Understanding Safety Based on Urban Perception

Felipe Moreno-Vera$^{1[0000-0002-2477-9624]}$

Universidad Católica San Pablo;
felipe.moreno@ucsp.edu.pe

Abstract. Currently, one important field on machine learning is Urban Perception Computing is to model the way in which humans can interact and understand the environment that surrounds them. This process is performed using convolutional models to learn and identify some insights which define the concept of perception of a place (e.g. a street image). One approach of this field is urban perception of street images, we will focus on this approach to study the safety perception of a city and try to explain why and how the perception can be predicted by a mathematical model. As result, we present an analysis about the influence and impact of the visual components on the safety criteria and also an explanation about why a certain decision on the perception of the safety of the streets, such as safe or unsafe.

Keywords: Urban Perception · Urban Computing · Interpretability · LIME · Computer Vision · Perception Computing · Deep Learning · Street-level imagery · Visual processing · Street View · Cityscape · Perception Learning· Grad-CAM.

1 Introduction

"Cities are designed to shape and influence the lives of their inhabitants" [13]. Various studies have shown that the visual appearance of cities plays a central role in human perception and reaction to said environment such as "The image of the city" [15]. A notable example is the Broken Window Theory [39] which suggests that visual signs of environmental disruption, such as broken windows, abandoned cars, trash, and graffiti, can induce social outcomes like increase crime levels. This theory has had a great influence on public policy strategies that lead to aggressive police tactics to control the manifestations of social and physical disorder. For example, in social experiments and studies on the perceived quality of life in the streets of New York, comparing impeccable places such as shopping malls (clean walls, orderly, quiet) with other places in which graffiti or garbage is presented [10, 28, 37, 15] concluding that in places where "the rules are violated" it means that in the long term, 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 the spaces of a city affect the psychological state of its inhabitants [13, 9]; Other studies show that the impact of green areas in urban cities [38, 27] has a positives relation to safety perception. In this study, we present a deep learning-based methodology and a model to predict and understand human perceptions of the physical setting of a place. The approach is able to predict, understand and explain predictions of the security perception accurately for a new urban region. Second, we studied the relationship between urban visual elements and perceptions, and tried to determine "the importance of visual elements and their influence over a specific perception". This result helps urban planners and researchers to understand the positive or negative impact of various visual components by exploring urban patterns. The present document is organized by the following: section II is about Related works; section III we present our Methodology; section IV Discussions and results; and finally, Conclusions of this work. Our main contributions is a methodology to train and explain urban perceptions from street-images.

2 Related Works

Previous works have a difficulty to explain the direct relation between visual appearance of a city and their corresponding non-visual attributes. These works made an study focus on find a relation between datasets like reports like crimes statistic, robbery rate, house prices, population density, graffiti presence (local reports), and a danger perception survey; with visual appearance of one city.

2.1 Urban Perception

There is a selected works based on urban perception and how to determine using computational methods. The main goal of these works is to correlate a visual appearance of a city with their non-visual attributes like crimes, house prices, perception surveys, etc. These works are solving questions like "What makes Paris look like Paris?" [6] to compare, differentiate and correlate the visual appearance (features) between 12 cities. A similar approach was proposed to answer "What Makes London Look Beautiful, Quiet, and Happy?" [22] exploring 700,000 street-images through a online web survey. [4] studied the correlation between visual non-attributes from city and their visual appearance using several dataset like crimes statistic, robbery rate, house prices, population density, graffiti presence (local reports), and a danger perception survey.

In addition, MIT Media Lab releases the PlacePulse dataset [25] which is compose by a street images from difference main cities like New York, Boston, Linz, and Salzburg; and a corresponding perceptual score associated. This work was born from the attempt to relate people's perception of a street through an online survey. This dataset conduced new studies like urban mapping [20] which performs a classification/regression task using and comparing the performance of features extractors like Gist, SIFT + Fisher Vectors, and DeCAF [7]. StreetScore [17] compares 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. Following this methodology, a similar study was performed over the city Bogotá, Colombia called Wmodi [1].

In summary, these works have difficulty in extracting information about the natural image because they use classical image features including Hog+Color descriptor, Locality-Sensitive Hashing, Gist, [4], SIFT Fisher Vectors, DeCAF features [20], geometric classification map, color Histograms, HOG2x2, and Dense SIFT [17]. Other works use non-linear models to predict images like SVM [5] and Linear Regression [20], Support Vector Regression was used in [17], RankingSVM was used in [21], SVR was implemented in [4], Multi Task Learning, Transfer Learning based models on ImageNet, and pre-trained networks in [18, 8, 40, 16, 11, 1].

2.2 Model Interpretation

Model interpretation methods helps us to get insights and understand our learning process and the behavior of a model. In Interpretable Machine Learning, there are several works whose purpose is to understand and explain predictions. Usually, models like CNN are called "black-box" due to they have a large number of parameters distributed in hidden layers with unknown information shared through each layer. Previous works such as LIME [23], SHAP [14], and Anchor [24] explain a model based on their local and global level features components. Other approach based on gradient attribution methods to generate feature maps of an input to provide a visual idea about the explanation like Saliency Maps [32], Gradient [31], Integrated Gradients [36], DeepLIFT [30], CAM [41], Grad-CAM [29], Guided Back Propagation [35], Guided gradCAM[29], and SmoothGrad [3]. These methods are useful to explain simple or complex black box models identifying the dependence of variables and determine if one of them can be isolated or not, in addition to which one has a better representation for prediction depending on the input type.

In this work, our approach is to understand the behavior of the urban perception trained on a convolutional network based model using the PlacePulse dataset, composed by images and associate perceptual scores. We want to understand which features impact positive and negative in the perception of safety in street images.

3 Methodology

Our methodology was divided into three parts: (i) dataset pre-processing, (ii) Model training and evaluation, and (iii) Model Interpretation.

3.1 Dataset

PlacePulse has two versions, the first one is Placepulse 1.0 is a dataset composed by a set of images and their correspond perceptual scores. The second one, PlacePulse 2.0 [8] is a dataset composed of a set of comparisons between 2 images, containing the latitude and longitude for each image. In addition, each comparison has the respective winner (or draw).

Place Pulse 1.0 At the end of 2013, Place Pulse 1.0 contains 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.

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			

Table 1. Data summary about Place Pulse 1.0 and their respective category mean.

Place Pulse 2.0 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 an algorithm proposed by [26] to pre-process all comparisons in the dataset: for each compared image i with other images j many times in different categories, we define as 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 (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 when you evaluate perceptions [19]. In this scores, 10 represents the highest possible score for a given question. As an example, if an image receives a calculated score of 0 for the question "Which place looks safer?" that means that specific image is perceived as the least safe image in the dataset.

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)

Table 2. Statistics obtained after process all comparisons from Place Pulse, containing information about images per cities in each continent and the mean score for each category asked.

3.2 Experiments

In this work, we adapted the VGG16 [33] architecture and our adapted a modification to GAP [12] called VGG16-GAP, we modify the last layer of the block-conv5, taking the Max-Pooling layer and replacing for a GAP layer. This modification aims to extract more informative and high-level features from input images through Global Average Pooling. Once we extract features with this architecture presented in Figure 1. Then we remove last 2 Fully Connected layer from original model architecture after layer 13, we call the output of this layer as the features extracted from VGG-GAP. We train our Place Pulse dataset focus on two main cities: Boston and New York with perceptual metric safety (we study other metric as well). We select both cities because these cities have the most images quantity and comparisons between images.

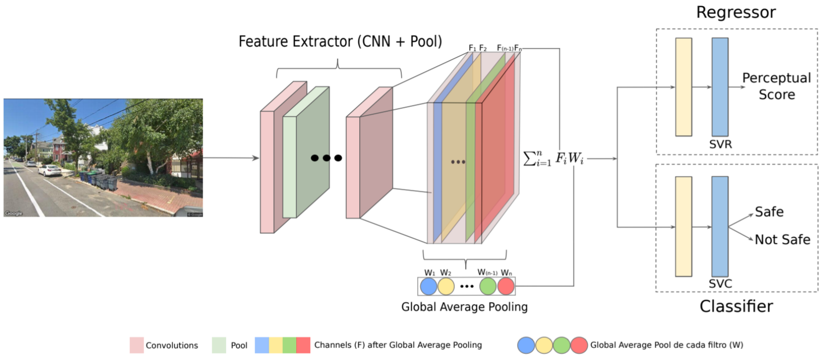

Regressor
Feature Extractor (CNN + Pool)
F_1 F_2 $F_{(n-1)}$ F_n
Perceptual Score
SVR
$\sum_{i=1}^{n} F_i W_i$
Safe
Not Safe
W_1 W_2 $W_{(n-1)}$ W_n
SVC
Global Average Pooling
Classifier
Convolutions
Pool
Channels (F) after Global Average Pooling
Global Average Pool de cada filtro (W)

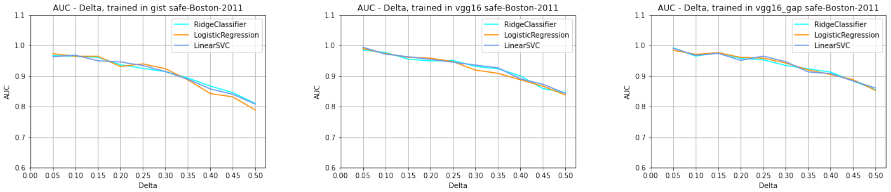
Fig. 1. Our adaptation of VGG16-GAP for both tasks.

We will focus our study comparing three main feature extractors: VGG16, VGG16-GAP, and GIST. Due to the reported results in several preivous works in street images described above, we select GIST as baseline feature extractor. To train our VGG16-GAP model, we used a Transfer Learning fine-tuned strategy in the classification and regression task. We use the pre-trained weights of ImageNet dataset, which contains millions of images across 1,000 classes. Then, we freeze our 10 first layers (from input until block4conv3), training only the Block-Conv5. To train the dataset, we make two experiments. The first one is the classification task: To perform this task we divided our images into two labels for each category in the dataset: e.g. safe and not safe. To select which set of images will be safe or unsafe, we define a parameter called δ with a value between 0,05 - 0,5. This delta will creates a subset using the binary labels $y_{i,k} \in \{1, -1\}$ for both training and testing as:

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

We parameterize the classification problem by a variable δ and calculate performance as we adjust δ. As we move the value of our parameter δ the problem becomes more difficult since the visual appearance of the positive and negative images starts to become less evident up to the point when $\delta =$ 0,5. At the same time when δ has smaller values the positive and negative images are easier to classify but we have access to less data.

We learn models to predict $y_{i,k}$ from input image representations x_i using the following methods to extract features: VGG16 based-line, VGG16-GAP, and GIST. We train and compare the behavior of linear and non-linear models regularized with l_2 like *Logistic Regression*: $L(y, f(x)) = \sum_{i=1}^{n} log(e^{(-y_i * f(x_i))} + 1)$, *Linear SVC*: $L(y, f(x)) = \sum_{i=1}^{n} max(0, 1-y_i f(x_i))$, *Ridge Classifier*: $L(y, f(x)) = sgn(||y - f(x)||_2^2 + ||w||_2^2)$, VGG16-Softmax, and VGG16-GAP -Softmax both with loss function $L(y, f(x)) = \frac{e^{(y_i - f(x_i))}}{\sum_{k=1}^{n} e^{(y_k - f(x_k))}} + ||w||_2^2$.



(a) Gist (b) VGG16 (c) VGG16-GAP

Fig. 2. Test classification results: 10 KFolds cross-validation Avg AUC over the city Boston trained using GIST, VGG16, and VGG16-GAP as feature extractor. VGG16-GAP achieves higher metric values than Gist and VGG16 along the different values of our parameter δ.

We evaluated our binary classifier model behavior using the Area Under the Curve (AUC) metric which depends on Precision-Recall as we report in Figure 2. We use a regularization l_2 to avoid overfit our model. We set the regularization parameter C using held-out data and learn w_k using training data $\{x_i, y_{i,k}\}$.

		Test Boston			Test New York		
Training	Métrica	*VGG16-GAP*	*VGG16*	*Gist*	*VGG16-GAP*	*VGG16*	*Gist*
Boston	*safety*	**71.428**	70.322	71.064	**67.741**	59.354	64.721
	wealthy	**67.741**	63.88	66.334	**65.897**	64.183	61.458
	uniquely	**63.354**	61.935	62.486	63.773	**63.858**	63.564
New York	*safety*	**66.459**	65.512	65.842	**66.968**	64.741	66.874
	wealthy	**64.748**	62.111	63.265	**63.8032**	60.001	62.997
	uniquely	**68.322**	64.748	66.349	62.895	62.468	**62.968**

Table 3. We report Test classification for δ=0,5 in two scenarios: a) training and testing perceptual prediction models on images from the same city, and b) training models on images from one city and testing on images from another city. We present that our VGG16-GAP has a better performance except in the perceptual category uniquely.

The second one is the regression task: To perform this task we divided our images in the same way as we divided before. In this case, we want to predict not the category but the perceptual score associate with an image which we calculated before (Equation 3). Here, our ground truth labels are $y_{i,k} = q_{i,k}$

for image i and perceptual measure k. Therefore, we make predictions $\hat{y}_{i,k}$ as a linear combination of the extracted features x_i corresponding to image i as follows:

$$\hat{y}_{i,k} = q_{i,k} \tag{5}$$

To perform our experiments, we train our set $(x_i, y_{i,k})$ using linear and non-linear methods regularized with l_2 like: *Ridge*: $L(y, f(x)) = ||y - f(x)||_2^2 + ||w||_2^2$, *Lasso*: $L(y, f(x)) = \frac{1}{2*n}||y - f(x)||_2^2 + ||w||_1$, *Linear SVR*: $L(y, f(x)) = \sum_{i=1}^{n} max(0, |y_i - f(x_i)|)$, and a simple *Linear Regression*: $L(y, f(x)) = ||y - f(x)||_2^2$. We choose the Pearson coefficient as a metric of regression models, we select this metric because we want to achieve a high correlation between extracted features from images with their correspond $y_{i,k}$:

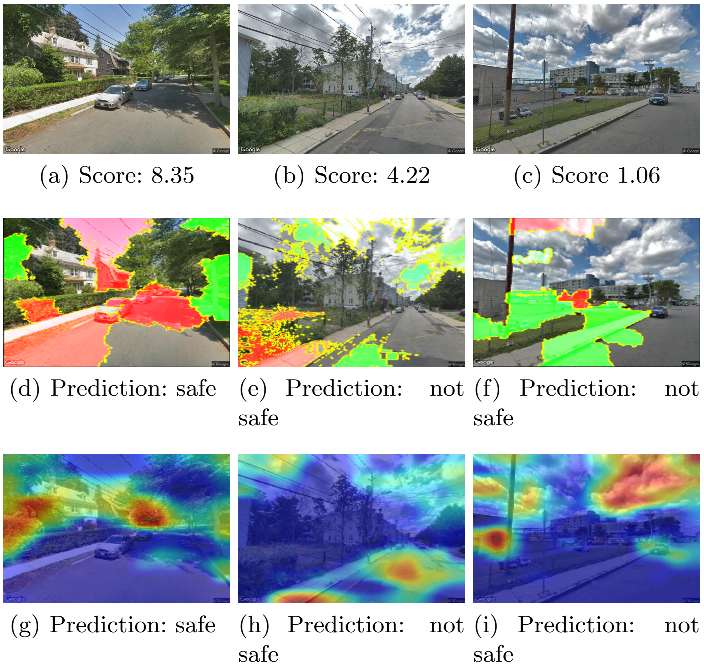
		Methods			
City	Feature Extractor	LinearSVR	LinearRegress	Lasso	Ridge
Boston	VGG16-GAP	**0.7095**	0.5717	**0.71342**	**0.7462**
	VGG16	0.6832	**0.6354**	0.70001	0.7163
	GIST	0.66643	0.41612	0.6569	0.6658
New York	VGG16-GAP	0.5649	**0.6196**	**0.6503**	**0.7209**
	VGG16	**0.6062**	0.60487	0.64531	0.70986
	GIST	0.59157	0.5734	0.61991	0.68732

Table 4. We report Test regression task trained with SVR-l_2. We note that regression task has a best behavior over Boston and VGG16-GAP provides the best results in both cities.

3.3 Model Explanation

In this work, we want to understand why our street images that are predicted as "safe" or "not safe". To do this, we compare two explainers: The first one is LIME, a local interpretable model-agnostic technique. LIME explains a black-box model by simulating local candidates close to the original prediction. Using these predictions, LIME generates a random distribution set of possible predictions based on L_2 distance called "local fidelity" taken as reference the original prediction. Then, LIME select which possibles random noises could be a good samples to evaluate using its Submodular Pick Algorithm (SP-LME) trained by a SVM.

The second one is Grad-CAM, this method presents a strong behavior in interpret convolutional networks [2]. In this work, a comparison of robustness against adversial attack was performed. This work shows that Grad-CAM is strong against adversarial attacks, unlike CAM [41], Gradient Input [31], Integrated Gradients [36], GBP [35], Smoothed Gradients [34], Grad-CAM and



(a) Score: 8.35 (b) Score: 4.22 (c) Score 1.06

(d) Prediction: safe (e) Prediction: not safe (f) Prediction: not safe

(g) Prediction: safe (h) Prediction: not safe (i) Prediction: not safe

Fig. 3. Images from Boston (first row) with their respective predicted scores and class predicted. Besides, we present LIME outputs (second row) in which green regions mean positive and red ones mean negatives impacts of the features of a prediction. Furthermore, Grad-CAM results (third row) only shows the highlighted positive regions with more importance for the prediction.

Guided Grad-CAM [29], and DeepLIFT [30]. As we can see on Figure 3, both methods highlight different regions for the same prediction sample. We can easily visualize which part of an input image was learned by the model and which regions are relevant to the prediction.

4 Discussions

This work presents a methodology to teach a machine how to learn features to differentiate perceptions using the Place Pulse dataset, and explain predictions about urban perception. This work was focused on safety perception processing and calculate the safe perceptual scores of street images. We adapt the VGG16 model modifying the MaxPool for a GAP operation layer. We compare VGG16, VGG16-GAP, and GIST performance in regression and classification task varying the quantify of images depending on our parameter δ varying from 0,05 to 0,5. For evaluations, we calculate the Area Under the Curve (AUC) for classification task. For regression, we trust in the Pearson Correlation Coefficient which report the correlation between a image and their associated perceptual score (see Tables 3 and 4).

To understand our model predictions, we use and compare two model explainer like LIME and grad-CAM. With these both methods, we analyze the resulting highlighted regions about safety perception predicted per image and visualize the impact of important features as you can see in Figure 3. For unsafe predictions, Grad-CAM highlights asphalt, fence, and walls. Instead of LIME, which presents a random behavior over the regions, sometimes highlight sky, asphalt, trees, grass, cars, earth or fences. For safe predictions, Grad-CAM highlighted regions are associated with green areas (trees and grass) as well as LIME. Nonetheless, LIME has a lack of importance due to the main features which have a positive impact on safe prediction usually are shadows, clouds, or asphalt as well.

Limitations : We found three main limitations in this work. The first one is about the Place Pulse dataset that was constructed using a online survey. Here each volunteer chooses 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 which has millions of samples, our total is not above of 100,000 generating a lack of robustness when training a model with few sample data. The last limitation is the impossibility of creating a general city perceptual predictor, due to the large difference between cities and their unique visual appearance.

5 Conclusions

In this work, we propose a methodology that allows us to understand the behavior of the urban perception of safety from street images. To do this, we pre-process the dataset Place Pulse 2.0 analyzing the 110 thousand images obtained by comparisons and calculated their corresponding perception scores in six categories. We focus our study on the safety scores to analyze which parts of the images are impacting positively and negatively in the predictions. To understand this predictions, we use and compare two model explainers LIME and Grad-CAM which show us the impact of the features extracted from the image. We conclude from this work that our model is capable to predict the safety perception from street image. Besides, we show the correlation between high safety perception with the presence of trees or green areas.

References

1. Acosta, S.F., Camargo, J.E.: Predicting city safety perception based on visual image content. In: Iberoamerican Congress on Pattern Recognition. pp. 177–185. Springer (2018)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps (2018)

3. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: A unified view of gradient-based attribution methods for deep neural networks. ETH Zurich (2017)
4. Arietta, S.M., Efros, A.A., Ramamoorthi, R., Agrawala, M.: City forensics: Using visual elements to predict non-visual city attributes. IEEE transactions on visualization and computer graphics **20**(12), 2624–2633 (2014)
5. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 144–152. ACM (1992)
6. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.: What makes paris look like paris? (2012)
7. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. pp. 647–655 (2014)
8. Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A.: Deep learning the city : Quantifying urban perception at A global scale. CoRR **abs/1608.01769** (2016), http://arxiv.org/abs/1608.01769
9. Kaplan, R., Kaplan, S.: The experience of nature: A psychological perspective. Cambridge university press (1989)
10. Keizer, K., Lindenberg, S., Steg, L.: The spreading of disorder. Science (New York, N.Y.) **322**, 1681–5 (12 2008). https://doi.org/10.1126/science.1161405
11. León-Vera, L., Moreno-Vera, F.: Car monitoring system in apartments' garages by small autonomous car using deep learning. In: Annual International Symposium on Information Management and Big Data. pp. 174–181. Springer, Springer International Publishing (2018)
12. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
13. Lindal, P.J., Hartig, T.: Architectural variation, building height, and the restorative quality of urban residential streetscapes. Journal of Environmental Psychology **33**, 26–36 (2013)
14. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
15. Lynch, K.: Reconsidering the image of the city. In: Cities of the Mind, pp. 151–161. Springer (1984)
16. Moreno-Vera, F.: Performing deep recurrent double q-learning for atari games. In: 2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI). pp. 1–4 (2019). https://doi.org/10.1109/LA-CCI47412.2019.9036763
17. Naik, N., Philipoom, J., Raskar, R., Hidalgo, C.: StreetScore: predicting the perceived safety of one million streetscapes. 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (2014)
18. Naik, N., Raskar, R., Hidalgo, C.A.: Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. American Economic Review **106**(5), 128–32 (2016)
19. Nasar, J.L.: The evaluative image of the city (1998)
20. Ordonez, V., Berg, T.L.: Learning high-level judgments of urban perception. European Conference on Computer Vision (ECCV) (2014)
21. Porzi, L., Rota Bulò, S., Lepri, B., Ricci, E.: Predicting and understanding urban perception with convolutional neural networks (10 2015). https://doi.org/10.1145/2733373.2806273

22. Quercia, D., O'Hare, N.K., Cramer, H.: Aesthetic capital: what makes london look beautiful, quiet, and happy? In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. pp. 945–955. ACM (2014)
23. Ribeiro, M.T., Singh, S., Guestrin, C.: 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. pp. 1135–1144. ACM (2016)
24. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
25. Salesses, M.P.: Place Pulse: Measuring the collaborative image of the city. Ph.D. thesis, Massachusetts Institute of Technology (2012)
26. Salesses, P., Schechtner, K., Hidalgo, C.A.: The collaborative image of the city: Mapping the inequality of urban perception. PLOS ONE (2013)
27. Sampson, R.J., Morenoff, J.D., Gannon-Rowley, T.: Assessing "neighborhood effects": Social processes and new directions in research. Annual review of sociology **28**(1), 443–478 (2002)
28. Schroeder, H.W., Anderson, L.M.: Perception of personal safety in urban recreation sites. Journal of leisure research **16**(2), 178–194 (1984)
29. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
30. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. arXiv preprint arXiv:1704.02685 (2017)
31. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences (2016)
32. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR) (2014)
34. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise (2017)
35. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
36. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks (2017)
37. Tokuda, E.K., Silva, C.T., Jr., R.M.C.: Quantifying the presence of graffiti in urban environments. CoRR **abs/1904.04336** (2019), http://arxiv.org/abs/1904.04336
38. Ulrich, R.S.: Visual landscapes and psychological well-being. Landscape research **4**(1), 17–23 (1979)
39. Wilson, J.Q., Kelling, G.L.: Broken windows. Atlantic monthly **249**(3), 29–38 (1982)
40. Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H.H., Lin, H., Ratti, C.: Measuring human perceptions of a large-scale urban region using machine learning. Landscape and Urban Planning **180**, 148–160 (2018)
41. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)