What Makes a Place Feel Safe? Analyzing Street View Images to Identify Relevant Visual Elements

Felipe Moreno-Vera
Fundação Getúlio Vargas (FGV)
Rio de Janeiro, Brazil
felipe.moreno@fgv.br

Bruno Brandoli
Fundação Getúlio Vargas (FGV)
Rio de Janeiro, Brazil
bruno.brandoli@fgv.br

Jorge Poco
Fundação Getúlio Vargas (FGV)
Rio de Janeiro, Brazil
jorge.poco@fgv.br

***Abstract*—Over the past four decades, urban perception has become a vital area of research that intersects multiple fields, such as criminology, psychology, and urban planning. This interdisciplinary approach seeks to understand and interpret how people perceive urban environments and how these perceptions shape their behavior. The surge in data collection methods, driven by modern web technologies and services, has enabled researchers to apply techniques from various domains to better quantify and analyze urban perception. In this study, we present the UrbanFormer, a vision transformer-based model, to address the task of urban perception analysis, leveraging the widely-used Place Pulse 2.0 dataset. Our focus is on the safety category, a key issue in urban perception, while employing vision transformer and explainability methods to provide insights into the decision-making process behind perception analysis.**

***Index Terms*—urban perception, urban computing, computer vision, deep learning, street view images, human perception, built environment**

I. Introduction

Studies have shown that the visual aesthetics of urban environments strongly influence human perception and behavior [13]. The “Broken Window Theory” [41] suggests that signs of neglect, such as broken windows, graffiti, and litter, contribute to negative social outcomes and higher crime rates. Currently, with the advance of deep learning techniques and street view imagery services, some studies gather data from websites and online surveys to study urban perception, such as MIT Media Lab’s Place Pulse “Which looks more safe?” [31], *scenic-or-not* [36], “What makes London beautiful?” [30], and City-SAFE [6]. Other research quantifies greenery, identifying green areas and their influence on urban perception [16]. In addition, some studies explore the relationship between violence levels, the presence of trees, and the human development index [3], [29], as well as the correlation between graffiti and perceptions of urban safety [14], [38]. Moreover, they categorize cities based on the most common types of objects or visual elements (e.g., trees, garbage, buildings, fences, graffiti) and the associated perception of safety [21], [23]. Although these studies analyze the visual appearance of cities and correlate them with demographic factors, no one can explain the behavior of human perception in street view images; To address this challenge, this paper aims to investigate the correlation between human perception of safety and the impact of the visual appearance of urban visual environment on the misperception of safety.

Contributions. This work introduces a novel approach leveraging the OneFormer model for segmentation and a modified Vision Transformer (ViT) for classification to achieve high performance in both binary and 10-label classification tasks. Furthermore, we evaluate the importance of visual elements within images by measuring the intersection over union (IoU) between segmentation masks and model-generated explanations, providing deeper insights into model interpretability and feature relevance.

II. Related works

Urban perception is a crucial area in urbanism and urban planning. This research field aims not only to create highly accurate prediction models [22], [33] but also to understand the urban environment and its impact on residents [7], [40]. The main goal is to develop a model that maps a city’s visual appearance and determines its uniqueness. For example, “What makes Paris look like Paris?” [8], or “What makes an outdoor space beautiful?” [36], or “What makes London look beautiful, quiet and happy?” [30] Additionally, some research incorporates other data, such as crime rates and robbery statistics [3], [15], [35], or aims to map the impact of graffiti in large cities and compares it with the human development index [2], [38].

The MIT Media lab introduced a significant dataset in urban perception, the MIT Place Pulse dataset [31]. This dataset consists of comparisons between pairs of images across various categories (e.g., safety, liveliness, wealth). This research aimed to perform urban mapping using urban perception scores and to localize these scores within the target city [25], [27]. Feature extractors such as GIST, DeCAF, and ImageNet were used to train image representations along with their respective perceptual scores [25], [47]. Other studies sought to extract more detailed information about the visual appearance of images using complex methods like convolutional neural networks (CNNs) [10], [29] and analyzed greenery areas in cities [16], [20]. Additionally, segmentation techniques have been employed to analyze the presence of visual elements and their correlation with safety perception [42], [45] or to understand the relationship between model predictions and human perception [17], [19], [21].

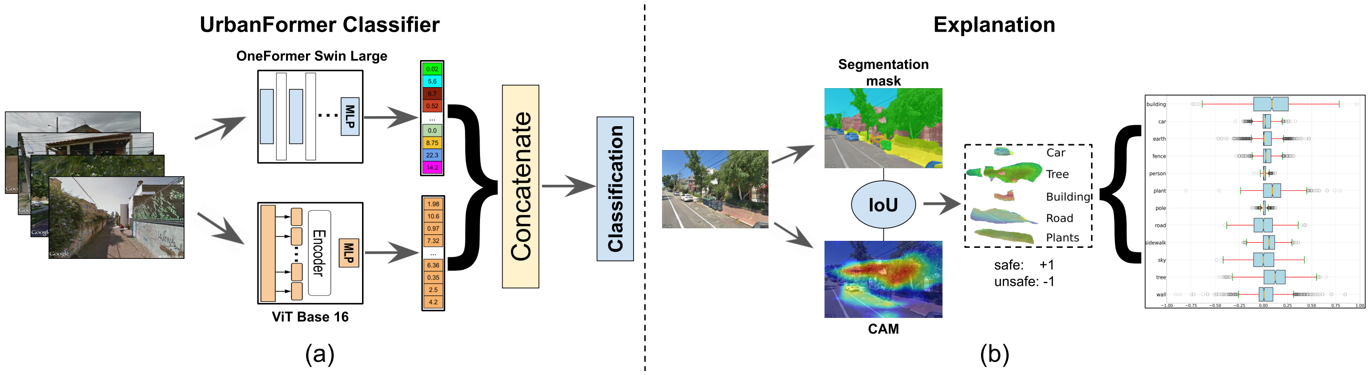


Fig. 1. (a) The proposed Urbanformer combines the OneFormer vectorized output with a modified ViT probability vector output to predict street human perception between safe and unsafe. (b) Our explanation technique combines the segmentation mask and the Grad-CAM method to analyze the relationship between safety perception and the presence of visual elements.

III. Methodology

Our methodology comprises three main steps: (i) urban perception quantification; we begin with an exploratory data analysis, quantifying urban perception scores derived from the street images to gain insights from the data; (ii) classifier model, Figure 1 (a) presents our proposed model called UrbanFormer. This model integrates the OneFormer segmentation model with a modified Vision Transformer to classify the perceived safety of street images; and (iii) model explanation, Figure 1 (b) illustrates our explanation technique. We use a class activation maps-based method to compute the intersection over union (IoU) between the segmentation mask from OneFormer and the model explanations, enabling us to assess the importance of each visual element.

A. Urban perception quantification

We study the MIT Place Pulse 2.0 dataset, composed of approximately 1.22 million random comparisons between pairs of 111,390 images, providing the image ID, latitude, longitude, and the respective winner. We implement the algorithm "strength of schedule" [28] to preprocess those comparisons. For each comparison between images i and j in the category k (e.g., safe), we define *intensity of perception* of the image i as the percentage of times that the image was selected and is affected by the intensity of the compared images j.

$$Q_{i,k} = \frac{10}{3}(W_{i,k} + \frac{1}{n_i}(\sum_{j_w}^{n_i} W_{j_w,k}) - \frac{1}{m_i}(\sum_{j_l}^{m_i} L_{j_l,k}) + 1) \quad (1)$$

The Equation 1 represents the perceptual score of image i, referred to as the *Q-score*, and denoted $Q_{i,k}$, within category k. Here, $W_{i,k}$ and $L_{i,k}$ represent the win and loss rates of image i in category k. In addition, n_i is the number of images j that image i has won against, and m_i is the number of images j that image i has lost to. Finally, following previous studies on visual assessment [26], [31], the perceptual score Q is scaled to fit a range from 0 to 10, where an image with a score close to zero is perceived as very unsafe, and a score close to 10 is perceived as very safe. This scaling is achieved by adding a constant value of 1 and multiplying by $\frac{10}{3}$.

B. UrbanFormer classifier

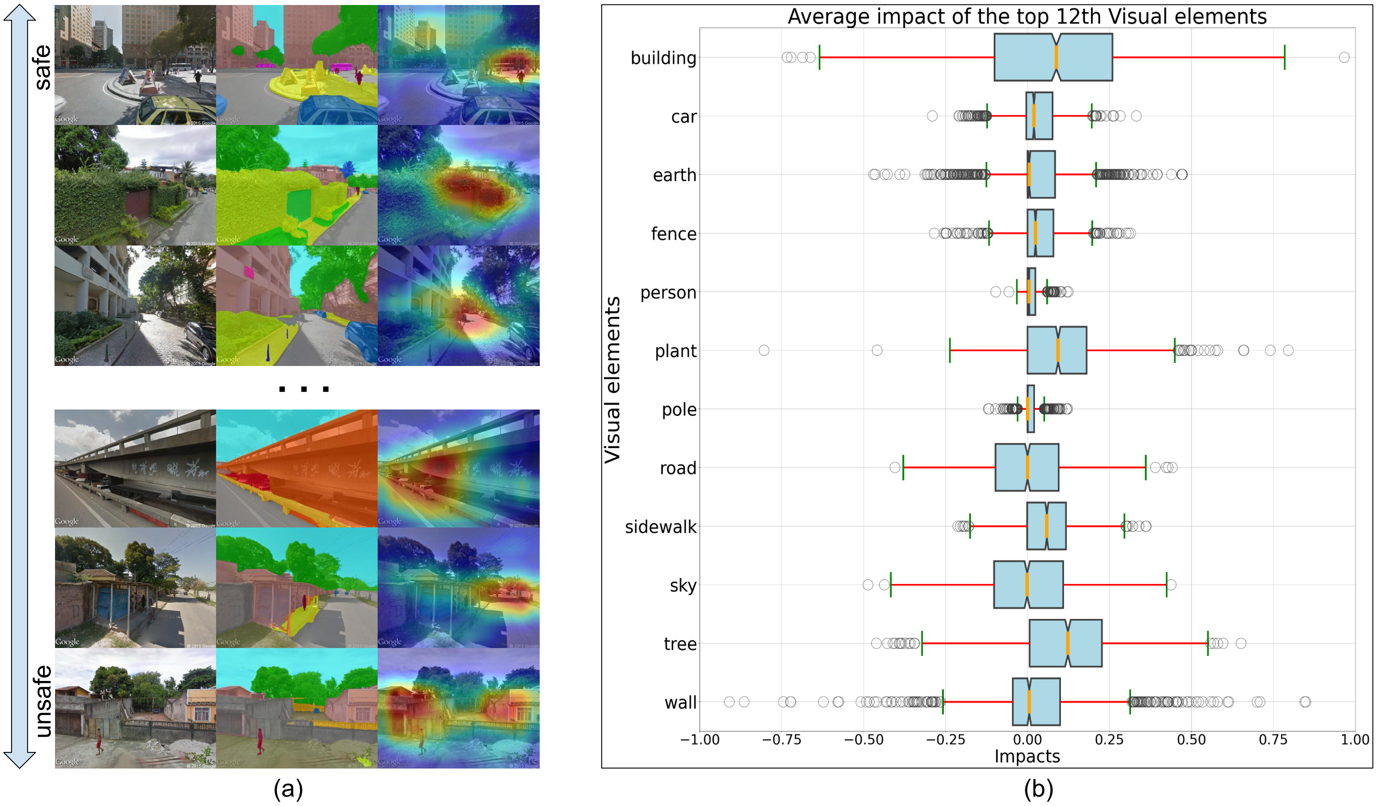
Figure 1 (a) shows our proposed classifier UrbanFormer, which concatenates the probability vector of the perception category extracted from the Vision Transformer (pre-trained on ImageNet [9]) with semantic features obtained from the OneFormer segmentation model (pre-trained on ADE20K [12]). We modify the ViT-B-16 model by adding 3 dense layers of 512, 512, and 1028, respectively. Then, we fused it with the pixel ratios vector of size 150 obtained by the OneFormer segmentation model. Then, we evaluate the output using the cross-entropy loss with the logits function.

C. Model explanation

We correlate human perception and the presence of visual elements in street images to explain predictions. For the explanation of convolutional networks, saliency map-based methods were widely adopted [24], [37], we select the Grad-CAM [34] due to the better behavior and performance against adversarial attacks or noise-adding techniques [1], [11]. Grad-CAM (Equation 2) highlights important image regions for model predictions by using target class gradients on the final convolutional layer's feature maps, followed by Global Average Pooling (GAP) to compute neuron importance.

$$\delta_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{GAP}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{grad-backprop}} \quad (2)$$

Where δ_k^c represents the neuron importance weights, c is the class, $Z = u \times v$ the size of the image, k is the kth feature map, A_{ij}^k is the feature map, y^c the score for class c, and $\frac{\partial y^c}{\partial A_{ij}^k}$ is the gradient vector obtained by backpropagation. We assess the importance of visual elements by calculating the IoU between the segmentation mask and CAM areas, identifying object overlap percentages, and determining relevance based on CAM values.



(a) (b)

Fig. 2. We present the outputs of the Grad-CAM algorithm and the OneFormer output: (a) The top 3 safest and the top 3 less safe images, segmentation, and Grad-CAM outputs. (b) The average impact of the top 12 visual elements with presence in almost 90% of the 108,820 images from the dataset. We obtain this relevance by doing an IoU between the activations and the segmentation mask, where positive values correspond to safe predictions and negative values correspond to no safe predictions.

IV. Discussions and Results

This work presents a methodology for analyzing urban perception using the MIT Place Pulse 2.0 dataset, focusing on the prediction of safety categories. In addition, we analyze the visual explanation using the Grad-CAM method to obtain information about which visual elements are the most relevant.

A. Dataset exploration and preparation

We identified that 2,471 locations (samples with the same latitude and longitude) have more than one ID assigned. For example, in Santiago (Chile), we identify 130 repeat locations, 126 in Berlin (Germany), 112 in Montreal (Canada), and so on. Lastly, Tel Aviv (Israel) and Seattle (USA) with 5 repeated locations, respectively. This repetition reduces the total number of images evaluated from 111,390 to 108,820 images. In addition, we noted the imbalance of sample sizes from different cities, e.g., Amsterdam has 622 images, while Atlanta has 3,965 images. Thus, we use the entire dataset for experiments, divided into 75% for training+validation and 25% for testing. Although several works use MIT Place Pulse 2.0, most of them perform regression tasks [5], [45] or pairwise learning using SiameseNets [4], [17], in this study we perform two classification experiments: (i) binary classification (safe versus no safe) and (ii) 10-label classification, that is, scores between 0-1 as label 0, 1-2 as label 1, and so on.

B. Model training and performance

We perform a classification task using the OneFormer segmentation model fused with our adapted ViT-based model. We initialize all weights using the Xavier uniform criteria and freeze all encoder layers. We perform the experiments using binary classification and 10-label classification. In addition, we employed grid search 5-fold stratified cross-validation on the training+validation set to maintain the proportion of categories during training. Table I and Table II reports the average classification metrics from five cross-validation runs for our model, as well as for previous works that perform classification tasks using either binary or 10-label approaches. Notably, most of these works report only accuracy, disregarding other performance metrics. We report the results of using the modified ViT only and fused with OneFormer, showing that for the 10-label classification, only our ViT-OneFormer has better performance. Further, we include the AUC metric to demonstrate our model's effectiveness and robustness in correctly identifying and differentiating between the categories.

C. Model explanations

Figure 2 (a) shows the 3 safest images and the 3 least safe images. It also shows the segmentation mask and the Grad-CAM obtained for the ground truth label (e.g., if the

TABLE I

ACCURACY REPORT USING BINARY CLASSIFICATION

Model	Acc
PspNet+VGG [21]	48.38
DeepLabV3+VGG [21]	51.93
DSAPN+ResNet [43]	64.87
MTDRALN-LC [19]	65.07
MTDRALN-TC [19]	65.82
VGG+ImageNet [22]	65.72
VGG-GAP+ImageNet [22]	66.09
VGG+Places365 [22]	66.46
VGG-GAP+Places365 [22]	66.96
VGG19+ImageNet [4]	67.01
PSPNet+SVR [44]	70.63
DeiT+ResNet50 [32]	71.16
ViT-nn (Ours)	**71.29**
ViT-nn+OneFormer (Ours)	**75.68**

sample is safe, we calculate the Grad-CAM for the safe prediction class). We set the Grad-CAM threshold at 0.3 to keep relevant information, complement information, and avoid irrelevant information [39]. We apply the Intersection over the Union (IoU) method between the segmentation mask and the Grad-CAM obtained. On average, Grad-CAM regions overlap the most present visual elements, such as trees, buildings, roads, sidewalks, walls, fences, earth, and the sky. These visual elements are present in almost 96% of the 108,820 images on the dataset. Then, we assign a positive weight to positive samples (e.g., safe images) and a negative weight for negative samples. Figure 2 (b) shows the average impact of the 12 most prevalent visual elements in the images. We observed that IoU scores for trees, sidewalks, plants, fences, and buildings have a stronger association with perceptions of safety. Visual elements such as walls, dirt, and trashcans appear more frequently in negative samples, suggesting their presence is associated with unsafe streets. Of the 150 identified visual elements, about 120 appear in less than 1% of images (e.g., flowers, pots, pools), contributing a zero average impact to this analysis. However, visual elements with high presence and a zero mean average indicate relevance to both safe and unsafe categories. In particular, the absence of these elements is linked to an unsafe perception of the streets.

D. Limitations

We found that it is not possible to study specific cities; this happens because comparisons were made randomly. In addition, the number of image comparisons is not balanced; most images were compared only three times, while others were up to 78 times. Further, the number of images collected per city is not proportional; we found cities with less than 600 images and cities with more than 3,500 images. In addition, after calculating the perceptual score, we obtained an imbalance across classes in both binary and 10-label cases, which could skew the model's performance by favoring overrepresented classes. Although data augmentation was applied, a more balanced dataset could provide a more accurate representation of the model's ability to generalize across all classes.

TABLE II

ACCURACY REPORT USING 10-LABEL CLASSIFICATION

Model	Acc
ResNet50 [18]	71.33
SegFormerB5+RF [46]	42.8
VGG19 [46]	75.2
ConvNeXt-B [46]	76.4
SFB5+ConvNeXt-B+RF [46]	78.1
ViT-nn (Ours)	74.97
ViT-nn+OneFormer (Ours)	**78.68**

Moreover, due to limited access to high-performance computational resources, particularly GPUs and memory, the ability to run extensive fine-tuning and parameter tuning was restricted. This limitation affected model training time, prevented testing on larger model architectures, and constrained batch sizes, potentially affecting the model's generalization performance. Additionally, limited resources restricted experimentation with multiple configurations to identify optimal hyperparameters. For experiments, data from all cities is necessary to ensure good model performance. These situations prevent the creation of a general model for all cities, as there are not enough training samples or comparisons for each city.

V. CONCLUSIONS

In conclusion, this study applies Deep Learning techniques—specifically semantic segmentation, classification models, and the Grad-CAM explainer—to explore urban safety perception. Semantic segmentation revealed spatial relationships between objects within street images, while Grad-CAM provided insights into their role in safety perception prediction. Our analysis shows that objects like trees, cars, fences, the sky, and buildings significantly influence safety perceptions, though this may partly result from their prevalence in most images. Our findings underscore the need for a nuanced approach to understanding urban environments, suggesting that future research should explore the contextual significance of these objects beyond their mere presence. This work lays the foundation for more informed urban planning and design initiatives to enhance safety and improve the overall quality of urban life.

ACKNOWLEDGEMENT

This work was supported by the National Council for Scientific and Technological Development (CNPq, grant #311144/2022-5), Carlos Chagas Filho Foundation for Research Support of Rio de Janeiro State (FAPERJ, grant #E-26/204.593/2024), São Paulo Research Foundation (FAPESP, grant #2021/07012-0) and the School of Applied Mathematics at Fundação Getulio Vargas.

References

[1] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps (2018)

[2] Alzate, J.R., Tabares, M.S., Vallejo, P.: Graffiti and government in smart cities: a deep learning approach applied to medellin city, colombia. In: International Conference on Data Science, E-learning and Information Systems 2021. pp. 160–165 (2021)

[3] 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) (2014)

[4] Beaucamp, B., Leduc, T., Tourre, V., Servieres, M.C.J.: The whole is other than the sum of its parts: Sensibility analysis of 360° urban image splitting. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2022)

[5] Buil-Gil, D., Solymosi, R.: Using crowdsourced data to study crime and place. SocArXiv (2020)

[6] Correia, G.P., da Costa, C.: City-safe: Estimating urban safety perception (2019)

[7] De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., Lepri, B.: The death and life of great italian cities: a mobile phone data perspective. In: Proceedings of the 25th international conference on world wide web. pp. 413–423 (2016)

[8] Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.: What makes paris look like paris? (2012)

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2020)

[10] Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A.: Deep learning the city : Quantifying urban perception at A global scale (2016)

[11] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning (2018)

[12] Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2989–2998 (2023)

[13] Keizer, K., Lindenberg, S., Steg, L.: The spreading of disorder. Science (New York, N.Y.) **322**, 1681–5 (12 2008)

[14] Lavi., B., Tokuda., E., Moreno-Vera., F., Nonato., L., Silva., C., Poco., J.: 17k-graffiti: Spatial and crime data assessments in são paulo city. In: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,. pp. 968–975. INSTICC, SciTePress (2022)

[15] 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)

[16] Li, X., Zhang, C., Li, W.: Does the visibility of greenery increase perceived safety in urban areas? evidence from the place pulse 1.0 dataset. ISPRS Int. J. Geo Inf. **4**, 1166–1183 (2015)

[17] Li, Z., Liu, P., Shi, J., Xing, Y.: Research on street space quality combined with attention multi-task deep learning. 2021 2nd International Conference on Big Data Economy and Information Management (BDEIM) pp. 434–441 (2021)

[18] Ma, Z.: Deep exploration of street view features for identifying urban vitality: A case study of qingdao city. Int. J. Appl. Earth Obs. Geoinformation **123**, 103476 (2023)

[19] Min, W., Mei, S., Liu, L., Wang, Y., Jiang, S.: Multi-task deep relative attribute learning for visual urban perception. IEEE Transactions on Image Processing **29**, 657–669 (2019)

[20] 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)

[21] Moreno-Vera, F.: Understanding safety based on urban perception. In: International Conference on Intelligent Computing. pp. 54–64. Springer (2021)

[22] Moreno-Vera., F, Lavi., B., Poco, J.: Quantifying urban safety perception on street view images. In: International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (2021)

[23] Moreno-Vera, F, Lavi, B., Poco, J.: Urban perception: Can we understand why a street is safe? In: Mexican International Conference on Artificial Intelligence. pp. 277–288. Springer (2021)

[24] Moreno-Vera., F., Medina., E., Poco., J.: Wsam: Visual explanations from style augmentation as adversarial attacker and their influence in image classification. In: Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,. pp. 830–837. INSTICC, SciTePress (2023). https://doi.org/10.5220/0011795400003417

[25] Naik, N., Raskar, R., Hidalgo, C.A.: Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. The American Economic Review **106**, 128–132 (2016)

[26] Nasar, J.L.: The evaluative image of the city (1998)

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

[28] Park, J., Newman, M.: A network-based ranking system for us college football. Journal of Statistical Mechanics: Theory and Experiment **2005**, P10014 – P10014 (2005)

[29] 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

[30] 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)

[31] Salesses, M.P.: Place Pulse: Measuring the collaborative image of the city. Ph.D. thesis, Massachusetts Institute of Technology (2012)

[32] Sangers, R., van Gemert, J.C., van Cranenburgh, S.: Explainability of deep learning models for urban space perception. ArXiv (2022)

[33] Santani, D., Ruiz-Correa, S., Gatica-Perez, D.: Looking south: Learning urban perception in developing cities. ACM Transactions on Social Computing **1**(3), 1–23 (2018)

[34] 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)

[35] Sengupta, N., Vaidya, A., Evans, J.: In her shoes: Gendered labelling in crowdsourced safety perceptions data from india. ACM Conference on Fairness, Accountability, and Transparency (2023)

[36] Seresinhe, C.I., Preis, T., Moat, H.S.: Using deep learning to quantify the beauty of outdoor places. Royal Society open science **4**(7), 170170 (2017)

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

[38] Tokuda, E.K., Silva, C.T., Jr., R.M.C.: Quantifying the presence of graffiti in urban environments. CoRR **abs/1904.04336** (2019)

[39] Vago, N.O.P., Milani, F., Fraternali, P., da Silva Torres, R.: Comparing cam algorithms for the identification of salient image features in iconography artwork analysis. Journal of Imaging **7** (2021)

[40] Wendt, M.: The importance of death and life of great american cities (1961) by jane jacobs to the profession of urban planning. New Visions for Public Affairs **1**, 1–24 (2009)

[41] Wilson, J.Q., Kelling, G.L.: Broken windows. Atlantic monthly **249**(3), 29–38 (1982)

[42] Xu, X., Qiu, W., Li, W., Liu, X., Zhang, Z., Li, X., Luo, D.: Associations between street-view perceptions and housing prices: Subjective vs. objective measures using computer vision and machine learning techniques. Remote. Sens. **14**, 891 (2022)

[43] Zhang, C., Wu, T., Zhang, Y., Zhao, B., Wang, T., Cui, C., Yin, Y.: Deep semantic-aware network for zero-shot visual urban perception. International Journal of Machine Learning and Cybernetics **13**, 1197 – 1211 (2021)

[44] Zhang, F., Hu, M., Che, W., Lin, H., Fang, C.: Framework for virtual cognitive experiment in virtual geographic environments. ISPRS Int. J. Geo Inf. **7**, 36 (2018)

[45] 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)

[46] Zhao, X., Lu, Y., Lin, G.: An integrated deep learning approach for assessing the visual qualities of built environments utilizing street view images. Engineering Applications of Artificial Intelligence **130** (2024)

[47] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)