ISSN 2395-8618

Urban Perception: Can We Understand Why a Street is Safe?

Felipe Moreno-Vera, Bahram Lavi, and Jorge Poco

***Abstract*—The importance of urban perception computing is relatively growing in machine learning, particularly in related areas to Urban Planning and Urban Computing. This field of study focuses on developing systems to analyze and map discriminant characteristics that might directly impact the city's perception. In other words, it seeks to identify and extract discriminant components to define the behavior of a city's perception. This work will perform a street-level analysis to understand safety perception based on the "visual components". As our result, we present our experimental evaluation regarding the influence and impact of those visual components on the safety criteria and further discuss how to properly choose confidence on safe or unsafe measures concerning the perceptional scores on the city street levels analysis.**

***Index Terms*—Urban perception, urban computing, interpretability, LIME computer vision, perception computing, deep learning, street-level imagery, visual processing, street view, cityscape, ADE20K, place pulse, perception learning, segmentation.**

I. Introduction

"Cities are designed to shape and influence the lives of their inhabitants" [1]. Various studies have shown that the visual appearance of cities plays a key role in human perception that could cause variant reactions (e.g., abnormality) in the city's environments, such as "The image of the city" [2]. A notable example is the Broken Window Theory [3] which delivers that visual signs of environmental disruption, such as broken windows, abandoned cars, trash, and graffiti, can induce social outcomes like an increase in crime levels.

This theory has greatly influenced on public policy makers that lead to aggressive police tactics to control the manifestations of social and physical disorders. For example, in social experiments and studies on the perceived quality of life in the streets of New York, [4] reports the high correlation between graffiti or garbage presence and dangerous places. On the other hand, clean places present high correlation with safety places. Similar results were reported by [5], [6], [2] concluding that in places where "the rules are violated", none of the rules will be fulfilled in that place negatively influenced by the environment (e.g., graffiti, garbage). In addition, other studies have shown that the visual aspect of environments of a city affects the psychological state of its inhabitants [1], [7]. Other studies show that the impact of green areas in urban cities has a positives relation to safety perception [8], [9], [10].

In this study, we present a methodology to analyze the influence of objects on a street image by taking into account their corresponding perceptual scores. Furthermore, we investigate machine learning techniques to alleviate the relationship between urban visual components and their perceptions score.

This paper is organized as follows: Section II explains the related works; Section III introduces our methodology in perception score analysis; Section IV presents our experiments and discussion about our achieved results by providing some signs over the limitations on this research field; and finally, Section V concludes our work.

Manuscript received 15/06/2019, accepted for publication on 28/08/2019.

F. Molino Vera is with the Universidad Católica San Pablo, Peru (felipe.moreno@ucsp.edu.pe).

B. Lavi and J. Poco are with the Fundação Getulio Vargas, Brazil ({bahram.lavi, jorge.poco}@fgv.br).

II. Related Works

Previous works have difficulty explaining the direct relation between the visual appearance of a city and its corresponding non-visual attributes. Therefore, these works focused on finding the relation between the data from police records and census statistics (e.g., robbery rate, house prices, population density, graffiti existence) with the the visual appearance of a city area. In the following, we will highlight and discuss them in details.

A. Urban Perception

Some studies have addressed urban perception analysis by examining different methods in computation and extracting knowledge from various resources (visual and non-visual components) – aiming to seek a proper correlation among them. The work proposed in [11] attempts to address the key question on the appearance of the Paris, "What makes Paris look like Paris?". The work was developed to compare, differentiate, and correlate the visual representation between 12 cities. Similarly, the work in [12] addressed another proposal as "What Makes London Look Beautiful, Quiet, and Happy?", which explores nearly to 700,000 street images through an online web survey.

In [13], the work studied the correlation between non-visual-attributes from the city along with its visual appearance using some datasets containing the data from crimes statistics, robbery rate, house pricing rate, population density, graffiti presence, and a perception survey.

ISSN 2395-8618

In addition, MIT Media Lab releases the PlacePulse dataset [14] which is composed of images over different streets from many capital cities like New York, Boston, Linz, and Salzburg. They also provided the associated perceptual scores for each of the images.

This work was born from the attempt to relate people's perception of a street through an online survey. This dataset conducted new studies for the problems like urban mapping [15] which performs as a classification/regression task to compare the performance of features extractors like Gist, SIFT+Fisher Vectors, and DeCAF [16]. In [17], a StreetScore approach proposed to compare a set of low-level features such as GIST, Geometric Probability Map, Text on Histograms, Color Histograms, Geometric Color Histograms , HOG 2x2, Dense SIFT, LBP , Sparse SIFT histograms, and SSIM features extractors doing a similar research on urban perception analysis. Following their methodology, a similar study was performed over the city of Bogotá, Colombia [18].

In summary, these works have difficulty in extracting information about the natural image because they use traditional image representations including Hog+Color descriptor, Locality-Sensitive Hashing, Gist, HOG+color [13], SIFT Fisher Vectors, DeCAF features [15], geometric classification map, color Histograms, HOG2x2, and Dense SIFT [17]. Besides, to train those features non-linear methods are used like SVM [19], Linear Regression [15], SVR [17], RankingSVM [20], Multi Task Learning [21], Transfer Learning based models and pre-trained networks in [10,18,22,23,24,25,26].

B. Model Interpretation

Model interpretation methods allow us to get insights and understand the behavior of the learning model in its training phase. In the line with those methods, there are several works whose purpose is to understand and explain predictions. Previous works such as LIME [22], SHAP [23], and Anchor [24] explain a model based on their local and global level feature components.

Other approach based on gradient attribution methods used to generate feature maps of an input to provide a visual idea about the explanation like Saliency Maps [25], Gradient [26], Integrated Gradients [27], DeepLIFT [28], Grad-CAM [29], Guided Back Propagation [30], Guided gradCAM [29], and SmoothGrad [31]. These methods can ease and assist us in explaining simple/complex models aiming to identify the dependence of variables and determine whether one of them can be isolated or not; to ascertain which one has a better representation for prediction depending on the input sample.

In this work, our primary goal is to understand the behavior of urban perception. First, we extract the objects segmented from Place Pulse images denominating our visual components. Then, we train those components, considering them as a feature vector using the standard classifiers: SVC model with RBF and Linear kernels, Logistic Regression, and Ridge Classifier. Finally, we aim to understand the impact of the visual component representations by adopting the LIME (black-box) method to analyze the behavior of the predictions.

TABLE I
DATA SUMMARY ABOUT PLACE PULSE 1.0 AND THEIR RESPECTIVE CATEGORY MEAN

Place Pulse 1.0				
City	# images	*safe mean*	*wealth mean*	*unique mean*
Linz	650	4.85	5.01	4.83
Boston	1237	4.93	4.97	4.76
New York	1705	4.47	4.31	4.46
Salzburg	544	4.75	4.89	5.04
Total	4136			

III. METHODOLOGY

In this section, we describe our methodology in urban perception analysis. Our main goal is to analyze the key components that mainly affect a security perception, such as safety. To this end, we first explain our utilized urban datasets, mainly they contain the perceptional scores for the safety criteria. Then we explain a possible solution for learning the key components to understand the importance of their presence in the images, which they have assigned as safe or unsafe.

A. Datasets and Data Pre-Processing

PlacePulse has two versions. The first one is Placepulse 1.0 which composed by a set of street views images and provides their corresponding perceptual scores [14]. At the end of 2013, Place Pulse 1.0 was organized with a total of 73,806 comparisons of 4,109 images from 4 cities: New York City (including Manhattan and parts of Queens, Brooklyn and The Bronx), Boston (including parts of Cambridge), Linz and Salzburg of two countries (US and Austria) and three types of comparisons: *safe*, *wealth*, y *unique*. This dataset has been pre-processed for quick use, containing information on the position of each image (latitude and longitude), perception score for each category, an image identifier and the city to which said image belongs.

The second dataset is PlacePulse 2.0 [32] that contains a set of comparisons between image pairs, and include the latitude and longitude points for each. In addition, each comparison has the respective winner (or draw). In 2016, Place Pulse 2.0 already contained around 1.22 million comparisons of 111,390 images of 56 cities in 28 countries across the 5 continents and six types of comparisons: *safe*, *wealth*, *depress*, *beautiful*, *boring*, and *lively*. This dataset contain 8 columns: image ID (left and right), latitude and longitude (of each image), the result of the comparison, and the respective evaluated category.

We perform the method proposed by [33] to pre-process all comparisons in the dataset: for each compared image i with other images j many times in different categories, we define the intensity of perception of any image i as the percentage of times that the image was selected. Besides, the intensity of j affects i intensity. Due to this, we define the positive rate W_i

ISSN 2395-8618

(1) and the negative rate L_i (2) of an image i corresponding to a specific category:

$$W_i = \frac{w_i}{w_i + d_i + l_i}, \tag{1}$$

$$L_i = \frac{l_i}{w_i + d_i + l_i}, \tag{2}$$

where, w_i is the number of wins, l_i number of loses, and d_i draws; From the equations 1 and 2 we can calculate the perceptual score associated for each an image i called *Q-score* with notation $q_{i,k}$ in a category k:

$$q_{i,k} = \frac{10}{3}(W_{i,k} + \frac{1}{n^w_{i,k}}(\sum_{j_1} W_{j_1,k}) - \frac{1}{n^l_{i,k}}(\sum_{j_2} L_{j_2,k}) + 1). \tag{3}$$

The Equation 3 is the perceptual score of the image i to be ranked, where j is an image compared to i, n^w_i is equal to the total number of images i beat and n^l_i is equal to the total number of images to which i lost. Besides, j_1 is the set of images that loses against the image i and j_2 is the set of images that wins against the image i. Finally, Q is normalized to fit the range 0 to 10; this scale is a standard measurement whereby one can evaluate the perceptions [34]. In this score, 10 represents the highest possible score for a given question. For example, if an image receives a calculated score of 0 for the question "Which place looks safer?" indicating that specific image is perceived as the least safe image in the dataset.

TABLE II
STATISTICS OBTAINED AFTER PROCESS ALL COMPARISONS FROM PLACE PULSE, CONTAINING INFORMATION ABOUT IMAGES PER CITY IN EACH CONTINENT AND THE MEAN SCORE FOR EACH REQUESTED CATEGORY

Place Pulse 2.0		
Continent	# cities	# images
America	22	50,028
Europe	22	38,747
Asia	7	11,417
Oceania	2	6,097
Africa	3	5,101
Total	56	111,390

(a)

Place Pulse 2.0		
Category	# comparisons	*mean*
Safety	368,926	5.188
Lively	267,292	5.085
Beautiful	175,361	4.920
Wealthy	152,241	4.890
Depressing	132,467	4.816
Boring	127,362	4.810
Total	1,223,649	

(b)

B. Visual Components Extraction

In this work, we will focus on Boston city for our experiments. We use two segmentation Network: (i) PSPNet [35] and (ii) DeepLabV3+ [36]. We define as "visual components" the object pixel presence extracted from (i) and (ii) per image of the city. As we show in Figure 1, our main idea is to transform the percentage of objects present in each image. As we know, different images could present different % pixel ratio segmented (see Figure 1 (a), (b), (c), and (d), we perform a Standardization of all features obtained by segmentation. We perform both extractions to compare these methods using the PSPNet as the baseline.

Both object segmentation extractors were trained in ADE20K [37], using networks like ResNet101 and Xception as backbone, respectively. We prefer to use the ADE20K dataset pre-trained weights instead of Pascal, COCO, or CityScapes [38], due to the number of classes between the datasets. ADE20K present 150 classes and a hierarchical tree of indoor-outdoor classes, CityScapes provides 50 classes (most of them contained in ADE20K), COCO 91 classes, and Pascal-VOC 20 classes.

Next, after extract our features, we train them using a Support Vector Classifier with 10 KFold cross-validations varying our regularization parameter l_2. To perform our classification task to predict is a street image is safe or not safe. In order to classify, we need to select subsets from each city dataset. To do this, we define a parameter called δ with a value between 0,05 - 0,5. This delta will create a subset using the binary labels $y_{i,k} \in \{1, -1\}$ for both training and testing as:

$$y_{i,k} = \begin{cases} 1 & \text{if } (q_{i,k}) \text{ in the top } \delta\%, \\ \text{-1} & \text{if } (q_{i,k}) \text{ in the bottom } \delta\%. \end{cases} \tag{4}$$

Since we know from previous works results [15], [14], [39], we focus our study on the worst case reported, corresponding to $\delta = 0.5$ using all labels divided into safe and not safe.

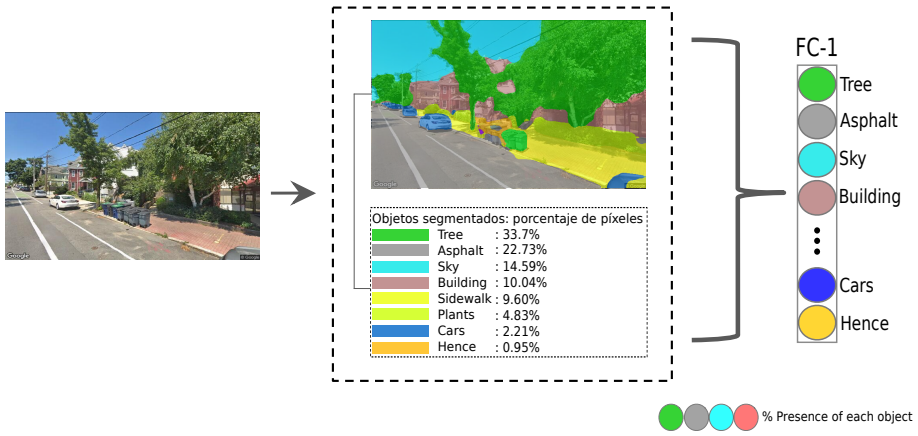
IV. EXPERIMENTS AND DISCUSSIONS

This work presents a methodology to learn and explain which features have more impact on the prediction of safe or not safe categories. We perform our experiments in Boston city (1327 images) from Place Pulse 2.0 dataset. We focus only in the safety perception due to the larger number of image compared in that category per city (see Table II (b)). We extract our visual components using both methods as mentioned above (PSP-Net, and DeepLabV3+). In order to visualize our distribution of visual components, we perform a mini-process in our pixel ratio extracted for each feature $X_{i,k}$:

$$X_{i,k} = \begin{cases} 1 & \text{if object } (k) \text{ is present in image } X_i, \\ 0 & \text{if object } (k) \text{ is not present in image } X_i. \end{cases} \tag{5}$$

Then, sum by column and divided by the total number of images, we got an object distribution presence in the whole Boston city for each method (see Figure 2). Then, we train both visual components extracted by both methods (PSP-Net and DeepLabV3), we can see that both features yielded a poor performance on the dataset (see Table IV).

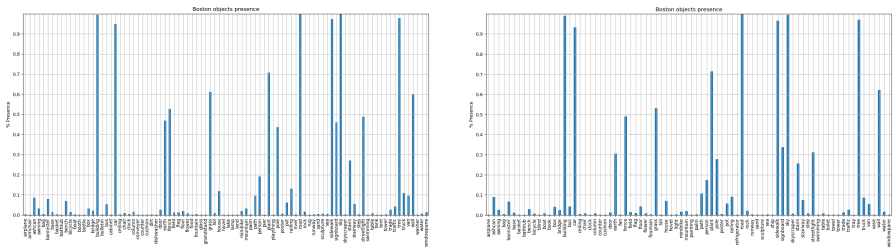
ISSN 2395-8618



(a) PoV 1

(b) PoV 1 (c) PoV 2 (d) DeepLabV3+ (e) PSPNet

Fig. 1. (a) Input image and output features (based on the object pixel ratio segmented). (a) y (b): Different Point of View of the images, each object presence will depends of the image evaluated. (c) y (d): Different pixel ratios extracted by DeepLabV3+ and PSP-Net. As you can see in (d) DeepLabV3+ detects the street light, but failed to detect by PSPNet



(a) DeepLabV3+ (b) PSPNet

Fig. 2. (a) DeepLabV3+ object distribution presence. (b) PSPNet Object distribution presence. As we can see, there are objects detected by only DeepLabV3 method

Model Explanation

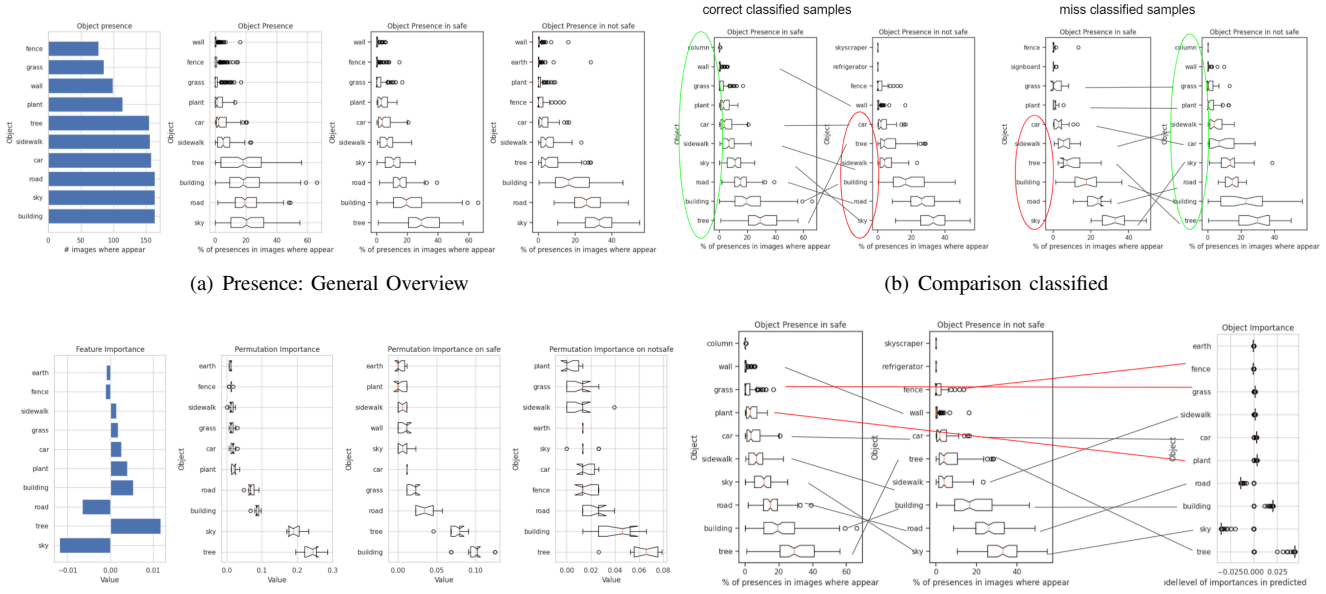
In this work, we want to understand why our street images are predicted as "safe" or "not safe". To do this, we use the black-box model explainer LIME: Local Interpretable Model-agnostic technique. LIME explains a black-box model by simulating local candidates close to the original prediction. By using these prediction outcomes, LIME generates a random distribution set of possible predictions based on L_2 distance called "local fidelity" taken as a reference to the original prediction.

First, based on the results of training using different models like Logistic Regression, Ridge Classifier, SVC with kernel Linear and RBF presented in the Table IV. Then, we choose the Logistic Regression model to analyze the feature importance and influence. We started analyzing the object presence in both subsets divided by safe and not safe categories.

Second, we analyze the Feature Presence in whole images and the respective subsets corresponding to safety and not safety. At this step, we divided in two sub-subsets: miss-classified and correct classified subsets (see Figure 3). Then, we perform the following Feature Importance, Permutation Importance in the whole dataset and the subsets divided by safe and not safe categories. In Figure 3-(a), we show the global and divided object presence, we can see that the first four objects in safety are the reverse of not safety.

In Figure 3-(b), we note that object presence in miss-classified samples correspond more to the opposite class (e.g., miss-classified as safe that has presented like not safe), these regions are highlighted in red and green: red correspond to not safe class, which is present in miss-classified

ISSN 2395-8618



(a) Presence: General Overview

(b) Comparison classified

(c) Permutations: Object Influence

(d) LIME: Object Importance

Fig. 3. (a) General object presence in the whole dataset divided by safe and not safe categories. (b) sub-subsets of correct/miss classified. (c) Object Influence after permutations on features. (d) Object importance calculated by LIME, red lines mean the match between presence in safety/not safety and the importance in predictions

TABLE III

WE REPORT TEST CLASSIFICATION FOR $\delta = 0,5$ (WORST CASE) FOR EACH FEATURE EXTRACTOR METHOD. WE NOTE THAT DEEPLAB AND PSPNET HAVE A POOR PERFORMANCE IN THIS TASK, BUT DEEPLAB PRESENT A BETTER LEARNING PROCESS

Features	Metric	RBF-SVC	Linear SVC	Logistic Regression	Ridge
PSPNet	*AUC*	0.46465	0.47036	0.465	0.48551
	ACC	0.44516	0.48065	0.47097	0.48387
	F1	0.42282	0.5752	0.50602	0.47712
DeepLabV3	*AUC*	**0.55553**	0.51255	0.51895	**0.56066**
	ACC	0.5129	0.50323	**0.52258**	0.51935
	F1	0.47018	**0.59043**	0.53459	0.52396

safe images, the same happens with miss-classified not safety highlighted in green. In Figure 3 (c) we present the object influence, this report obtained by the methods mentioned above.

The results show the Influence of permute features or visual components and the relevance in the predictions, as the component 'tree'. The feature of 'tree' component has higher permutation importance in not safe class.

Our last step is to use the model explainer LIME to determine the object importance of predictions. This step shows the main results of LIME. The most highlighted ones are tree, sky, building, and road which are the most discriminant features to the predictions method in comparison with cars and sidewalk. This can be explained due to the high frequently appearance of these objects in all the images. Besides, objects like grass, plants, earth, and fence are very related to a particular category. In safety class, we have grass and plants, while in not safe class, it observes more with earth and fence components.

Limitations: We found three main limitations in this work. The first one is about the Place Pulse dataset that constructed using an online survey. Each volunteer chose between two images that are the most "safe" depending on their biased personal perception criteria. The second limitation is the small number of sample images per city. Comparing with other dataset with millions of samples, in total is not above of 100,000 that yields the method to have week performance when a lower number of data samples are available. The last limitation is the impracticality of creating a general city perceptual predictor due to the significant difference between cities and their unique visual appearance.

V. CONCLUSIONS

In this work, we propose a methodology that allows us to understand the behavior of the urban safety perception on street view images. To this end, we pre-processed the dataset Place Pulse 2.0, analyzing the 110 thousand images obtained by comparisons and calculated their corresponding perception scores in six different categories. In our study, we focused on Boston city with its safety scores. We investigated and analyzed which visual components are impacting positively and negatively in the predictions. To

ISSN 2395-8618

understand the predictions, we used LIME to determine the importance of feature components. From the result, we conclude that our model is capable to predict the safety perception from street view images. In addition, we showed the correlation between higher safety perception with the presence of trees or green areas, skies, and roads.

REFERENCES

[1] P. J. Lindal and T. Hartig, "Architectural variation, building height, and the restorative quality of urban residential streetscapes," *Journal of Environmental Psychology*, vol. 33, pp. 26–36, 2013.

[2] K. Lynch, "Reconsidering the image of the city," in *Cities of the Mind*. Springer, 1984, pp. 151–161.

[3] J. Q. Wilson and G. L. Kelling, "Broken windows," *Atlantic monthly*, vol. 249, no. 3, pp. 29–38, 1982.

[4] K. Keizer, S. Lindenberg, and L. Steg, "The spreading of disorder," *Science*, vol. 322, no. 5908, pp. 1681–1685, 2008.

[5] H. W. Schroeder and L. M. Anderson, "Perception of personal safety in urban recreation sites," *Journal of leisure research*, vol. 16, no. 2, pp. 178–194, 1984.

[6] E. K. Tokuda, C. T. Silva, and R. M. C. Jr., "Quantifying the presence of graffiti in urban environments," *CoRR*, vol. abs/1904.04336, 2019. [Online]. Available: http://arxiv.org/abs/1904.04336

[7] R. Kaplan and S. Kaplan, *The experience of nature: A psychological perspective*. Cambridge university press, 1989.

[8] R. S. Ulrich, "Visual landscapes and psychological well-being," *Landscape research*, vol. 4, no. 1, pp. 17–23, 1979.

[9] R. J. Sampson, J. D. Morenoff, and T. Gannon-Rowley, "Assessing "neighborhood effects": Social processes and new directions in research," *Annual review of sociology*, vol. 28, no. 1, pp. 443–478, 2002.

[10] X. Li, C. Zhang, W. Li, R. Ricard, Q. Meng, and W. Zhang, "Assessing street-level urban greenery using google street view and a modified green view index," *Urban Forestry & Urban Greening*, vol. 14, no. 3, pp. 675–685, 2015.

[11] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros, "What makes paris look like paris?" 2012.

[12] D. Quercia, N. K. O'Hare, and H. Cramer, "Aesthetic capital: what makes london look beautiful, quiet, and happy?" in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014.

[13] S. M. Arietta, A. A. Efros, R. Ramamoorthi, and M. Agrawala, "City forensics: Using visual elements to predict non-visual city attributes," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2624–2633, 2014.

[14] M. P. Salesses, "Place Pulse: Measuring the collaborative image of the city," Ph.D. dissertation, Massachusetts Institute of Technology, 2012.

[15] V. Ordonez and T. L. Berg, "Learning high-level judgments of urban perception," *European Conference on Computer Vision (ECCV)*, 2014.

[16] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning*, 2014, pp. 647–655.

[17] N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo, "StreetScore: predicting the perceived safety of one million streetscapes." *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

[18] S. F. Acosta and J. E. Camargo, "Predicting city safety perception based on visual image content," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2018, pp. 177–185.

[19] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[20] L. Porzi, S. Rota Bulò, B. Lepri, and E. Ricci, "Predicting and understanding urban perception with convolutional neural networks," 10 2015.

[21] W. Min, S. Mei, L. Liu, Y. Wang, and S. Jiang, "Multi-task deep relative attribute learning for visual urban perception," *IEEE Transactions on Image Processing*, vol. 29, pp. 657–669, 2019.

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[25] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013.

[26] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," 2016.

[27] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," 2017.

[28] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *arXiv preprint arXiv:1704.02685*, 2017.

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

[30] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[31] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "A unified view of gradient-based attribution methods for deep neural networks." ETH Zurich, 2017.

[32] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo, "Deep learning the city : Quantifying urban perception at A global scale," *CoRR*, 2016.

[33] P. Salesses, K. Schechtner, and C. A. Hidalgo, "The collaborative image of the city: Mapping the inequality of urban perception," *PLOS ONE*, 2013.

[34] J. L. Nasar, "The evaluative image of the city," 1998.

[35] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[36] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[37] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal on Computer Vision*, 2018.

[38] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[39] F. Moreno-Vera, "Understanding safety based on urban perception," in *International Conference on Intelligent Computing*. Springer, 2021, pp. 54–64.