for improvement in future work.

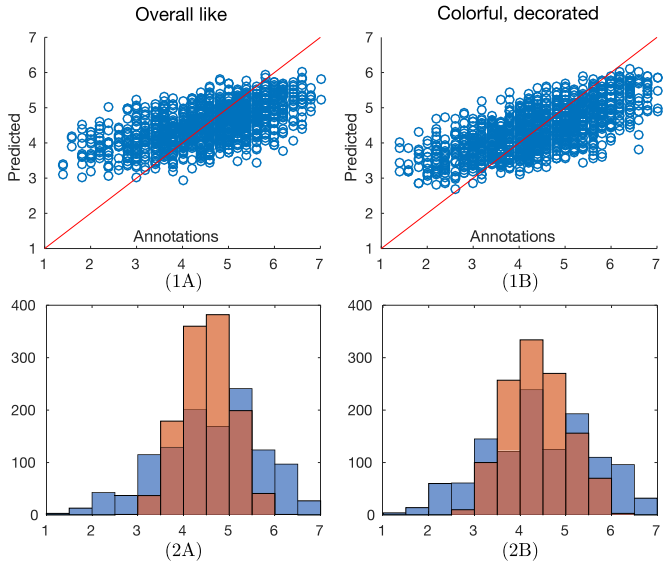

Fig. 4. Visualization examples for (A) *overall Like* ($r = 0.59$, $R^2 = 0.34$) and (B) *colorful, decorated* ($r = 0.69$, $R^2 = 0.46$), using GoogLeNet activation features trained on Places, with *3-Image Subset*, and *VLAD (k=1)*. (1) Scatter plots for predicted vs. annotated scores; (2) histogram plots for predicted scores (in red) and annotated scores (in blue).

VIII. Conclusion

In this work, we analyzed the impressions that can be formed from Airbnb places, using images as one of the site's most important modalities. We first collected Airbnb data, using Mexico and Switzerland as case studies. In total, 350 K images were collected from 22 K listings. The analysis of the image categories indicates that most images indeed depict indoor home environments, which partly validates the use of Airbnb data to study personal living spaces. To understand the types of impressions that could be formed from Airbnb properties, 200 listings were annotated (using manually selected images to represent the place), and results showed that a high level of agreement could be achieved. Based on our cluster analysis, these annotations can be grouped into three main clusters: *positive*, *negative*, and *decorated/unique*. Furthermore, the generation of MTurk annotations on automatically selected images to represent the home environments has demonstrated that annotations of Airbnb listings can be reliably crowdsourced, and that automatic image selection can be used, opening new perspectives for large-scale home environment research. Last, we demonstrated the feasibility to automatically infer high-level impressions of Airbnb listings (up to a certain level) using image features. Overall, the best image representations were activation layers of deep CNNs trained on the Places dataset.

Several directions for future work to improve the accuracy of the automatic prediction of ambiance and physical attributes exist: they include the analysis and fusion of other modalities to the image features, such as text (in the form of listing description or comments) or location; the use of sample weights in the training process to reduce the prediction bias towards the median score; fine-tuning of existing CNNs to our specific task; the use of different pooling methods for feature aggregation; and conducting the prediction task at the image level instead of the listing level.

Moreover, several important aspects were not addressed, which we would like to investigate in the future. First, we would like to understand the basis on which observers form their impressions from Airbnb places; in other words, what are the cues used by raters to make judgments on places? Second, little is known about the inter-cultural differences (e.g., cities *vs.* rural areas, or across countries) on Airbnb despite its worldwide penetration. Inter-cultural differences might also be found in the way raters perceive pictures of shared homes, and Airbnb (in combination with online crowdsourcing methods) constitutes a setting that enables to investigate the norms and expectations of raters based on factors including geographical location, age, gender, personality, or socio-economical status. Third, one might argue that Airbnb is not solely composed of home environments, as the platform is also used by traditional bed and breakfasts and rental houses; it would be interesting to analyze differences of listings. Fourth, Airbnb constitutes an opportunity to study relationships between people (hosts and guests) and the perception of indoor spaces as presented in social media.

To our knowledge, this work constitutes the first study analyzing first impressions from Airbnb pictures. Airbnb constitutes an unprecedented opportunity to study home environments at scale, which could benefit to both the multimedia and psychology communities. For multimedia, a clear direction for future work is the collection and annotation of a large dataset of hundreds of thousands listings to train predictive models of homes, in a similar fashion to the Places dataset, but focusing on home environments. For psychologists, analyzing Airbnb places at large scale will allow to better understand the structure of ambiance attributes to ultimately find a small number of possibly orthogonal factors, clearing the path for the development of a short and comprehensive psychometric instrument, not unlike the Ten-Item Personality Inventory (TIPI) for personality [11]. In any case, we believe in the benefits of the collaboration between the two communities in this specific context, as complementary knowledge can be found at both sides.

Acknowledgment

The authors would like to thank D. Santani (Idiap) for discussions and all the MTurk workers for contributing their impressions.

References

[1] Caffe Model Zoo. [Online]. Available: https://github.com/BVLC/caffe/wiki/Model-Zoo. Retrieved on: Jul. 4, 2016.
[2] Inside Airbnb. Adding data to the debate. [Online]. Available: http://insideairbnb.com/. Retrieved on: Mar. 27, 2016.
[3] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 584–599.
[4] S. Bakhshi, D. Shamma, and E. Gilbert, "Faces engage us: Photos with faces attract more likes and comments on Instagram," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2014, pp. 965–974.
[5] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2007, pp. 401–408.
[6] L. Breiman, "Random forests," *J. Mach. Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[7] D. Carney, J. Jost, S. Gosling, and J. Potter, "The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind," *Pol. Psychol.*, vol. 29, no. 6, 2008, pp. 807–840.
[8] Oxford Dictionary, "Definition of ambiance in English," [Online]. Available: https://en.oxforddictionaries.com/definition/ambience. Retrieved on Mar. 14, 2017.
[9] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 647–655.
[10] B. Edelman and M. Luca, "Digital discrimination: The case of Airbnb.com," Harvard Bus. School Working Paper 14-054, Jan. 2014.
[11] S. Gosling, "A very brief measure of the Big-Five personality domains," *Res. Pers.*, vol. 37, pp. 504–528, 2003.
[12] S. Gosling, K. Craik, N. Martin, and M. Pryor, "The personal living space cue inventory: An analysis and evaluation," *Environ. Behav.*, vol. 37, pp. 683–705, 2005.
[13] S. Gosling, R. Gifford, and L. McCunn, "The selection, creation, and perception of interior spaces: An environmental psychology approach," in *The Handbook of Interior Architecture and Design*. G. Brooker and L. Weinthal, Eds. London, U.K.: Bloomsbury, pp. 278–290, 2013.
[14] S. Gosling, S. Ko, T. Mannarelli, and M. Morris, "A room with a cue: Personality judgments based on offices and bedrooms," *Pers. Soc. Psychol.*, vol. 82, no. 3, pp. 379–398, 2002.
[15] L. Graham and S. Gosling, "Can the ambiance of a place be determined by the user profiles of the people who visit it?" in *Proc. Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 145–152.
[16] L. Graham, S. Gosling, and C. Travis, "The psychology of home environments: A call for research on residential space," *Perspectives Psychol.*, vol. 10, pp. 346–356, 2015.

[17] D. Guttentag, "Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector," *Current Issues Tourism*, vol. 18, pp. 1192–1217, 2015.
[18] J. Hays and A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. Comput. Vis. Pattern Recogn.*, 2008, pp. 1–8.
[19] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the intrinsic memorability of images," in *Proc. Neural Inf. Process. Syst.*, 2011, pp. 2429–2437.
[20] Y. J. Lee, A. Efros, and M. Hebert, "Style-aware mid-level representation for discovering visual connections in space and time," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 1857–1864.
[21] H. Jégou, D. Matthijs, and P. P. Schmid, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2010, pp. 3304–3311.
[22] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[23] B. Jin, M. V. Ortiz Segovia, and S. Süsstrunk, "Image aesthetic predictors based on weighted CNNs," in *Proc. Int. Conf. Image Process.*, 2016, pp. 2291–2295.
[24] A. Kittur, E. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2008, pp. 453–456.
[25] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[26] A. Lampinen, "Hosting via Airbnb: Motivations and financial assurances in monetized network hospitality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2016, pp. 1669–1680.
[27] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. Comput. Vis. Pattern Recogn.*, 2006, pp. 2169–2178.
[28] D. Lee *et al.*, "An analysis of social features associated with room sales of Airbnb," in *Proc. ACM Conf. Companion Comput. Supported Cooperative Work Social Comput.*, 2015, pp. 219–222.
[29] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating Image Aesthetics Using Deep Learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, Nov. 2015.
[30] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1784–1791.
[31] L. Nguyen and D. Gatica-Perez, "Hirability in the wild: Analysis of online conversational video resumes," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1422–1437, Jul. 2016.
[32] R. Nisbett and T. D. Wilson, "The halo effect: Evidence for unconscious alteration of judgments," *Pers. Soc. Psychol.*, vol. 35, pp. 250–256, 1977.
[33] V. Ordonez and T. Berg, "Learning high-level judgments of urban perception," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 494–510.
[34] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.
[35] L. Porzi, S. Rota Bulo, B. Lepri, and E. Ricci, "Predicting and understanding urban perception with convolutional neural networks," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 139–148.
[36] G. Quattrone and D. Proserpio, "Who benefits from the "sharing" economy of Airbnb?" in *Proc. Int. World Wide Web Conf.*, 2016, pp. 11–15.
[37] S. Rahimi, X. Liu, and C. Andris, "Hidden style in the city: An analysis of geolocated Airbnb rental images in ten major cities," in *Proc. ACM SIGSPATIAL Workshop*, 2016, pp. 1–7.
[38] M. Redi, D. Quercia, L. Graham, and S. Gosling, "Like partying? Your face says it all. Predicting the ambiance of places with profile pictures," in *Proc. Int. Conf. Web Social Media*, 2015, pp. 347–356.
[39] H. Rheingold and K. Cook, "The content of boys' and girls' rooms as an index of parents' behavior," *Child Develop.*, vol. 46, no. 2, pp. 459–463, 1975.
[40] O. Russakovsky *et al.* "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[41] P. Salesses, K. Schechtner, and C. Hidalgo, "The collaborative image of the city: Mapping the inequality of urban perception," *PloS One*, vol. 8, no. 7, 2013, Art. no. e68400.
[42] D. Santani and D. Gatica-Perez, "Loud and trendy: Crowdsourcing impressions of social ambiance in popular indoor urban places," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 211–220.
[43] D. Santani, R. Hu, and D. Gatica-Perez, "InnerView: Learning place ambiance from social media images," in *Proc. ACM Multimedia Conf.*, 2016, pp. 451–455.
[44] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1338–1350, Jul. 2016.
[45] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 1–9.
[46] E. Tang and K. Sangani, "Neighborhood and price prediction for San Francisco Airbnb listings," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2015.
[47] X. Tian, Z. Dong, K. Yang, and T. Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2035–2048, Nov. 2015.
[48] G. Zervas, D. Proserpio, and J. Byers, "A first look at online reputation on Airbnb, where every stay is above average," Boston Univ., Tech. rep., 2015.
[49] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using Places database," in *Proc. Neural Inf. Process. Syst.*, 2014, vol. 27, pp. 487–495.

Laurent Son Nguyen received the Ph.D. degree in 2015 from École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, on the automated analysis of human behavior in job interviews. He is currently a Postdoctoral Researcher in the Social Computing Group, Idiap Research Institute, Martigny, Switzerland. His research interests include computational approaches to understand the formation of first impressions in face-to-face interactions, online videos, and social media.

Salvador Ruiz-Correa (M'00) received the Ph.D. degree in electrical engineering from the University of Washington, Seattle, WA, USA. He is currently an Adjucnt Professor and a Researcher in the Instituto Potosino de Investigacin Cientfica y Tecnolgica, San Luis Potosí, Mexico, where he is also the lead of the Youth Innovation Laboratory (You-i Lab). He codirects the Center for Mobile Life (Ce Mobili) research initiative in Mexico. His research interests range from machine learning and computer vision applications to social computing in development contexts, data for social good, and citizen innovation. He is a member of the Sistema Nacional de Investigadores of Consejo Nacional de Ciencia y Tecnologa (CONACYT).

Marianne Schmid Mast received the Ph.D. degree in psychology from the University of Zurich, Zurich, Switzerland, and pursued research at Northeastern University, Boston, MA, USA. She was an Assistant Professor of social psychology at the University of Fribourg and a Full Professor in the Department of Work and Organizational Psychology, University of Neuchatel. She is currently a Full Professor of organizational behavior in HEC, University of Lausanne, Lausanne, Switzerland. Her research interests include how individuals in power hierarchies interact, perceive, and communicate, how first impressions affect interpersonal interactions and evaluations, and how people form accurate impressions of others. She is currently an Associate Editor for the *Journal of Nonverbal Behavior* and on the Editorial Board of the journal *Leadership Quarterly*.

Daniel Gatica-Perez (S'01–M'02) is currently the Head of the Social Computing Group, Idiap, Martigny, Switzerland, and a Professeur Titulaire in the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. His research interests include social computing, social media, ubiquitous computing, and crowdsourcing. He was an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA.